Interpreting and Presenting Statistical Results

Mike Tomz   Jason Wittenberg
Harvard University

APSA Short Course
September 1, 1999

A Reader's Nightmare

Table 19: Vote for George W. Bush? Heteroskedastic FIML Model

	Model 1	Model 2	Model 3
INCOME	5.417 ***	5.746 ***	6.185 **
	(2.10)	(2.21)	(2.81)
GENDER	9.542 *		9.826 **
	(5.02)		(4.27)
SUBURBS	7.945	7.135	8.107 *
	(6.10)	(5.49)	(4.14)
RACE	5.207 **	4.217 **	
	(2.60)	(1.83)	
CONST	-4.061 ***	-5.598 ***	-5.323 ***
	(1.35)	(1.60)	(1.53)
Pseudo R^2	0.37	0.34	0.35

Note: N=3017. Estimated coefficients are given with standard errors in parentheses underneath. * *p*<.05, ** *p*<.01, *** *p*<.001

Our method helps researchers

- Convey results in a reader-friendly manner
- Uncover new facts about the political world

The method does *not* require

- collecting new data
- changing the statistical model
- introducing new assumptions

This course has three parts.

1. The problem
2. A solution
3. Examples, with software

Good methods of interpretation should satisfy three criteria:

1. Convey numerically precise estimates of the quantities of substantive interest

2. Include reasonable estimates of uncertainty about those estimates

3. Require no specialized knowledge to understand

The most common methods of interpretation do not satisfy our criteria.

1. Listing coefficients and se's
   - not intrinsically interesting
   - hard to understand

1. Computing expected values or first differences exclusively
   - ignores sampling error and fundamental uncertainty

Best current practice

- “Fitted”, “predicted” (expected) values

  Compute an expected value of the dependent variable given the estimated coefficients and interesting values of the explanatory variables.

- First Differences

  Compute the difference between two expected values.

Both ignore sampling error and fundamental uncertainty.

Two kinds of uncertainty

- Sampling error

  How much do estimated quantities differ from sample to sample?

- Fundamental uncertainty

  How much do unmodeled random factors influence the outcome?

A better method

Our Goal

To obtain precise quantities of interest and estimates of uncertainty that are easy to understand.

The technique

Use simulation to extract all available information from a statistical model.

What is simulation?

1. Simulation is analogous to survey sampling.

Survey Sampling	Simulation
Learn about a population by taking a random sample from it	Learn about a distribution by taking random draws from it
Use the random sample to estimate a feature of the population	Use the random draws to approximate a feature of the distribution
The estimate is arbitrarily precise for large N	The approximation is arbitrarily precise for large M
Example: estimate the mean of the population	Example: approximate the mean of the distribution

2. Example: Approximating the mean of a distribution is like estimating the mean of a population.

$$\frac{1}{\mathrm{n}}\sum_{i=1}^{n} y_i \quad \text{versus} \quad \frac{1}{m}\sum_{j=1}^{m} \tilde{y}_j$$

What is a model?

A statistical model is a representation of the social process that produces the outcomes of interest.

For example, linear regression:

$$Y_i = \beta_{0i} + \beta_{1i} X + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

Equivalently:

$$Y_i \sim N\left(\mu_i, \sigma^2\right)$$
$$\mu_i = \beta_{0i} + \beta_{1i} X_i$$

Logit and other models

Logistic regression:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$
$$\pi_i = \frac{1}{1+e^{-X_i \beta}}.$$

Most models can be written as:

$$Y_i \sim f(\theta_i, \alpha)$$
$$\theta_i = g(X_i, \beta)$$

What parts of the model do we simulate?

Our Goal: Generate simulations of the outcome variable that account for both sampling error and fundamental uncertainty

Consider the Logit:

$$Y_i \sim \text{Bernoulli}(\pi_i)$$

$$\pi_i = \frac{1}{1+e^{-X_i\beta}}.$$

Simulate the uncertain quantities.

How do we simulate the parameters?

1. Obtain the estimated coefficients and variance matrix

2. Draw (simulate) the parameters from a multivariate normal distribution

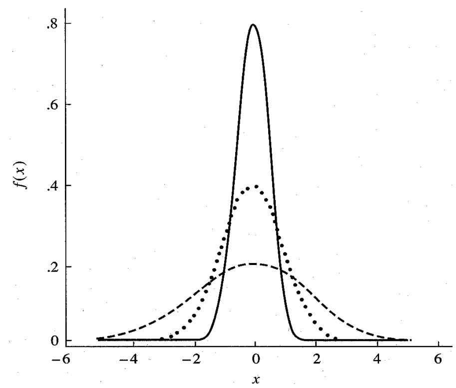
Obtain the estimated coefficients and variance matrix

```
reg yreg x1

  Source |       SS       df       MS              Number of obs =    1000
---------+------------------------------           F(  1,   998) =   52.75
   Model |  364.960926     1  364.960926           Prob > F      =  0.0000
Residual |  6904.20739   998  6.91804348           R-squared     =  0.0502
---------+------------------------------           Adj R-squared =  0.0493
   Total |  7269.16832   999  7.27644476           Root MSE      =  2.6302

------------------------------------------------------------------------------
    yreg |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |    2.15182    .296261      7.263   0.000       1.570454    2.733186
   _cons |  -1.174021   .1698267     -6.913   0.000       -1.50728   -.840763
------------------------------------------------------------------------------
```

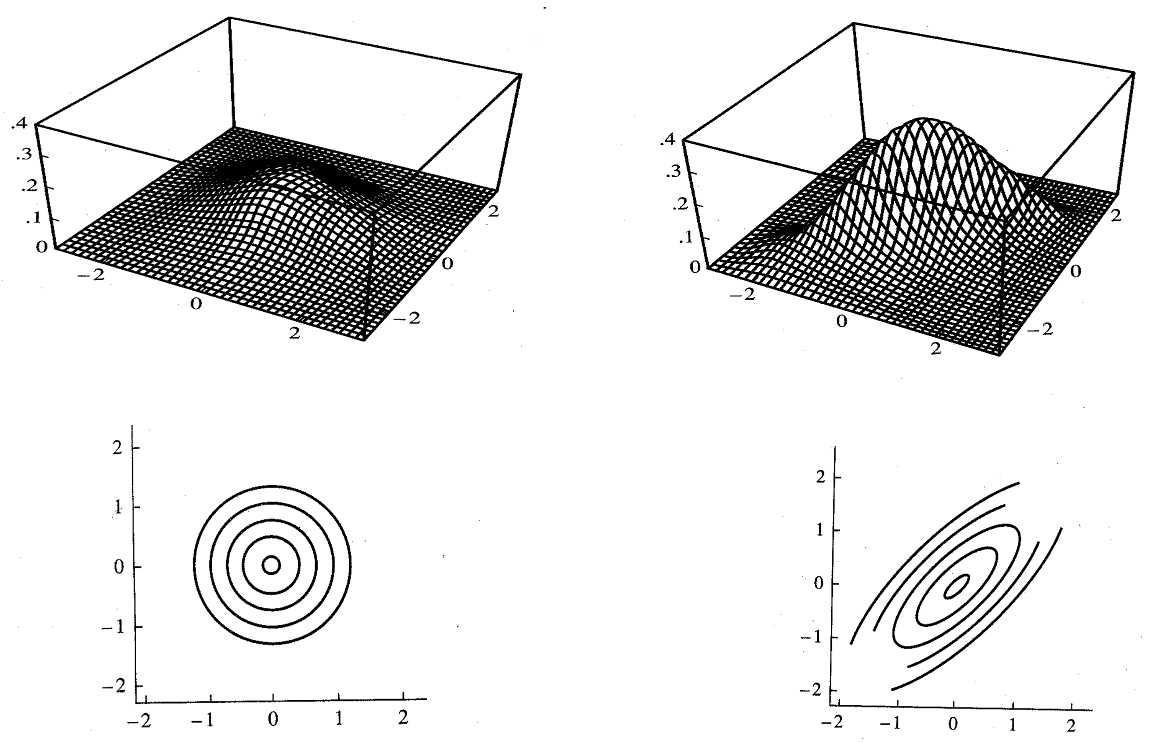
$$\hat{\gamma} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\alpha} \end{bmatrix} \quad \hat{V} = \begin{bmatrix} v_{\hat{\beta}_{11}} & v_{\hat{\beta}_{12}} & v_{\hat{\beta}_1\hat{\alpha}} \\ v_{\hat{\beta}_{21}} & v_{\hat{\beta}_{22}} & v_{\hat{\beta}_2\hat{\alpha}} \\ v_{\hat{\alpha}\hat{\beta}_1} & v_{\hat{\alpha}\hat{\beta}_2} & v_{\hat{\alpha}} \end{bmatrix}.$$

A Primer on Normal Distributions

1. Univariate normal distribution



2. Bivariate normal distributions



Draw (simulate) parameters from a multivariate normal distribution

$$\tilde{\gamma} \sim N(\hat{\gamma}, \hat{V})$$

Each draw will be a vector of simulated parameters:

$$\begin{bmatrix} \tilde{\beta}_{11} \\ \tilde{\beta}_{21} \\ \tilde{\alpha}_{1} \end{bmatrix} \begin{bmatrix} \tilde{\beta}_{12} \\ \tilde{\beta}_{22} \\ \tilde{\alpha}_{2} \end{bmatrix} \cdots \begin{bmatrix} \tilde{\beta}_{1M} \\ \tilde{\beta}_{2M} \\ \tilde{\alpha}_{M} \end{bmatrix}$$

To simulate *one* value of *Y* from

$$Y_i \sim f(\theta_i, \alpha), \quad \theta_i = g(X_i, \beta)$$

1. Choose a scenario, X_c.
2. Draw one value of $\tilde{\gamma} = [\tilde{\beta} \ \tilde{\alpha}]$.
3. Compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$.
4. Draw $\tilde{Y}_c$ from $f(\tilde{\theta}_c, \tilde{\alpha})$.

[Repeat steps 2-4 many times to approximate the distribution of $Y|X_c$]

Example of simulating $\tilde{Y}_c$

\

Regress income on education, as in

$$Income \sim N\left(\mu\ ,\sigma^2\right)$$

$$\mu = \beta_0 + \beta_1 \times education.$$

To simulate one value of *income,*

1. Choose a scenario for $education$,
   for example $education_c = 12$ years.

2. Draw one value of $\tilde{\gamma} = \left[\tilde{\beta}_0\ \ \tilde{\beta}_1\ \ \tilde{\sigma}^2\right]$.

3. Compute $\tilde{\mu}_c = \tilde{\beta}_0 + \tilde{\beta}_1 \times education_c$.

4. Draw one value of $income$, conditional
   on $education_c$, from $income_c \sim N(\tilde{\mu}_c, \tilde{\sigma}^2)$.

[Repeat steps 2-4 many times to approximate the distribution of income|education=12 years]

With $\tilde{Y}_c$ we can compute any quantity, including:

- Predicted values
- Expected values
- First differences

To simulate *one* expected value,

1. Choose a scenario, X_c.

2. Draw one value of $\tilde{\gamma} = [\tilde{\beta} \;\; \tilde{\alpha}]$.

3. Compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$.

4. Draw m values of $\tilde{Y}_c \sim f(\tilde{\theta}_c, \tilde{\alpha})$.

5. Calculate the mean

$$\tilde{E}(Y_c) = \sum_{\text{all } \tilde{Y}_c} \frac{\tilde{Y}_c}{m}.$$

For *one* first difference,

1. Choose starting scenario, X_s.
2. Calculate $\tilde{E}(Y_s)$.
3. Choose an ending scenario, X_e.
4. Calculate $\tilde{E}(Y_e)$.
5. Compute $\tilde{E}(Y_e) - \tilde{E}(Y_s)$.

With *many* draws of the quantity of interest, we can calculate:

- Average values
- Confidence intervals
- *Anything else we want!*

Tricks for simulating parameters

1. Simulate betas *and* ancillary parameters.

2. Transform parameters to make them unbounded and symmetric.

Tricks for simulating the quantity of interest

1. Increase simulations for more precision, reduce for computational speed.

2. Reverse transformations of the dependent variable.

3. Advanced users can take shortcuts to simulate the expected value and other quantities.

The method in practice
(Please try this at home!)

There are three main steps.

1. Estimate the model and simulate the parameters.
2. Choose a scenario for the explanatory variables.
3. Simulate quantities of interest.

We provide software to get you started.

How to use CLARIFY

- Clarify works with Stata version 5.0+
- Issue three simple commands.

`estsimp`	estimates the model and simulates the parameters
`setx`	sets values for explanatory variables (the X's)
`simqi`	simulates quantities of interest

Basic Syntax

The commands have an intuitive syntax.

```
estsimp model depvar indvars
setx indvar1 value1 indvar2 value2 …
simqi
```

Consider a hypothetical example:

```
estsimp logit y x1 x2 x3
setx x1 mean x2 p20 x3 .4
simqi
```

Here is the intermediate output.

Stata Results

```
. use testlog

. estsimp logit ylog x1 x2 x3

Iteration 0:   log likelihood = -693.04918
Iteration 1:   log likelihood = -486.16969
Iteration 2:   log likelihood = -467.79199
Iteration 3:   log likelihood = -466.72909
Iteration 4:   log likelihood = -466.72345

Logit estimates                                   Number of obs   =       1000
                                                  LR chi2(3)      =     452.65
                                                  Prob > chi2     =     0.0000
Log likelihood = -466.72345                       Pseudo R2       =     0.3266

------------------------------------------------------------------------------
    ylog |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
      x1 |   2.171535   .3031034      7.164   0.000       1.577463    2.765607
      x2 |   4.379989   .3345488     13.092   0.000       3.724285    5.035692
      x3 |  -4.238202   .3255733    -13.018   0.000      -4.876314    -3.60009
   _cons |  -1.093521    .239139     -4.573   0.000      -1.562225   -.6248177
------------------------------------------------------------------------------

Simulating main parameters.  Please wait....
% of simulations completed: 25% 50% 75% 100%

Number of simulations  : 1000
Names of new variables : b1 b2 b3 b4

.
```

Here is the final output.

Stata Results

```
. setx x1 mean x2 p20 x3 .4

. setx

You have set the following values for the explanatory variables:

-----------+-------------------------
 Variable |     Value      Description
-----------+-------------------------
       x1 |    .499777        mean
       x2 |   .1935453        p20
       x3 |         .4         .4
-----------+-------------------------

. simqi


      Quantity of Interest |       Mean        Std. Err.     [95% Conf. Interval]
---------------------------+----------------------------------------------------
               Pr(ylog==0) |    .7024219      .0243127        .65635    .7477913
               Pr(ylog==1) |    .2975781      .0243127      .2522087      .34365

.
```

Using estsimp to *estimate*

Type “estsimp” before a standard command

Which models does it estimate?

Model Name	Type of Y
regress	continuous
logit, probit	binary
ologit, oprobit	ordered
mlogit	categorical
poisson, nbreg	count

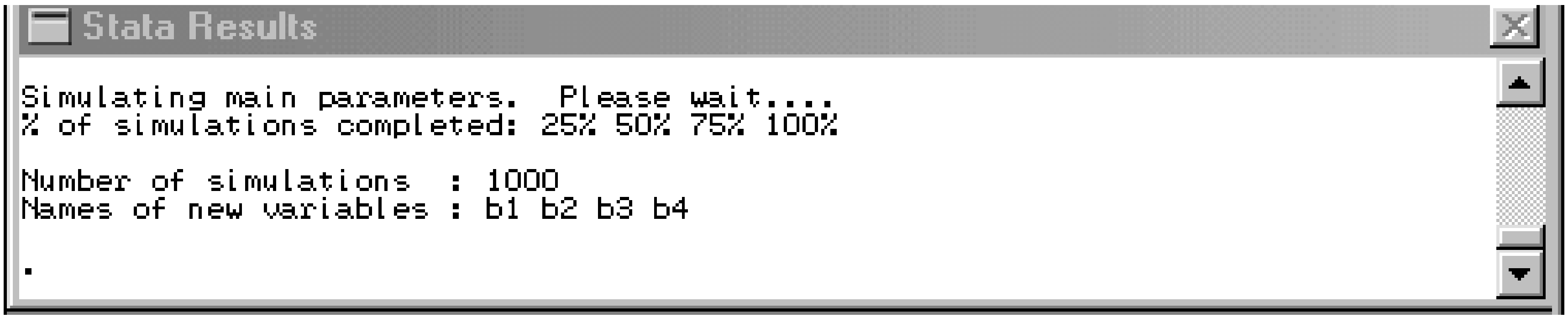
Using estsimp to *simulate*

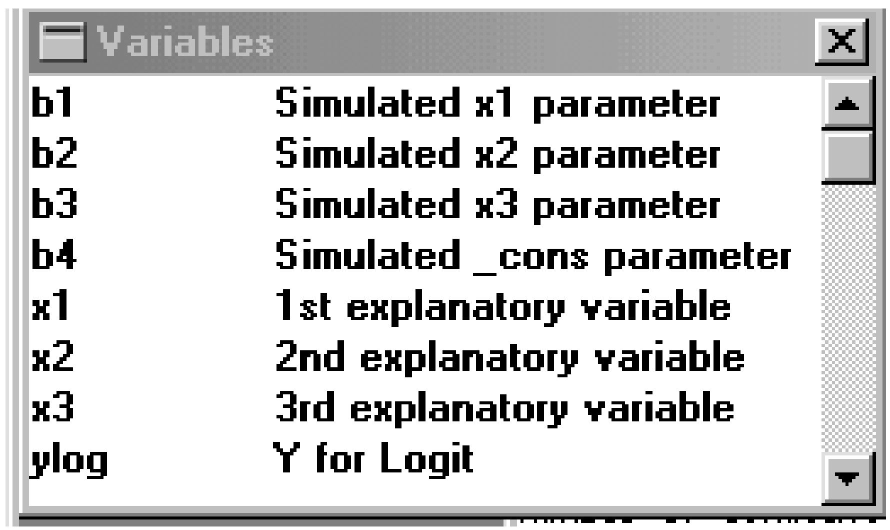
estsimp simulates *all* the parameters and stores them as new variables

obs	y	x
1	1	4.0
2	0	2.4
3	0	7.1
4	0	6.4
5	0	7.0
6	0	4.6
7	1	3.8
8	0	4.3
9	1	5.1
10	0	0.1

obs	y	x	b1	b2
1	1	4.0	0.09	-0.37
2	0	2.4	0.14	-0.67
3	0	7.1	0.17	-0.81
4	0	6.4	0.17	-0.93
5	0	7.0	0.09	-0.38
6	0	4.6	0.12	-0.66
7	1	3.8	0.17	-0.88
8	0	4.3	0.12	-0.60
9	1	5.1	0.13	-0.62
10	0	0.1	0.13	-0.56
11	.	.	0.15	-0.75
12	.	.	0.16	-0.79
13	.	.	0.12	-0.58
14	.	.	0.14	-0.66
15	.	.	0.16	-0.75

Ways to verify what clarify simulated





Or summarize the simulated parameters and compare to point estimates

Using setx

Use setx to choose a real or hypothetical value for each explanatory variable. Options include:

Value	Syntax
arithmetic mean	mean
median	median
minimum	min
maximum	max
#th percentile	p#
math expression	5*5
numeric value	#
contents of macro	\`macro'
value in #th obs	[#]

You can set each value individually or assign values to groups of variables.

Using simqi

By default, simqi displays sensible quantities of interest for each model. For example:

Model	Quantity of Interest
regress	$E(Y \mid X_c)$
logit	$Pr(Y=1 \mid X_c)$
oprobit	$Pr(Y=j \mid X_c)$ for all j

Simqi allows many options for displaying and saving quantities of interest.

How do education and age affect voter turnout?

Dependent variable:

Did the person vote? (1=yes,0=no)

Explanatory variables:

age, education, income, race, age^2

Logit model:

$$turnout_i \sim \text{Bernoulli}(\pi_i)$$

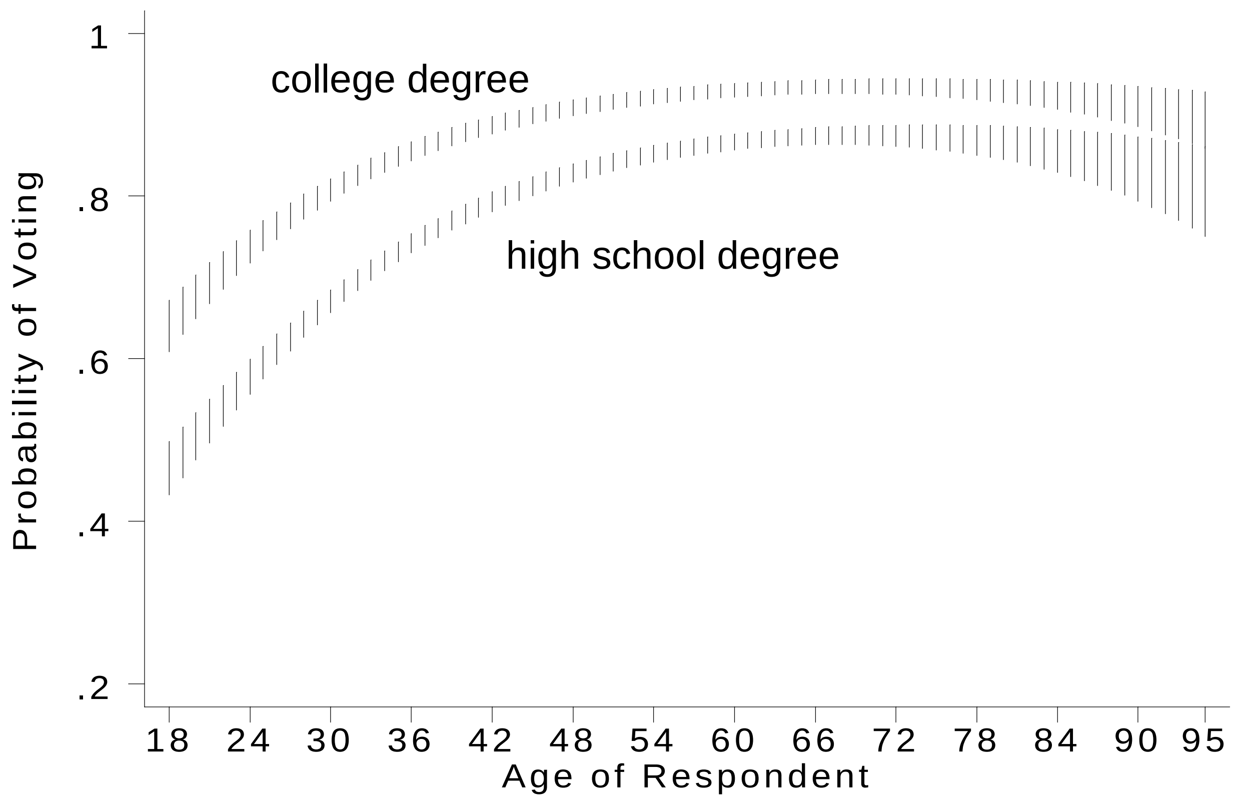
$$\pi_i = \frac{1}{1 + e^{-X_i \beta}}.$$

One way of presenting logit results

Explanatory Variable	Estimated Coefficient	Standard Error
Education	0.181 **	0.007
Age	0.109 **	0.006
$Age^2/100$	-0.078 **	0.007
Income	0.151 **	0.010
White	0.116 **	0.054
Constant	-4.715 **	0.174

Better ways to present logit results

"Other things equal, someone with a college degree is 9-12% more likely to vote than someone with only a high school education."



Calculating the probability of voting

Suppose we are interested in the probability of voting for the following scenario:

30 year old, college-educated black with an annual salary of $50,000

How would we simulate that?

```
estsimp logit turnout age agesqrd educate white income

setx age 30 agesqrd 30^2/100 educate 16 white 0 income 50

simqi
```

Calculating *changes* in the probability of voting

Suppose we wanted to know:

> For a typical American, how would the probability of voting change if we increased age from 20 to 40 years?

How would we simulate the answer?

```
setx age 45.4 agesqrd 45.4^2/100 educ mean white 1 inc mean

simqi, fd(pr) changex(age 20 40 agesqrd 20^2/100 40^2/100)
```

Calculating *percentage* changes in the probability of voting

Remember the example:

> “Other things equal, someone with a college degree is 9-12% more likely to vote than someone with only a high school education.”

Here is the code:

```
setx educ 12
simqi, prval(1) genpr(ed12)
setx educ 16
simqi, prval(1) genpr(ed16)
generate qoi = (ed16-ed12)*100/ed12
summarize qoi
```

Calculating and graphing probabilities for many scenarios

1. Estimate the model and simulate 1000 sets of parameters
2. Choose a scenario for the explanatory variables
3. Calculate1000 values of

$$\pi_c \equiv \Pr(turnout = 1) \mid \tilde{\beta}, X_c$$

4. Save the 95% confidence interval
5. Repeat steps 2-4 for many different scenarios
6. Graph the confidence intervals

Making the graph in Clarify

```
generate plo = .
generate phi = .
generate ageaxis = _n + 17 in 1/78
setx educate 12 white 1 income mean
local a = 18
while `a' <= 95 {
	setx age `a' agesqrd `a'^2/100
	simqi, prval(1) genpr(pi)
	_pctile pi, p(2.5,97.5)
	replace plo = r(r1) if ageaxis==`a'
	replace phi = r(r2) if ageaxis==`a'
	drop pi
	local a = `a' + 1
}
sort ageaxis
graph plo phi ageaxis, s(ii) c(||)
```

How does partisanship affect employment in state government?

Dependent variable:

ln(employment in state government)

Explanatory variables:

ln(state population),
ln(proportion of Democrats in House)

Regression model:

$$\ln(employment)_i \sim N\left(\mu_i, \sigma^2\right)$$
$$\mu_i = X_i \beta$$

Two ways of presenting regression results

Explanatory Variable	Estimated Coefficient	Standard Error
Lpop	0.779 **	0.026
Ldem	0.312 **	0.095
Constant	-2.057 **	0.228

or

Increasing Democratic control from half to two-thirds of the lower house tends to raise government employment by 9% (± 5%). Decreasing control to one-third would cut employment by 12% (± 6%).

Simulating state employment

1. Estimate the model and simulate 1000 sets of parameters.
2. Set ldem = ln(1/2) and lpop = ln(mean).
3. Simulate 1000 expected values of *employment*. To obtain one,

   draw lots of $\ln(employment) \mid \tilde{\beta}, X_s$

   exponentiate to recover $\tilde{e}mployment$

   take the mean of the $\tilde{e}mployment$'s

   this gives one value of $\tilde{E}(employment) \mid \tilde{\beta}, X_s$

4. Repeat steps 2-3 with ldem = ln(2/3) to simulate 1000 values of $\tilde{E}(employment) \mid \tilde{\beta}, X_e$
5. Subtract the expected values to get first differences

Code (available in the next release of Clarify)

```
estsimp regress lemp lpop ldem
summarize pop, meanonly
local popmean = r(mean)
setx lpop ln(`popmean') ldem ln(.5)
simqi, tfunc(exp) fd(ev) changex(ldem ln(.5) ln(2/3))
```

Why did Salinas win the Mexican election of 1988?

Dependent variable:

Vote for Salinas, Cardenas, Clouthier (3x1 vector)

Explanatory variables:

Attitude toward PRI
Many other variables

Multinomial logit model:

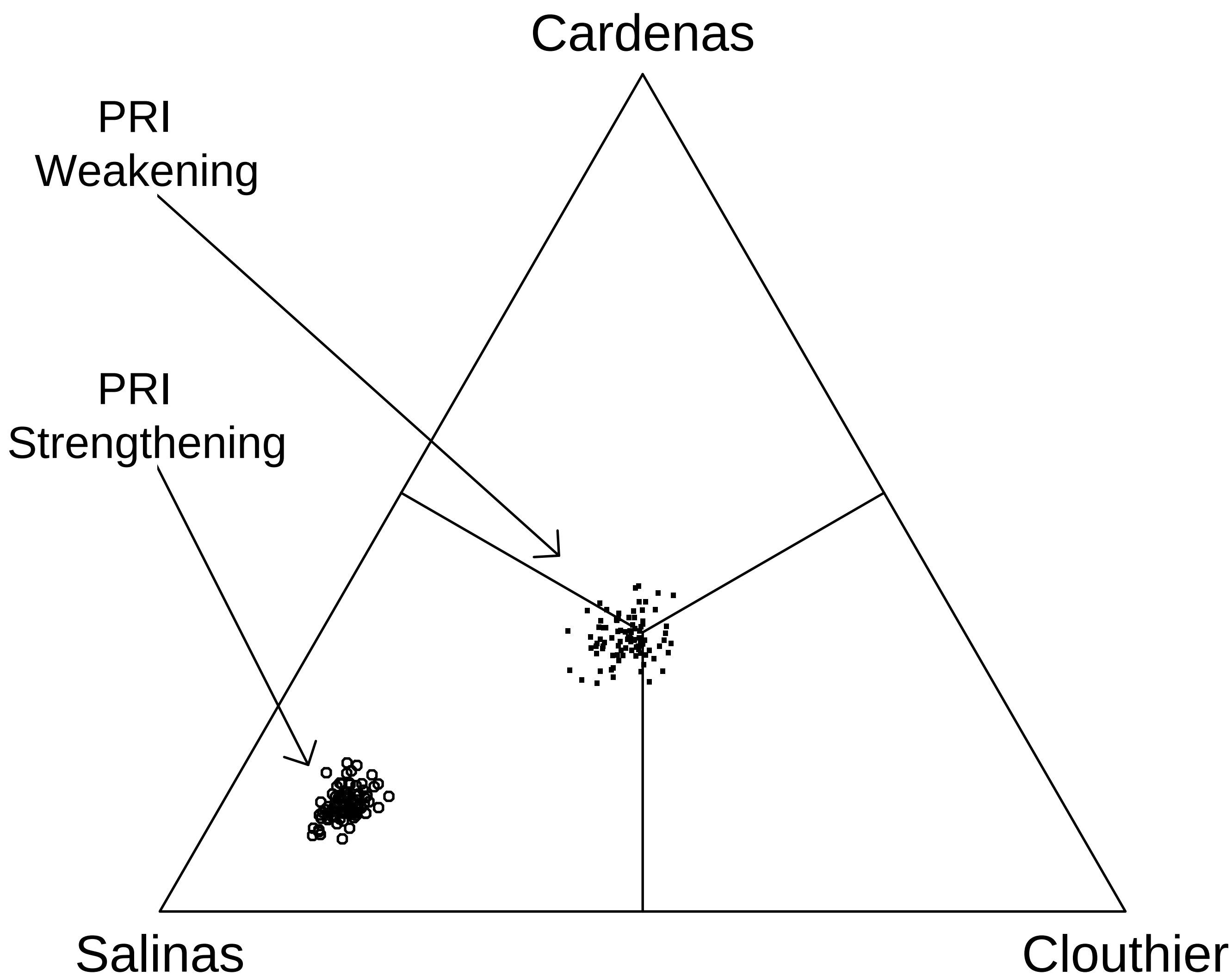
$$vote_i \sim Multinomial(\pi_i)$$

$$\pi_i = \frac{e^{X_i \beta_h}}{\sum_{k=1}^{3} e^{X_i \beta_k}}$$

One way to present multinomial logit results

	Salinas				Cardenas			
Variable	Coefficient	S.E.			Coefficient	S.E.	
pri82	-0.701	0.290	**	-1.137	0.244	**
pan82	2.483	0.377	**	1.027	0.368	**
novote82	0.110	0.347		-0.083	0.304	
deathok	0.072	0.104		0.152	0.094	
forinvok	0.209	0.121		0.139	0.107	
limimp	-0.055	0.106		-0.043	0.095	
paydebt	0.110	0.124		-0.224	0.106	**
keepind	0.034	0.112		0.162	0.098	
polint	-0.034	0.106		0.015	0.095	
auth	0.073	0.131		-0.051	0.118	
natecon	0.067	0.144		0.022	0.129	
futecok	-0.314	0.147	**	-0.060	0.130	
persecon	0.043	0.173		-0.141	0.151	
futperok	-0.208	0.103	**	-0.064	0.091	
school	-0.035	0.057		0.084	0.053	
age	-0.012	0.010		-0.005	0.009	
female	-0.098	0.237		-0.158	0.211	
prof	-0.267	0.345		-0.726	0.314	**
working	-0.385	0.347		0.121	0.283	
union	-0.805	0.278	**	-0.178	0.228	
townsize	0.099	0.078		0.020	0.068	
north	-0.100	0.308		-0.699	0.292	**
south	-0.800	0.341	**	-0.259	0.278	
zmgm	-0.246	0.312		0.158	0.273	
religion	0.089	0.076		-0.114	0.066	
pristr	-0.417	0.132	**	-0.341	0.116	**
othecok	0.912	0.157	**	0.927	0.139	**
othsocok	0.431	0.208	**	0.276	0.186	
ratemdm	-0.093	0.048		-0.137	0.043	**
traitmjc	1.070	0.104	**	0.139	0.123	
traitccs	0.165	0.097		0.751	0.073	**
one	-2.663	1.231	**	-1.228	1.069	

Another way to present multinomial logit results



Simulating the results of multi-candidate elections

1. Estimate the model, simulate 1000 sets of parameters
2. For each voter,
   (a) Assume the PRI is weakening and set other X's to their true values
   (b) Draw 1000 predicted values of vote88, one for each set of simulated parameters

   This gives us 1000 simulated elections.
3. For each simulated election, calculate the percentage of votes going to each party.
4. Repeat steps 2-3 assuming the PRI is strengthening
5. Graph the results

Simulating elections in Clarify

```
estsimp mlogit vote88 pri82 pan82 novote82  …
gen salinas = 0 in 1/1000
gen clouthie = 0 in 1/1000
gen cardenas = 0 in 1/1000
local nvoter=1
while `nvoter' <= 1359 {
     setx [`nvoter']
     setx pristr 1 othcok 3 othsocok 2
     simqi, pv genpv(vote)
     replace salinas = salinas + 1 if vote==1
     replace clouthie = clouthie + 1 if vote==2
     replace cardenas = clouthie + 1 if vote==3
     drop vote
     local nvoter = `nvoter' + 1
}
triplot clouthie cardenas salinas
```

How can I set the values of “interaction terms”?

Set the values by hand

```
setx x1x2 10*13
```

Use macros

```
summarize x1
local meanx1 = r(mean)
summarize x2
local meanx2 = r(mean)
setx x1x2 `meanx1’*`meanx2’
```

What if I want a different confidence level?

Use the level(#) option in simqi

```
simqi, level(90)
```

Use the sumqi command

```
simqi, genpr(myvar)
replace myvar = sqrt(myvar)
sumqi myvar, level(90)
```

Why does Clarify give slightly different results every time?

- It uses random numbers
- You can check the precision of your results

  Rerun the analysis and see if anything of importance changes.

- You can increase the precision of your results

  Simply increase the number of simulations and take a coffee break!