Regression modeling using Stata for continuous, binary, and count outcomes

Chris Cheng, Ph.D.
StataCorp LLC

College Station, TX, USA
June 4th, 2024

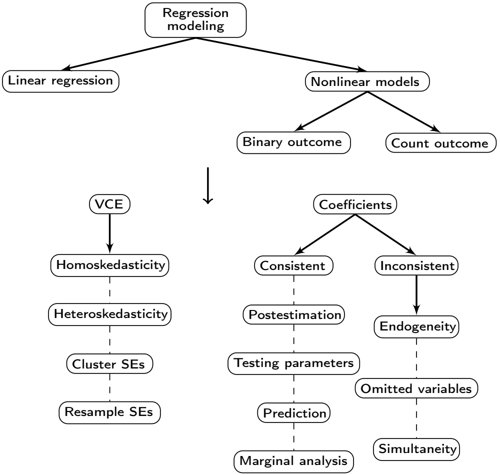
Outline

1. Basic concepts
   - Incentives
   - Stata tools
   - Data structure
   - Modeling intention
2. Linear regression
   - Properties of estimators
   - VCE estimates
   - Modeling consideration
   - Marginal analysis
3. Nonlinear model
   - Binary outcome
   - Count outcome
4. Instrumental estimation
   - Endogeneity
   - 2SLS
   - Diagnosis

Basic concepts

- Motivations: quantitative analysis is based on our conceptualization of an object of interest whose full characterization is unknown
- Conditional quantities: mean wage, probability of have a disease, number of counts
- The way of testing and exploring the concepts is through statistics
- Populative vs sample datasets

Road map


Regression modeling
Linear regression
Nonlinear models
Binary outcome
Count outcome
VCE
Homoskedasticity
Heteroskedasticity
Cluster SEs
Resample SEs
Coefficients
Consistent
Postestimation
Testing parameters
Prediction
Marginal analysis
Inconsistent
Endogeneity
Omitted variables
Simultaneity

Cross-sectional data

- A random sample of units from a population taken at a moment in time
- Sample observations are independently and identically distributed
- Example: Survey of households over a given year

Other data types

- Repeated measures/panel data/longitudinal data datasets – see **help xtset**
- Time-series datasets – see **help tsset**
- Survival time datasets – see **help stset**
- Datasets arising from complex survey designs (called survey datasets) – see **help svyset**

Linear regression

The linear relationship

- Question: What determines babies' birthweights?
- Assuming a linear relationship

$$bwt_i = \beta_0 + \beta_1 age_i + \beta_2 race_i + \beta_3 smoke_i + \varepsilon_i$$

How it looks

```
. webuse lbw
(Hosmer & Lemeshow data)
. list bwt age race smoke in 1/20, noobs sep(0)
```

bwt	age	race	smoke
2523	19	Black	Nonsmoker
2551	33	Other	Nonsmoker
2557	20	White	Smoker
2594	21	White	Smoker
2600	18	White	Smoker
2622	21	Other	Nonsmoker
2637	22	White	Nonsmoker
2637	17	Other	Nonsmoker
2663	29	White	Smoker
2665	26	White	Smoker
2722	19	Other	Nonsmoker
2733	19	Other	Nonsmoker
2750	22	Other	Nonsmoker
2750	30	Other	Nonsmoker
2769	18	White	Smoker
2769	18	White	Smoker
2778	15	Black	Nonsmoker
2782	25	White	Smoker
2807	20	Other	Nonsmoker
2821	28	White	Smoker

White	1
Black	2
Other	3

Nonsmoker	0
Smoker	1

Descriptive statistics

- In Stata 18, we introduced a new command `dtable` to make descriptive statistics easily and nicely, for instance

```
. dtable bwt age i.race i.smoke
```

	Summary
N	189
Birthweight (grams)	2,944.286 (729.016)
Age of mother	23.238 (5.299)
Race	
White	96 (50.8%)
Black	26 (13.8%)
Other	67 (35.4%)
Smoked during pregnancy	
Nonsmoker	115 (60.8%)
Smoker	74 (39.2%)

Note: Tables can be exported to .xlsx, .pdf, .docx, .tex, and more.

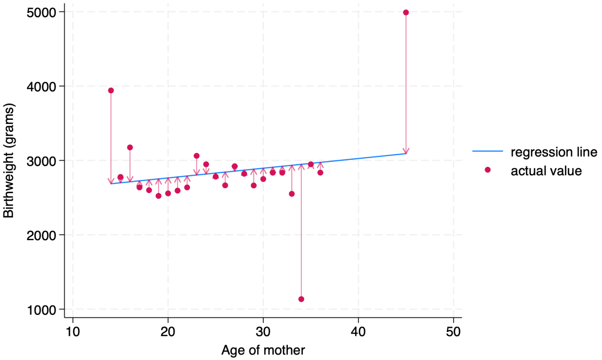
OLS parameters

```
. regress bwt age i.race i.smoke
      Source |       SS           df       MS      Number of obs   =       189
-------------+----------------------------------   F(4, 184)       =      6.50
       Model |  12366825.4         4  3091706.34   Prob > F        =    0.0001
    Residual |  87548473.2       184   475806.92   R-squared       =    0.1238
-------------+----------------------------------   Adj R-squared   =    0.1047
       Total |  99915298.6       188  531464.354   Root MSE        =    689.79

------------------------------------------------------------------------------
         bwt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   1.998899   9.767361     0.20   0.838    -17.27152    21.26932
             |
        race |
      Black  |  -444.6489   156.1404    -2.85   0.005    -752.7047   -136.5931
      Other  |   -449.481   118.9765    -3.78   0.000    -684.2147   -214.7474
             |
       smoke |
     Smoker  |  -425.5563   109.9505    -3.87   0.000    -642.4822   -208.6304
       _cons |   3284.964   260.5749    12.61   0.000     2770.865    3799.062
------------------------------------------------------------------------------
. estimates store linear
```

- Add the base option or turn on base level for all estimation:

```
. set showbaselevels on, perm
(set showbaselevels preference recorded)
```

Graph of regression



Note: graph produced by `twoway`, `lfit`, `scatter`, and `pcarrowi`.

Heteroskedastic regression

- Question: What if the assumption of homoskedasticity is violated?

```
. regress bwt age i.race i.smoke, base vce(robust)
Linear regression                               Number of obs     =        189
                                                F(4, 184)         =       5.92
                                                Prob > F          =     0.0002
                                                R-squared         =     0.1238
                                                Root MSE          =     689.79

------------------------------------------------------------------------------
             |               Robust
         bwt | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   1.998899   11.41526     0.18   0.861    -20.52274    24.52053
             |
        race |
      White  |          0  (base)
      Black  |  -444.6489   146.8476    -3.03   0.003    -734.3704   -154.9274
      Other  |   -449.481   128.4989    -3.50   0.001    -703.0016   -195.9604
             |
       smoke |
  Nonsmoker  |          0  (base)
     Smoker  |  -425.5563   112.6523    -3.78   0.000    -647.8126   -203.3001
             |
       _cons |   3284.964   293.1682    11.21   0.000      2706.56    3863.367
------------------------------------------------------------------------------

. estimates store robust
```

Heteroskedastic regression

- If your data natually come from clusters, we can use the `vce(cluster clustervar)` option to allow intra-group correlation (within clusters).
- Starting in Stata 18, `vce(cluster clustvarlist)` is supported for `regress`, `areg`, and `xtreg, fe` allowing multiway clustering.

Resampling

- Bootstrap: resampling with replacement; random process
- Jackknife: level-one-out resampling

Bootstrap estimates

```
. regress bwt age i.race i.smoke, vce(bootstrap, reps(100) seed(123))
(running regress on estimation sample)
Bootstrap replications (100): .........10.........20.........30.........40.....
> ....50.........60.........70.........80.........90.........100 done
Linear regression                                       Number of obs =      189
                                                        Replications  =      100
                                                        Wald chi2(4)  =    25.47
                                                        Prob > chi2   =   0.0000
                                                        R-squared     =   0.1238
                                                        Adj R-squared =   0.1047
                                                        Root MSE      = 689.7876

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
         bwt | coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |   1.998899   10.76285     0.19   0.853    -19.09589    23.09369
             |
        race |
       White |          0  (base)
       Black |  -444.6489   151.9961    -2.93   0.003    -742.5558    -146.742
       Other |   -449.481   128.4691    -3.50   0.000    -701.2758   -197.6863
             |
       smoke |
   Nonsmoker |          0  (base)
      Smoker |  -425.5563   117.5088    -3.62   0.000    -655.8694   -195.2433
             |
       _cons |   3284.964   288.5131    11.39   0.000     2719.488    3850.439
------------------------------------------------------------------------------
. estimates store boot
```

Comparing standard errors

```
. etable, estimates(linear robust boot)
```

	bwt	bwt	bwt
Age of mother	1.999	1.999	1.999
	(9.767)	(11.415)	(10.763)
Race			
Black	-444.649	-444.649	-444.649
	(156.140)	(146.848)	(151.996)
Other	-449.481	-449.481	-449.481
	(118.977)	(128.499)	(128.469)
Smoked during pregnancy			
Smoker	-425.556	-425.556	-425.556
	(109.951)	(112.652)	(117.509)
Intercept	3284.964	3284.964	3284.964
	(260.575)	(293.168)	(288.513)
Number of observations	189	189	189

Note: `etable` is introduced in Stata 17

Questions about my model

- I would like to know the effect of an explanatory variable on the dependent variable
- I would like to know the elasticity of the dependent variable with respect to a particular explanatory variable
- I would like to test different variables and functional forms for my model
- I would like to use my estimates to test a particular hypothesis
- Which model is the best?

Introducing interactions

```
. regress bwt c.age##c.age i.race##i.smoke, base
      Source |       SS           df       MS      Number of obs   =       189
-------------+----------------------------------   F(7, 181)       =      5.27
       Model |  16926486.9         7  2418069.55   Prob > F        =    0.0000
    Residual |  82988811.7       181  458501.722   R-squared       =    0.1694
-------------+----------------------------------   Adj R-squared   =    0.1373
       Total |  99915298.6       188  531464.354   Root MSE        =    677.13

------------------------------------------------------------------------------
         bwt | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -145.2719   62.92151    -2.31   0.022    -269.4259   -21.11789
             |
 c.age#c.age |   2.870236   1.238137     2.32   0.022     .4271967    5.313274
             |
        race |
       White |          0  (base)
       Black |  -594.6642   206.6937    -2.88   0.004    -1002.503    -186.825
       Other |  -592.2127   142.1154    -4.17   0.000    -872.6286   -311.7967
             |
       smoke |
   Nonsmoker |          0  (base)
      Smoker |  -584.2795   142.5358    -4.10   0.000     -865.525   -303.0339
             |
  race#smoke |
Black#Smoker |   261.5206    314.689     0.83   0.407    -359.4102    882.4514
Other#Smoker |   516.6908   258.9457     2.00   0.048     5.750271    1027.631
             |
       _cons |   5163.883   785.1612     6.58   0.000     3614.637    6713.129
------------------------------------------------------------------------------
```

Testing parameters

- Testing individual parameters

```
. test 2.race == 3.race
( 1)  2.race - 3.race = 0
      F(  1,   181) =    0.00
           Prob > F =    0.9900
```

- Testing interaction terms

```
. testparm race#smoke
( 1)  2.race#1.smoke = 0
( 2)  3.race#1.smoke = 0
      F(  2,   181) =    2.04
           Prob > F =    0.1331
```

Counterfactuals

```
. margins, at(smoke = 0)
Predictive margins                                     Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
At: smoke = 0
```

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
_cons	3126.408	65.78038	47.53	0.000	2996.613	3256.203

```
. margins, at(smoke = 0 race = 1)
Predictive margins                                     Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
At: race  = 1
    smoke = 0
```

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
_cons	3418.152	105.8946	32.28	0.000	3209.205	3627.099

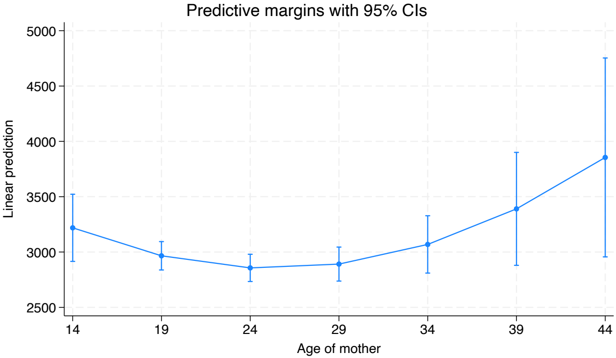
Counterfactuals - across a range

```
. margins, at(age=(14(5)45))
Predictive margins                                      Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
1._at: age = 14
2._at: age = 19
3._at: age = 24
4._at: age = 29
5._at: age = 34
6._at: age = 39
7._at: age = 44
```

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
_at						
1	3218.776	153.7864	20.93	0.000	2915.331	3522.221
2	2966.005	65.03792	45.60	0.000	2837.675	3094.335
3	2856.746	62.27762	45.87	0.000	2733.863	2979.63
4	2890.999	77.87498	37.12	0.000	2737.34	3044.659
5	3068.764	131.3441	23.36	0.000	2809.601	3327.926
6	3390.04	258.6442	13.11	0.000	2879.695	3900.386
7	3854.828	455.4711	8.46	0.000	2956.112	4753.544

Visualizing the quadratic relationship

```
. marginsplot
Variables that uniquely identify margins: age
```



Combination of parameters

- We'd like to estimate the age at which the babies are the lightest, on average

  Given $y = ax^2 + bx + c$, where $a > 0$, the minimum is at $x = -b/2a$

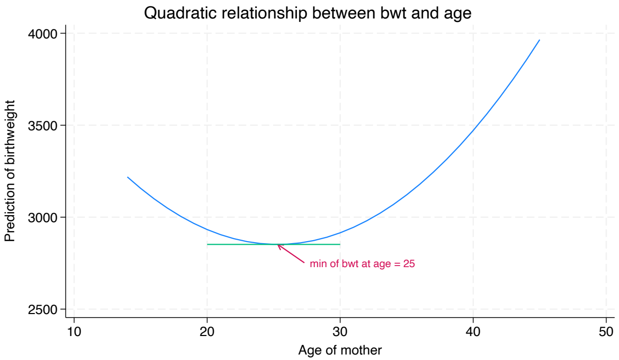
- This is a nonlinear combination of coefficients, so we use `nlcom`

```
. nlcom -_b[age]/(2*_b[age#age])
      _nl_1: -_b[age]/(2*_b[age#age])
```

bwt	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
_nl_1	25.30662	1.723374	14.68	0.000	21.92887	28.68437

- For linear combinations, see `lincom`

Visualizing the quadratic relationship - more



Note: graph produced by `marginsplot`, `pci`, and `pcarrowi`.

Average marginal effects (AMEs)

- In our model:

$$bwt_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2 + ... + \varepsilon_i$$

For a continuous variable (age):

$$\frac{\partial bwt}{\partial age} = \beta_1 + 2 * \beta_2 * age_i$$

```
. margins, dydx(age)
Average marginal effects                              Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
dy/dx wrt:  age
```

	dy/dx	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
age	-11.87431	10.89633	-1.09	0.277	-33.37449	9.625873

```
. quietly margins, dydx(age) atmeans
```

- For marginal effects evaluated at sample means, add the `atmeans` option

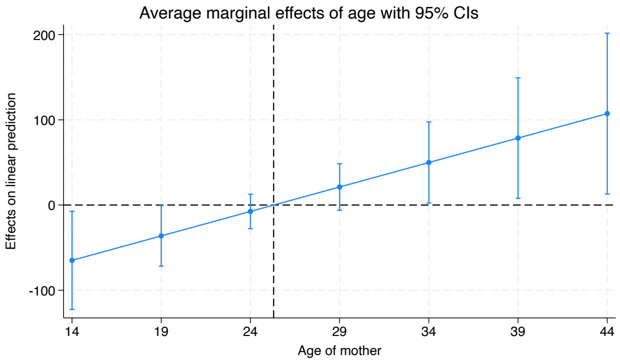
AMEs - across a range

```
. margins, dydx(age) at(age=(14(5)45))
Average marginal effects                              Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
dy/dx wrt:  age
1._at: age = 14
2._at: age = 19
3._at: age = 24
4._at: age = 29
5._at: age = 34
6._at: age = 39
7._at: age = 44
```

	dy/dx	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
age						
_at						
1	-64.90532	29.19551	-2.22	0.027	-122.5127	-7.297996
2	-36.20297	18.03779	-2.01	0.046	-71.79436	-.6115755
3	-7.500615	10.24415	-0.73	0.465	-27.71393	12.7127
4	21.20174	13.82461	1.53	0.127	-6.07639	48.47987
5	49.9041	24.1639	2.07	0.040	2.224933	97.58326
6	78.60645	35.82268	2.19	0.029	7.922673	149.2902
7	107.3088	47.84592	2.24	0.026	12.9013	201.7163

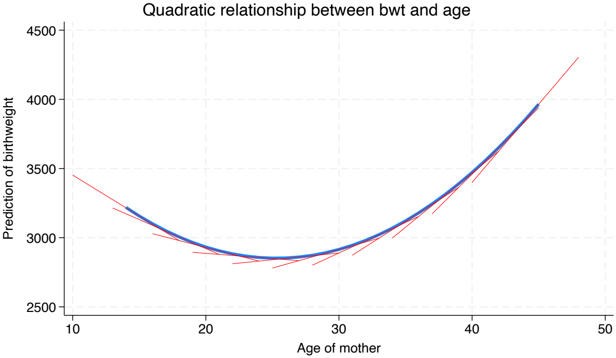
Visualizing the marginal effects

```
. marginsplot, yline(0) xline(25.3)
Variables that uniquely identify margins: age
```



Visualizing the marginal effects

- The previous marginal effects indicate this



Note: graph produced by `marginsplot`, `twoway function`.

Counterfactuals - interaction

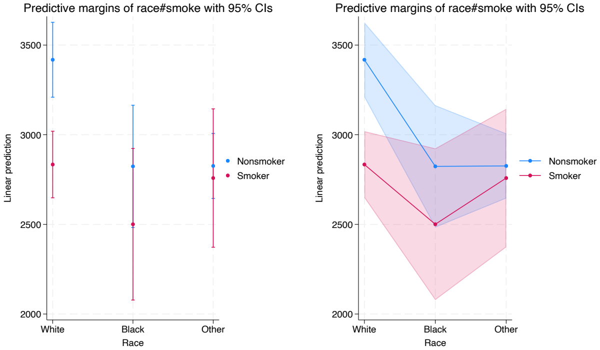
- We can investigate interaction effect (moderation) using `margins`

```
. margins race#smoke
Predictive margins                                    Number of obs = 189
Model VCE: OLS
Expression: Linear prediction, predict()
```

	Margin	Delta-method std. err.	t	P>|t|	[95% conf.	interval]
race#smoke						
White#Nonsmoker	3418.152	105.8946	32.28	0.000	3209.205	3627.099
White#Smoker	2833.872	94.03684	30.14	0.000	2648.323	3019.422
Black#Nonsmoker	2823.488	172.7865	16.34	0.000	2482.553	3164.422
Black#Smoker	2500.729	214.3039	11.67	0.000	2077.874	2923.584
Other#Nonsmoker	2825.939	91.87685	30.76	0.000	2644.652	3007.227
Other#Smoker	2758.35	195.6079	14.10	0.000	2372.385	3144.316

Visualizing the interaction effect

```
 . marginsplot, recast(scatter) name(m1, replace)
Variables that uniquely identify margins: race smoke
. marginsplot, recastci(rarea) ciopts(fcolor(%20) acolor(%20)) name(m2, replace)
Variables that uniquely identify margins: race smoke
. graph combine m1 m2, iscale(0.7)
```



Nonlinear models

Maximum likelihood

The linear model we studied was characterized by the following:

- A linear form for the relationship between the regressors and the dependent variable
- Assumptions about the conditional expectation and the conditional variance
- A minimization of the mean squared error

Maximum likelihood models

- The relationships between the dependent variable and the explanatory variables can be linear but usually are highly nonlinear (exponential family)
- Assumptions are made about the densities of the unknown random disturbance
- The solution is a maximization of the "likelihood" that the data fit your distributional assumptions

Probit and logit models

- Probit and logit models are models for conditional probabilities
- Conditional expectation model may violate some of the conditions that a probability should have (values outside [0,1])
- Assumptions are made over the entire distribution and not only the first two moments

Probit and logit models

- By construction, $P(y_i = 1|x_i) = F(x_i'\beta + \varepsilon)$
- We make an assumption on the distribution of ε, f_ε
  1. If $F(.)$ is the standard normal distribution, we have a **probit**
  2. If $F(.)$ is the logistic distribution, we have a **logit** model

Binary outcome - nonlinear relationship

- Question: What determines the probability of having a low-birthweight baby?

- Assuming a standard normal dist. for the error term

$$Pr(low_i = 1|age_i, race_i, ...) = \Phi(\beta_0 + \beta_1 age_i + \beta_2 race_i + ...)$$

How it looks

```
. list bwt low lwt age race smoke in 120/140, noobs sep(0)
```

bwt	low	lwt	age	race	smoke
3997	0	95	16	Other	Nonsmoker
3997	0	158	20	White	Nonsmoker
4054	0	160	26	Other	Nonsmoker
4054	0	115	21	White	Nonsmoker
4111	0	129	22	White	Nonsmoker
4153	0	130	25	White	Nonsmoker
4167	0	120	31	White	Nonsmoker
4174	0	170	35	White	Nonsmoker
4238	0	120	19	White	Smoker
4593	0	116	24	White	Nonsmoker
4990	0	123	45	White	Nonsmoker
709	1	120	28	Other	Smoker
1021	1	130	29	White	Nonsmoker
1135	1	187	34	Black	Smoker
1330	1	105	25	Other	Nonsmoker
1474	1	85	25	Other	Nonsmoker
1588	1	150	27	Other	Nonsmoker
1588	1	97	23	Other	Nonsmoker
1701	1	128	24	Black	Nonsmoker
1729	1	132	24	Other	Nonsmoker
1790	1	165	21	White	Smoker

White	1
Black	2
Other	3

Nonsmoker	0
Smoker	1

Fitting the model

```
 . probit low age i.race##c.lwt i.smoke, base
Iteration 0:  Log likelihood =   -117.336
Iteration 1:  Log likelihood =  -106.4612
Iteration 2:  Log likelihood = -106.38868
Iteration 3:  Log likelihood = -106.38866
Iteration 4:  Log likelihood = -106.38866
Probit regression                                       Number of obs =    189
                                                        LR chi2(7)    =  21.89
                                                        Prob > chi2   = 0.0026
Log likelihood = -106.38866                             Pseudo R2     = 0.0933
------------------------------------------------------------------------------
         low | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
         age |  -.0130096    .020609    -0.63   0.528    -.0534026    .0273833
             |
        race |
       White |          0  (base)
       Black |     .36389   1.256845     0.29   0.772    -2.099481    2.827261
       Other |   1.592299   1.197156     1.33   0.183    -.7540831    3.938681
             |
         lwt |  -.0061773   .0054959    -1.12   0.261    -.0169491    .0045945
             |
  race#c.lwt |
       Black |    .002564   .0087645     0.29   0.770     -.014614     .019742
       Other |  -.0084805   .0096368    -0.88   0.379    -.0273683    .0104072
             |
       smoke |
   Nonsmoker |          0  (base)
      Smoker |   .6577589   .2269254     2.90   0.004     .2129933    1.102524
             |
       _cons |  -.0090972   .8656977    -0.01   0.992    -1.705833    1.687639
------------------------------------------------------------------------------
```

Marginal effects - AMEs vs at sample means

- If you do not specify the option `atmeans`, you are getting the average marginal effect. $E(g(x)) \neq g(E(x))$ when g is not a linear function
- For a change in x_{ik} this is equal to

$$\frac{1}{N}\sum_{i=1}^{N} f\left(x_i'\beta\right)\beta_k$$

- Before, we were getting the effect for the average person
- If we do not specify `atmeans` we are getting the average effect over the sample

Marginal effect of age - nonlinear

$$\text{marginal (probability) effect} = \frac{dP(y_i = 1|x_i)}{dx_{ik}} = \beta_k * \phi\left(x_i'\beta\right)$$

```
. margins, dydx(age)
Average marginal effects                              Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
dy/dx wrt:  age
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
age	-.0041374	.0065331	-0.63	0.527	-.016942	.0086672

```
. margins, dydx(age) atmeans
Conditional marginal effects                          Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
dy/dx wrt:  age
At: age     =  23.2381 (mean)
    1.race  = .5079365 (mean)
    2.race  = .1375661 (mean)
    3.race  = .3544974 (mean)
    lwt     = 129.8201 (mean)
    0.smoke = .6084656 (mean)
    1.smoke = .3915344 (mean)
```

	dy/dx	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
age	-.0043718	.0069187	-0.63	0.527	-.0179322	.0091886

Counterfactuals and contrast

```
. margins smoke, at(age = 25 race = 1)
Predictive margins                                    Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
At: age  = 25
    race =  1
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
smoke						
Nonsmoker	.1319512	.0453406	2.91	0.004	.0430853	.2208172
Smoker	.3194439	.0589322	5.42	0.000	.2039388	.4349489

```
. margins r.smoke, at(age = 25 race = 1)
Contrasts of predictive margins                       Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
At: age  = 25
    race =  1
```

	df	chi2	P>chi2
smoke	1	8.99	0.0027

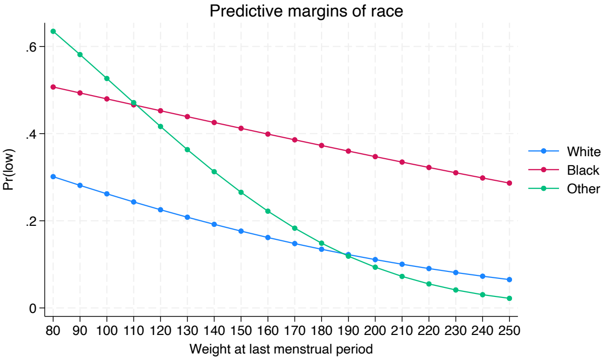
	Contrast	Delta-method std. err.	[95% conf.	interval]
smoke				
(Smoker vs Nonsmoker)	.1874927	.062535	.0649263	.310059

Contrast - two different at()

```
. margins, at(age=generate(age)) at(age=generate(age+4))
Predictive margins                                       Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
1._at: age =   age
2._at: age = age+4

-------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+-----------------------------------------------------------------
         _at |
          1  |   .3112646   .0316713     9.83   0.000     .2491901    .3733392
          2  |   .2949081   .0400258     7.37   0.000     .2164589    .3733573
-------------------------------------------------------------------------------

. margins, at(age=generate(age)) at(age=generate(age+4)) contrast(atcontrast(r))
Contrasts of predictive margins                          Number of obs = 189
Model VCE: OIM
Expression: Pr(low), predict()
1._at: age =   age
2._at: age = age+4

------------------------------------------------
             |         df        chi2     P>chi2
-------------+----------------------------------
         _at |          1        0.41     0.5215
------------------------------------------------

--------------------------------------------------------------
             |            Delta-method
             |   Contrast   std. err.     [95% conf. interval]
-------------+------------------------------------------------
         _at |
   (2 vs 1)  |  -.0163565   .0255186     -.066372     .0336589
--------------------------------------------------------------
```

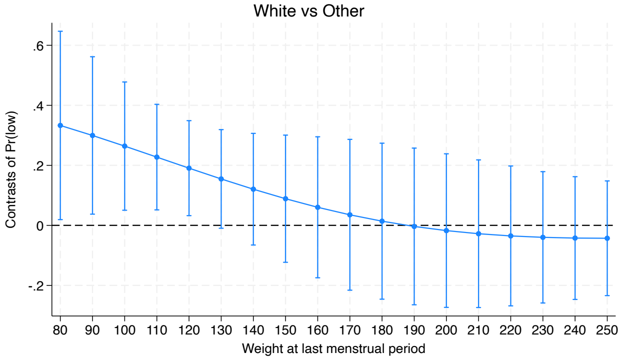
Visualizing the interaction effect

```
. quietly margins race, at(lwt=(80(10)250))
. marginsplot, noci
Variables that uniquely identify margins: lwt race
```



Visualizing the contrast

```
. quietly margins r.race if race == 1 | race == 3, at(lwt=(80(10)250))
. marginsplot, yline(0) title("White vs Other")
Variables that uniquely identify margins: lwt
```



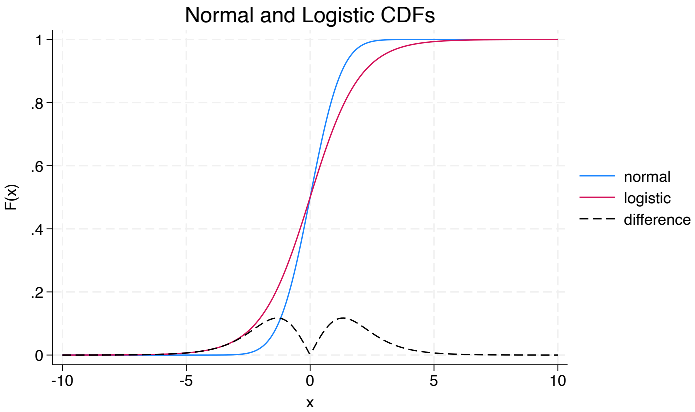
Logistic regression for the same outcome

```
. logit low age i.race##c.lwt i.smoke, base
Iteration 0:  Log likelihood =   -117.336
Iteration 1:  Log likelihood = -106.86243
Iteration 2:  Log likelihood = -106.60413
Iteration 3:  Log likelihood = -106.60373
Iteration 4:  Log likelihood = -106.60373
Logistic regression                                     Number of obs =    189
                                                        LR chi2(7)    =  21.46
                                                        Prob > chi2   = 0.0031
Log likelihood = -106.60373                             Pseudo R2     = 0.0915
```

low	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
age	-.0199142	.0342929	-0.58	0.561	-.0871271	.0472987
race						
White	0	(base)				
Black	.5769514	2.069831	0.28	0.780	-3.479843	4.633746
Other	2.783524	2.099436	1.33	0.185	-1.331294	6.898342
lwt	-.0100061	.0094562	-1.06	0.290	-.0285399	.0085277
race#c.lwt						
Black	.0043309	.0145609	0.30	0.766	-.0242078	.0328697
Other	-.0154112	.0170738	-0.90	0.367	-.0488753	.0180528
smoke						
Nonsmoker	0	(base)				
Smoker	1.076494	.3860288	2.79	0.005	.3198915	1.833097
_cons	-.0598411	1.482944	-0.04	0.968	-2.966358	2.846676

```
. quietly logistic low age i.race##c.lwt i.smoke, base
```

Graphical explanation: probit vs logit



Count outcomes

- Maximum likelihood assumes we know the entire distribution of the unobservables
  - Poisson or negative binomial regressions
- If our distribution is misspecified, we can still obtain consistent marginal effects under certain conditions
- An example is an exponential mean model using a Poisson model. Our model for the mean is correct, but the standard errors from the Poisson distribution are incorrect.

How it looks

- Question: What determines mortality rate?

```
. webuse dollhill3, clear
(Doll and Hill (1966))
. list deaths smoke agecat pyears, noobs sep(0)
```

deaths	smokes	agecat	pyears
32	1	35-44	52,407
104	1	45-54	43,248
206	1	55-64	28,612
186	1	65-74	12,663
102	1	75-84	5,317
2	0	35-44	18,790
12	0	45-54	10,673
28	0	55-64	5,710
28	0	65-74	2,585
31	0	75-84	1,462

Note: "pyears": person years, used as the exposure

Estimation

```
. poisson deaths smokes i.agecat, exposure(pyears) vce(robust)
Iteration 0:   Log pseudolikelihood = -33.823284
Iteration 1:   Log pseudolikelihood = -33.600471
Iteration 2:   Log pseudolikelihood = -33.600153
Iteration 3:   Log pseudolikelihood = -33.600153
Poisson regression                                      Number of obs =      10
                                                        Wald chi2(5)  = 6380.53
                                                        Prob > chi2   =  0.0000
Log pseudolikelihood = -33.600153                       Pseudo R2     =  0.9321

------------------------------------------------------------------------------
             |               Robust
      deaths | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      smokes |   .3545356    .123158     2.88   0.004     .1131504    .5959209
             |
      agecat |
      35-44  |          0  (base)
      45-54  |   1.484007   .2211923     6.71   0.000     1.050478    1.917536
      55-64  |   2.627505   .2102283    12.50   0.000     2.215465    3.039545
      65-74  |   3.350493   .2104029    15.92   0.000     2.938111    3.762875
      75-84  |   3.700096   .2372667    15.59   0.000     3.235062    4.165131
             |
       _cons |  -7.919326   .2509888   -31.55   0.000    -8.411255   -7.427397
  ln(pyears) |          1  (exposure)
------------------------------------------------------------------------------
```

margins after Poisson

- After `poisson`, `margins` can be used to predict the following:
  - `n` number of events; the default
  - `ir` incidence rate, exp(xb), n when the exposure variable = 1
  - `pr(n)` probability that $y = n$
  - `pr(a,b)` probability that $a \leq y \leq b$
  - `xb` the linear prediction

Counterfactuals

- Predicted probability that deaths = 5

```
. margins, predict(pr(5))
Predictive margins                                      Number of obs = 10
Model VCE: Robust
Expression: Pr(deaths=5), predict(pr(5))

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _cons |   .0134236   .0061924     2.17   0.030     .0012867    .0255605
------------------------------------------------------------------------------
```

- Predicted number of deaths across age categories

```
. margins agecat, predict(n)
Predictive margins                                      Number of obs = 10
Model VCE: Robust
Expression: Predicted number of events, predict(n)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      agecat |
      35-44  |   8.800113   1.856679     4.74   0.000      5.16109    12.43914
      45-54  |   38.81363   2.571338    15.09   0.000      33.7739    43.85336
      55-64  |   121.7865   .8360965   145.66   0.000     120.1478    123.4252
      65-74  |   250.9509   2.437698   102.95   0.000     246.1731    255.7287
      75-84  |   355.9752   37.88741     9.40   0.000     281.7172    430.2332
------------------------------------------------------------------------------
```

Instrumental estimation

Endogeneity

- When we fit our linear model we assumed that $E(\varepsilon|X) = 0$. This implies that the random disturbance does not affect the model.
- Endogeneity is the violation of this assumption. Mathematically:

$$E\left(X\varepsilon\right) = 0$$

- The regressors are related in some regard to the random disturbance:
  1. Omitted variables: unobserved confounding factors
  2. Simultaneity (original)

Two-stage least squares (2SLS)

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i \\
E\left(x_{2i}\varepsilon_i\right) &\neq 0 \\
E\left(x_{1i}\varepsilon_i\right) &= 0
\end{aligned}
$$

- The solution is to think about x_2 as having a component that is related to ε and a component that is unrelated
- The exogenous parts are referred to as instruments
- Instruments have an indirect effect on the dependent variable

2SLS solution

$$
\begin{aligned}
E(z_{2i}\varepsilon_i) &= 0 \\
E(\varepsilon_i \nu_i) &\neq 0 \\
x_{2i} &= \pi_1 x_{1i} + \pi_2 z_{2i} + \nu_i
\end{aligned}
$$

- $\hat{x}_2$ from a regression of x_2 on x_1 and z_2 is a function of exogenous components
- A regression of y on $\hat{x}_2$ and x_1 satisfies our regression assumptions

Properties of good instruments

- They should satisfy exogeneity $E(z_2'\varepsilon) = 0$ and indirect effect (exclusion restriction)
- $Cov(z_2, x_2) \neq 0$ A violation of this is known as weak instruments
- Stata has tests for validity of instruments and for weak instruments

How it looks

- What determines wage level? Is it related to job tenure? What if job tenure is endogenous?

```
. webuse nlswork, clear
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)
. describe ln_wage age race tenure union msp occ_code
Variable      Storage   Display    Value
    name         type    format    label      Variable label
-------------------------------------------------------------------------------------------------
ln_wage         float   %9.0g                 ln(wage/GNP deflator)
age             byte    %8.0g                 Age in current year
race            byte    %8.0g      racelbl    Race
tenure          float   %9.0g                 Job tenure, in years
union           byte    %8.0g                 1 if union
msp             byte    %8.0g                 1 if married, spouse present
occ_code        byte    %8.0g                 Occupation
```

Estimation

- Syntax

  `ivregress` *estimator depvar exogenous (endogenous = instruments)*

- The estimators are 2sls, liml, gmm

2sls and show first stage

```
. ivregress 2sls ln_wage c.age##c.age i.race (tenure = union msp occ_code)
Instrumental variables 2SLS regression            Number of obs     =     18,927
                                                  Wald chi2(5)      =     961.26
                                                  Prob > chi2       =     0.0000
                                                  R-squared         =          .
                                                  Root MSE          =     .71794

------------------------------------------------------------------------------
     ln_wage | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      tenure |   .1861536   .0081384    22.87   0.000     .1702027    .2021045
         age |   .0233494   .0079369     2.94   0.003     .0077933    .0389055
             |
 c.age#c.age |  -.0008669   .0001244    -6.97   0.000    -.0011108    -.000623
             |
        race |
       White |          0  (base)
       Black |  -.2043076    .011858   -17.23   0.000    -.2275489   -.1810663
       Other |   .1645503   .0501795     3.28   0.001     .0662004    .2629003
             |
       _cons |     1.2278   .1208876    10.16   0.000     .9908649    1.464736
------------------------------------------------------------------------------
Endogenous: tenure
Exogenous:  age c.age#c.age 2.race 3.race union msp occ_code

. ivregress 2sls ln_wage c.age##c.age i.race (tenure = union msp occ_code), first
(output omitted)
```

Diagnosis

- Instrument weakness

```
. estat firststage
First-stage regression summary statistics
--------------------------------------------------------------------------
             |            Adjusted      Partial
    Variable |   R-sq.       R-sq.        R-sq.    F(3,18919)   Prob > F
-------------+------------------------------------------------------------
      tenure |  0.1470      0.1467       0.0275       178.646    0.0000
--------------------------------------------------------------------------

Minimum eigenvalue statistic = 178.646
Critical Values                      # of endogenous regressors:    1
H0: Instruments are weak             # of excluded instruments:     3
---------------------------------------------------------------------
                                   |    5%     10%     20%     30%
2SLS relative bias                 |  13.91    9.08    6.46    5.39
-----------------------------------+---------------------------------
                                   |   10%     15%     20%     25%
2SLS size of nominal 5% Wald test  |  22.30   12.83    9.54    7.80
LIML size of nominal 5% Wald test  |   6.46    4.36    3.69    3.32
---------------------------------------------------------------------
```

Diagnosis

- Weak-instrument-robust tests (new in StataNow)

```
 . estat weakrobust
Weak-instrument-robust test
Model VCE: Unadjusted
 ( 1)  tenure = 0
Cond. likelihood ratio (CLR) test = 1807.26
                       Prob > CLR =  0.0000
Note: CLR test reported by default because
      model is overidentified.
```

just-identified models	Anderson Rubin (1949) test
overidentified models & unadjusted VCE	conditional likelihood-ratio (CLR) test (Moreira 2003)
overidentified models & robust VCE	generalized CLR test (Finlay and Magnusson 2009)

https://www.stata.com/statanow/
https://www.stata.com/statanow/inference-robust-to-weak-instruments/

Diagnosis

- Endogeneity

```
. estat endogenous
  Tests of endogeneity
  H0: Variables are exogenous
  Durbin (score) chi2(1)          =   956.15  (p = 0.0000)
  Wu-Hausman F(1,18920)           =  1006.65  (p = 0.0000)


. estat overid
  Tests of overidentifying restrictions:
  Sargan (score) chi2(2) =  185.731  (p = 0.0000)
  Basmann chi2(2)        =  187.492  (p = 0.0000)
```

Summary

1. Basic concepts
2. Linear regression
   - Properties of estimators: `regress, vce()`
   - Marginal analysis: `margins, at()/dydx()`
3. Nonlinear models
   - Binary outcome: `probit` and `logit/logistic`
   - Count outcome: `poisson/nbreg`
4. Instrumental estimation: `ivregress`

Reference

1. Aldrich, J. H., and F. D. Nelson. 1984. *Linear Probability, Logit, and Probit Models*. Newbury Park, CA: Sage.
2. Anderson, T. W., and H. Rubin. 1949. Estimation of the parameters of a single equation in a complete system of stochastic equations. *Annals of Mathematical Statistics* 20: 46–63
3. Basmann, R. L. 1960. On finite sample distributions of generalized classical linear identifiability test statistics. *Journal of the American Statistical Association* 55: 650–659.
4. Cameron, A. C., and P. K. Trivedi. 2010. *Microeconometrics Using Stata*. Rev. ed. College Station, TX: Stata Press.
5. Durbin, J. 1954. Errors in variables. *Review of the International Statistical Institute* 22: 23–32.
6. Finlay, K., and L. M. Magnusson. 2009. Implementing weak-instrument robust tests for a general class of instrumentalvariables models. *Stata Journal* 9: 398–421.
7. Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.
8. Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.
9. Moreira, M. J. 2003. A conditional likelihood ratio test for structural models. *Econometrica* 71: 1027–1048.
10. Sargan, J. D. 1958. The estimation of economic relationships using instrumental variables. *Econometrica* 26: 393–415.
11. Stock, J. H., and M. Yogo. 2005. Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, ed. D. W. K. Andrews and J. H. Stock, 80–108. New York: Cambridge University Press.
12. Wu, D.-M. 1974. Alternative tests of independence between stochastic regressors and disturbances: Finite sample results. *Econometrica* 42: 529–546.

Send questions to Tech Support

tech-support@stata.com

Upcoming webinars

https://www.stata.com/training/webinar/
https://www.stata.com/training/webinar/cluster-robust-inference-in-stata/

STATA® 18