Title

stata.com

Example 7 — Random-effects regression with continuous endogenous covariate

Description Remarks and examples Reference Also see

Description

In this example, we show how to estimate and interpret the results of an extended regression model with a continuous outcome, a continuous endogenous covariate, and random effects.

Remarks and examples

stata.com

We will use `nlswork.dta`, a subsample of the NLSY data (Center for Human Resource Research 1989) on young women aged 14–24 in 1968. These data are panel data; each individual was surveyed in multiple years ranging from 1968 to 1988.

Suppose that we want to study the relationship between the natural logarithm of wage (`ln_wage`) and the number of years at a job (`tenure`). We also model `ln_wage` with a quadratic effect of the individual's age (`age` and `c.age#c.age`), living in a metropolitan area (`not_smsa`), and whether the individual is African American (`2.race`). We suspect that the unobserved factors that influence the individual's job tenure are correlated with the unobserved factors that influence their wage, so we treat job tenure as an endogenous covariate. We use an individual's union status (`union`) and whether she lived in the southern United States (`south`) as instrumental covariates for tenure. Of course, these are not the instruments we would choose in real research, but they are useful for demonstrating how to use the commands below.

We also want to account for the within-panel correlation in our data, so we fit a random-effects model using `xteregress`. Before we can fit our model, we must use `xtset` to specify the panel identifier variable, in this case, `idcode`. Our data have already been `xtset`, so we type `xtset` to display the settings.

```
. use https://www.stata-press.com/data/r18/nlswork
(National Longitudinal Survey of Young Women, 14-24 years old in 1968)

. xtset

Panel variable: idcode (unbalanced)
 Time variable: year, 68 to 88, but with gaps
         Delta: 1 unit
```

We are now ready to fit our model. We want to make inferences about how our covariates affect the log wage in the population, not just in our sample. Therefore, we add the `vce(robust)` option so that subsequent calls to `margins` will consider our sample as a draw from the population.

By default, `xteregress` includes random effects for both `ln_wage` and `tenure` and allows these random effects to be correlated. Because of the complexity of this model, the command may take a few minutes to run.

```
. xteregress ln_wage age c.age#c.age i.not_smsa 2.race,
> endogenous(tenure = age c.age#c.age i.union 2.race i.south) vce(robust)
```

(*iteration log omitted*)

```
Extended linear regression                      Number of obs     = 19,007
Group variable: idcode                          Number of groups =  4,134

                                                Obs per group:
                                                             min =      1
                                                             avg =    4.6
                                                             max =     12

Integration method: mvaghermite                 Integration pts. =      7

                                                Wald chi2(5)      = 384.25
Log pseudolikelihood = -53601.41                Prob > chi2       = 0.0000

                            (Std. err. adjusted for 4,134 clusters in idcode)
```

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
ln_wage						
age	.0161086	.0134428	1.20	0.231	-.0102388	.042456
c.age#c.age	-.0011178	.0002402	-4.65	0.000	-.0015887	-.000647
1.not_smsa	-.172498	.0122743	-14.05	0.000	-.1965552	-.1484408
race						
Black	-.2374388	.0254533	-9.33	0.000	-.2873263	-.1875513
tenure	.2300781	.0277646	8.29	0.000	.1756605	.2844957
_cons	1.690136	.2077606	8.14	0.000	1.282933	2.097339
tenure						
age	.0892847	.0599348	1.49	0.136	-.0281852	.2067547
c.age#c.age	.0033688	.0009943	3.39	0.001	.0014199	.0053176
1.union	.5584566	.0740956	7.54	0.000	.4132318	.7036814
race						
Black	.4691202	.1101411	4.26	0.000	.2532476	.6849929
1.south	-.4024058	.0628545	-6.40	0.000	-.5255983	-.2792132
_cons	-2.929734	.8800349	-3.33	0.001	-4.65457	-1.204897
var(e.ln_w~e)	.3654205	.0786259			.2396866	.5571114
var(e.tenure)	6.656475	.1285168			6.409292	6.913189
corr(e.ten~e, e.ln_wage)	-.9055589	.0213219	-42.47	0.000	-.9395846	-.8538145
var(ln_~e[idc~e])	.3314414	.0736048			.2144748	.5121973
var(ten~e[idc~e])	7.593483	.3027546			7.022688	8.210672
corr(ten~e[idc~e], ln_~e[idc~e])	-.8299334	.0421356	-19.70	0.000	-.8963409	-.7271053

The first two sections of the output provide the estimated coefficients in the equations for `ln_wage` and `tenure`. Because this is a linear regression, we can interpret the coefficients in the usual way. For example, we expect an increase of 0.23 in log wage for an additional year of job tenure.

Next, we see the estimates of the observation-level error variances and their correlation with the dependent variable. This is followed by estimates of the variances of the random effects and an estimate of their correlation with the dependent variable. If at least one of these correlations is significantly different from zero, we can conclude that `tenure` is endogenous. In our case, the correlation between the observation-level errors is -0.91, and the correlation between the random effects is -0.83. Because both are negative and significantly different from zero, we conclude that `tenure` is endogenous and that unobserved individual-level factors that increase job tenure tend to decrease log wage. Additionally, unobserved observation-level (time-varying) factors that increase job tenure tend to also decrease log wage.

We may also want to ask questions about specific groups in the population. Below, we consider how log wages differ by age group. We will study people between 18 and 40. As we mentioned in [ERM] **Intro 7**, the effects that `margins` computes by default have a causal interpretation if the model is correctly specified. The reason they do is that `margins` conditions on the level of endogeneity. We type

```
. margins, over(age) subpop(if (age>=18)*(age<=40)) vce(unconditional)
  (output omitted)
```

and then graph the effects

```
. marginsplot
```

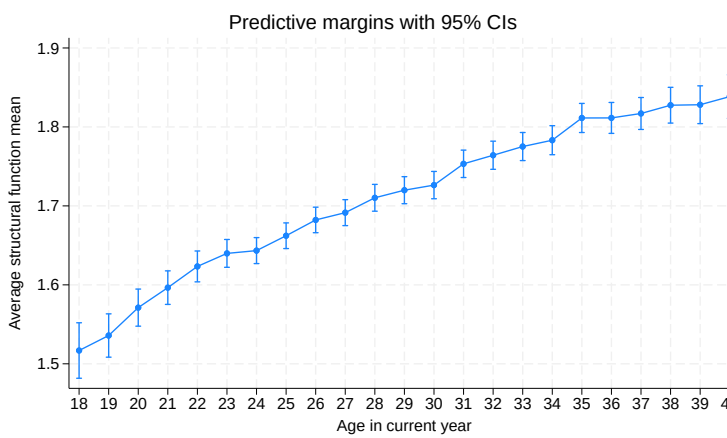

Figure 1.

What if we had not accounted for the level of endogeneity? This will not matter in cases where we average over the entire population and the effect of the unobservable becomes zero. It matters, however, when we look at effects over subpopulations. Below, we use `margins` to compute effects using the linear prediction, $x_{it}\beta$, by adding the option `predict(xb)`. The linear prediction is not conditioning on endogeneity.

```
. margins, over(age) predict(xb) subpop(if (age>=18)*(age<=40)) vce(unconditional)
  (output omitted)
. marginsplot
  (output omitted)
```

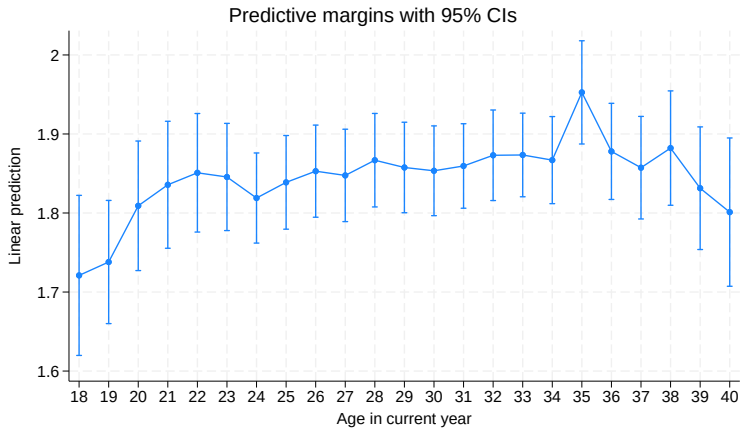

Figure 2.

Figure 1 shows wages increasing for each age group, whereas figure 2 shows wages decreasing after 35 years. Conditioning on the level of endogeneity to obtain structural effects matters even when we have linear models.

Reference

Center for Human Resource Research. 1989. *National Longitudinal Survey of Labor Market Experience, Young Women 14–24 years of age in 1968*. Columbus, OH: Ohio State University Press.

Also see

[ERM] **eregress** — Extended linear regression

[ERM] **eregress postestimation** — Postestimation tools for eregress and xteregress

[ERM] **Intro 3** — Endogenous covariates features

[ERM] **Intro 6** — Panel data and grouped data model features

[ERM] **Intro 9** — Conceptual introduction via worked example