University of Exeter

Applied Econometrics

Topic 3: Linear Regression Analysis

Jingya Zeng

University of Exeter

Table of contents

University of Exeter

Linear regression model refresher

University of Exeter

Linear regression model refresher

Population bivariate linear regression:

$$Y_i = a + bX_i + e_i$$

- **Y_i** the dependent variable, regressand, predicted **or simply the** left-hand variable;
- **X_i** : **the** independent/explanatory variable, regressor, predictor **or simply the** right-hand variable;
- **a+b X_i** : population regression line;
- **a** : the intercept of the population regression line;
- **b** : the slope of the population regression line;
- **e_i** : the (regression) residual.

University of Exeter

Linear regression model refresher

Minimises the mean squared-error

$$(a, b) = \underset{a_0, b_0}{argmin} E(Y_i - a_0 - b_0 X_i)^2$$

Computed from a sample of n observations of X_i and Y_i:

$$Y_i = \hat{a} + \hat{b} X_i + \hat{e}_i$$

- $\hat{a}$: estimated intercept, estimates of the unknown population intercept a
- $\hat{b}$: estimated slope, estimates of the unknown population slope b

University of Exeter

Assumptions for causal inference

Econometric model:

$$Y_i = \alpha + \beta X_i + u_i$$

where u_i is the unobservable error term.

- (1) (X_i, Y_i) $i = 1, ..., n$, are independent and identically distributed (i.i.d.) across observations
- (2) Large outliers are unlikely
- (3) $E(u_i|X_i) = 0$: the conditional distribution of u_i given X_i has a mean of zero.
- with large random sample and assumptions (1) and (2), we can invoke the LLN (Law of Large Numbers) and the CLT (Central Limit Theorem): $\hat{b} = b$
- with assumption (3), $b = \beta$

University of Exeter

Why use the OLS estimator?

- OLS is the dominant method used in practice: ”Speaking the same language” as other economists and statisticians.
- **Easy to use:** OLS formulas are built into virtually all spreadsheet and statistical software packages.
- **Desirable theoretical properties:** unbiased and consistent under assumptions.

University of Exeter

Applied regression analysis

University of Exeter

Data

We use the data file `GreeneF41_3.dta`

- 428 observations
- Married couples' labour market data and characteristics
- All women in the sample are in the labour force (willing and able to work)
- The labour force definition excludes students, pensioners, those unable to work due to disability, those who are not seeking work for other reasons (stay-at-home parent, independent wealth, in between jobs etc)

We want to estimate a model that explains wife's annual earnings as a function of her characteristics.

University of Exeter

Stata regression

Step 1: find the variable to be used in the regression.

Step 2: run the regression - which command to use?

Step 3: interpret the results - are the signs of the coefficients in line with your expectations?

Saved results

ereturn list

```
. ereturn list

scalars:
                  e(N) =  428
               e(df_m) =  5
               e(df_r) =  422
                  e(F) =  36.16472911677187
                 e(r2) =  .299961102900541
               e(rmse) =  3620.294132943347
                e(mss) =  2369970464.856518
                e(rss) =  5530955495.008137
               e(r2_a) =  .2916668031718744
                 e(ll) =  -4111.449382550107
               e(ll_0) =  -4187.765929484392
               e(rank) =  6

macros:
            e(cmdline) : "regress wearn ax ax2 we kids mtr"
              e(title) : "Linear regression"
          e(marginsok) : "XB default"
                e(vce) : "ols"
            e(depvar) : "wearn"
                e(cmd) : "regress"
         e(properties) : "b V"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "regress_estat"

matrices:
                  e(b) :  1 x 6
                  e(V) :  6 x 6
               e(beta) :  1 x 5

functions:
             e(sample)
```

Post-estimation

To obtain fitted values: `predict yhat, xb`

To obtain residuals: `predict ehat, resid`

To obtain the variance-covariance matrix of the estimator: `estat vce`

```
. estat vce

Covariance matrix of coefficients of regress model
```

e(V)	ax	ax2	we	kids	mtr	_cons
ax	5144.8755					
ax2	-145.47356	4.6235938				
we	-205.86439	7.0318469	7136.028			
kids	1010.2063	94.822753	-2046.0452	172837.7		
mtr	9546.1852	-273.95584	88141.225	-92454.45	6332946.9	
_cons	-37393.438	840.73257	-146812.48	-65653.522	-5345364.1	5796554.7

F tests / Wald tests

Test of overall significance: whether a set of independent variables are collectively significant.

Is the wife's experience significant in the regression?

`test ax ax2`

```
. test ax ax2

 ( 1)  ax = 0
 ( 2)  ax2 = 0

       F(  2,    422) =   24.97
            Prob > F =    0.0000
```

Linear combinations of coefficients

What is the predicted earnings for a woman with no children, 10 years of schooling, no job experience and a marginal tax rate of 0.2?

`lincom _cons + 10*we + 0.2*mtr`

```
. lincom _cons + 10*we + 0.2*mtr

 ( 1)  10*we + .2*mtr + _cons = 0
```

wearn	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
(1)	11330.73	1428.861	7.93	0.000	8522.153	14139.3

(There is a another way of doing this with `margins`. We will introduce it in a few weeks)

Non-linear combinations of coefficients

How to compute elasticities?

$$elasticity = \frac{\%\,\Delta\, in\, y}{\%\,\Delta\, in\, x_i} = \frac{dy/y}{dx_i/x_i} = \frac{x}{y}\frac{dy}{dx_i} = \frac{x}{y}\beta_i$$

Example

Simple linear regression example: `regress wearn mtr`

Elasticity of mrt:

e_{mtr} at a marginal tax rate of 0.5 = $\frac{0.5}{\beta_0 + 0.5\beta_1}\beta_1$

Plug in the estimates:

$\hat{e}_{mtr}$ at a marginal tax rate of 0.5 = $\frac{0.5}{\hat{\beta}_0 + 0.5\hat{\beta}_1}\hat{\beta}_1$

University of Exeter

Non-linear combinations of coefficients

How to calculate elasticities?

```
nlcom (_b[mtr]*0.5)/(_b[_cons] + 0.5*_b[mtr])
```

```
. nlcom (_b[mtr]*0.5)/(_b[_cons] + 0.5*_b[mtr])

       _nl_1: (_b[mtr]*0.5)/(_b[_cons] + 0.5*_b[mtr])

------------------------------------------------------------------------------
       wearn | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
       _nl_1 |  -3.692233   1.68397    -2.19   0.028    -6.992753   -.3917121
------------------------------------------------------------------------------
```

Coefficients after regression can be referred to by _b[ ] and standard errors by _se[ ]

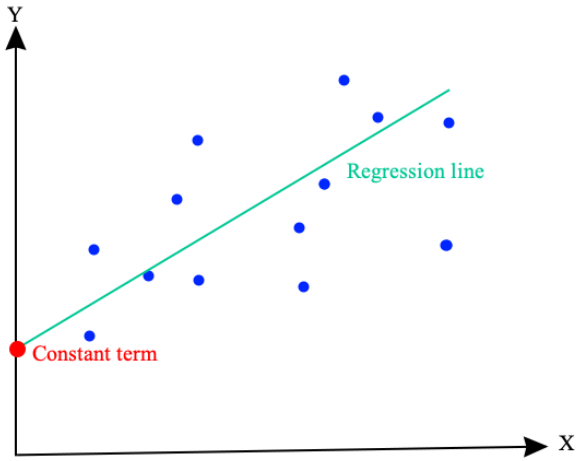
```
display _b[mtr]
```

```
display _se[mtr]
```

Again, you'll also be able to use `margins` to obtain elasticities.

Additional points related to OLS regression

- Constant term: expected mean value of Y when all X=0. Can only be removed if you have a good theoretical explanation for it to be forced to go through the origin (very rare).



- Scaling of variables: are the coefficients comparable? The magnitude of the estimated coefficients relates to the scale of the variables.
- Interpretation of dummy variables, interaction effects and confidence intervals for these (marginal effects). More details to follow.

University of Exeter

Stata tips and tricks

University of Exeter

Working with more than 1 regression

Postestimation commands like `test`, `predict`, `rvfplot`, `estat hettest` etc... always use the most recent estimation results for their calculations.

We often find that we estimate several different models, and want to be able to:

- keep track of all our different results, and switch between them easily;
- produce postestimation statistics for any one of them without having to re-run our models;
- create tables of coefficients comparing the different models.

We can save all our regression results in Stata's memory, and recall them again using the `estimates store` command.
We can save our results to disk using the `estimates save` command.

University of Exeter

Storing estimation results

```
regress wearn ax ax2 we kids mtr
estimates store model1

regress wearn ax we kids mtr
estimates store model2

estimates restore model1
```

Creating tables of estimation results

```
estimates table model1 model2, star stats(N r2)
```

```
. estimates table model1 model2, star stats(N r2)

----------------------------------------
    Variable |    model1        model2
-------------+--------------------------
          ax |  263.46956***   164.77352***
         ax2 | -3.1368616
          we |   252.4203**     257.19103**
        kids | -51.887831       12.444345
         mtr | -20225.378***   -20411.243***
       _cons |  12851.598***    13421.99***
-------------+--------------------------
           N |         428            428
          r2 |   .2999611      .29643073
----------------------------------------
Legend: * p<0.05; ** p<0.01; *** p<0.001
```

University of Exeter

Saving estimation results to disk

`estimates store` only works within the current Stata session. If we close Stata, even if we save the latest version of our data, the regression results will be lost. (We can of course re-run the models.)

Using `estimates save`, we can store results to disk and recall them in a future Stata session, for the purpose of postestimation tests.

After estimation, we can save our results to disk using the `estimates save` command: `estimates save model1`

In a later Stata session, `estimates use model1` recalls the results and allows us to conduct post-estimation tests.

This will work even if we don't have the original data in memory, but any command that relies on the original estimation sample will not work: diagnostic plots, for example.