

## Modelling and Testing for Structural Breaks

The (artificial) data set given in Johnston & DiNardo page 130 contains 15 observations on 2 variables (y and z) 2 periods (one of 5 years, one of 10).

```
list y z d1 z1 d2 z2
```

	y	z	d1	z1	d2	z2
1.	1	2	1	2	0	0
2.	2	4	1	4	0	0
3.	2	6	1	6	0	0
4.	4	10	1	10	0	0
5.	6	13	1	13	0	0
6.	1	2	0	0	1	2
7.	3	4	0	0	1	4
8.	3	6	0	0	1	6
9.	5	8	0	0	1	8
10.	6	10	0	0	1	10
11.	6	12	0	0	1	12
12.	7	14	0	0	1	14
13.	9	16	0	0	1	16
14.	9	18	0	0	1	18
15.	11	20	0	0	1	20

To stack the data make additional 15x1 column vectors,  $z_1$ ,  $z_2$ ,  $d_1$ ,  $d_2$ , where zeros appear in the rows relating to the **other** sample and the non-zero elements correspond to the original z values for the relevant sub-sample.

The unrestricted regression is equivalent to a regression of y on d1 z1 d2 z2

```
. reg y d1 d2 z1 z2, nocons (1)
```

Source	SS	df	MS	Number of obs = 15		
Model	505.839773	4	126.459943	F( 4, 11)	=	440.18
Residual	3.16022727	11	.287293388	Prob > F	=	0.0000
-----				R-squared	=	0.9938
-----				Adj R-squared	=	0.9915
Total	509.00	15	33.9333333	Root MSE	=	.536

  

	y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
d1		-.0625	.4831417	-0.13	0.899	-1.125888	1.000888
d2		.4	.366156	1.09	0.298	-.405904	1.205904
z1		.4375	.0599263	7.30	0.000	.305603	.569397
z2		.5090909	.0295057	17.25	0.000	.4441493	.5740325

(note the constant is suppressed in the regression command, because the model already contains two intercept terms: d1 is the intercept for the 1<sup>st</sup> period, d2 is the intercept for the 2<sup>nd</sup> period.)

Comparing with the separate regressions

reg y z1 if \_n<6 (2)

Source	SS	df	MS	Number of obs = 5		
Model	15.3125	1	15.3125	F( 1, 3)	=	66.82
Residual	.6875	3	.229166667	Prob > F	=	0.0038
				R-squared	=	0.9570
				Adj R-squared	=	0.9427
				Root MSE	=	.47871
-----						
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
z1	.4375	.0535218	8.17	0.004	.2671697	.6078303
_cons	-.0625	.4315066	-0.14	0.894	-1.435746	1.310746

. reg y z2 if \_n>=6 (3)

Source	SS	df	MS	Number of obs = 10		
Model	85.5272727	1	85.5272727	F( 1, 8)	=	276.71
Residual	2.47272727	8	.309090909	Prob > F	=	0.0000
				R-squared	=	0.9719
				Adj R-squared	=	0.9684
				Root MSE	=	.55596
-----						
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
z2	.5090909	.0306046	16.63	0.000	.4385167	.5796652
_cons	.4	.3797926	1.05	0.323	-.4758033	1.275803

Comparison with the separate regressions above show the coefficients are identical as are the sum of the residual sum of squares

$$2.47272727 + .6875 = 3.1602273 = \text{RSS}_{\text{stacked}}$$

- you can check this by saving (squared) residuals for each individual from the pooled regression)

(though the standard errors are not - while the addition of zeros does not change the inverse of the  $x'x$  matrix (inverse of diagonal matrix equals inverse of elements on main diagonal), the estimated ols residual variance are different - because the degrees of freedom in the denominator are not proportional

$$\text{RSS}_{\text{stacked}} = \text{RSS}_1 + \text{RSS}_2 \text{ but } \frac{\text{RSS}_{\text{stacked}}}{N-k} \neq \frac{\text{RSS}_1}{N-k_1} + \frac{\text{RSS}_2}{N-k_2}$$

Also the sum of the total sum of squares in each regression does not add to the stacked total,  $\sum y^2 \neq \sum y_1^2 + \sum y_2^2$  where  $y$  is in mean deviation form)

Then compare with restricted regression (4)

reg y z

Source	SS	df	MS	Number of obs = 15		
Model	127.443885	1	127.443885	F( 1, 13)	=	252.71
Residual	6.55611511	13	.504316547	Prob > F	=	0.0000
				R-squared	=	0.9511
				Adj R-squared	=	0.9473
				Root MSE	=	.71015
-----						
y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
z	.5244604	.0329917	15.90	0.000	.4531862	.5957347
_cons	-.0697842	.3678736	-0.19	0.852	-.8645268	.7249584

---

and do F test comparing RSS from (4) and (1) ( or the sum in (2) and (3) )