

Creating Indicator Variables in Stata

Example from Appendix C4 includes

Y = GPA for 1st year in College, X₁ = ACT test score (taken before admission)

Categorical variable = “Year” = year of admission, from 1996 to 2000 (5 categories)

To create four indicator variables:

xi i.Year

i.Year _IYear_1996-2000 (naturally coded; _IYear_1996 omitted)

To create indicator variables and run the regression analysis at the same time:

xi: regress GPA ACT i.Year

i.Year _IYear_1996-2000 (naturally coded; _IYear_1996 omitted)

Source	SS	df	MS	
Model	38.7250909	5	7.74501817	
Residual	244.723315	699	.350104886	
Total	283.448406	704	.402625577	

Number of obs =	705
F(5, 699) =	22.12
Prob > F =	0.0000
R-squared =	0.1366
Adj R-squared =	0.1304
Root MSE =	.5917

GPA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
ACT	.0579188	.0055663	10.41	0.000	.0469901 .0688475
_IYear_1997	.0791562	.0717559	1.10	0.270	-.0617266 .220039
_IYear_1998	.0822289	.0688475	1.19	0.233	-.0529438 .2174016
_IYear_1999	.0887545	.0703506	1.26	0.208	-.0493693 .2268783
_IYear_2000	.034314	.0708633	0.48	0.628	-.1048164 .1734444
_cons	1.498709	.1461276	10.26	0.000	1.211808 1.785611

To test whether the year of admission is needed in the equation – this tests all years together:

Method 1: List all four of the indicator variables by name:

testparm _IYear_1997 _IYear_1998 _IYear_1999 _IYear_2000

```
( 1)  _IYear_1997 = 0
( 2)  _IYear_1998 = 0
( 3)  _IYear_1999 = 0
( 4)  _IYear_2000 = 0
```

```
F( 4, 699) = 0.60
Prob > F = 0.6648
```

Method 2: Use shorthand code that includes them all:

testparm _IYear*

```
( 1)  _IYear_1997 = 0
( 2)  _IYear_1998 = 0
( 3)  _IYear_1999 = 0
( 4)  _IYear_2000 = 0
```

```
F( 4, 699) = 0.60
Prob > F = 0.6648
```

To create separate intercepts and slopes in the regression:

xi: regress GPA ACT i.Year*ACT

i.Year _IYear_1996-2000 (naturally coded; _IYear_1996 omitted)

i.Year*ACT _IYeaXACT_# (coded as above)

Source	SS	df	MS	

Number of obs =	705
F(9, 695) =	12.33

Model		39.0234908	9	4.33594343	Prob > F	=	0.0000
Residual		244.424915	695	.351690526	R-squared	=	0.1377
Total		283.448406	704	.402625577	Adj R-squared	=	0.1265
					Root MSE	=	.59304

	GPA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	ACT	.0522815	.0128546	4.07	0.000	.0270431	.0775199
_IYear_1997		.0821589	.4622083	0.18	0.859	-.8253331	.9896509
_IYear_1998		-.1078284	.4254001	-0.25	0.800	-.9430518	.727395
_IYear_1999		-.0590787	.4311573	-0.14	0.891	-.9056057	.7874483
_IYear_2000		-.3325866	.4748162	-0.70	0.484	-1.264833	.5996595
	ACT	(dropped)					
_IYeaXA~1997		-.0002628	.0187286	-0.01	0.989	-.0370342	.0365085
_IYeaXA~1998		.0077131	.0170561	0.45	0.651	-.0257745	.0412008
_IYeaXA~1999		.0059855	.0171826	0.35	0.728	-.0277506	.0397216
_IYeaXA~2000		.0149111	.0190717	0.78	0.435	-.0225339	.0523561
_cons		1.637894	.3212589	5.10	0.000	1.00714	2.268649

To test whether the interaction terms (separate slopes) are needed:

testparm _IYeaXACT*

- (1) _IYeaXACT_1997 = 0
- (2) _IYeaXACT_1998 = 0
- (3) _IYeaXACT_1999 = 0
- (4) _IYeaXACT_2000 = 0

F(4, 695) = 0.21
Prob > F = 0.9317

To test whether year is needed at all – separate intercepts and/or slopes:

testparm _IYear* _IYeaXACT*

- (1) _IYear_1997 = 0
- (2) _IYear_1998 = 0
- (3) _IYear_1999 = 0
- (4) _IYear_2000 = 0
- (5) _IYeaXACT_1997 = 0
- (6) _IYeaXACT_1998 = 0
- (7) _IYeaXACT_1999 = 0
- (8) _IYeaXACT_2000 = 0

F(8, 695) = 0.40
Prob > F = 0.9189

It looks like it does not help to take year of admission into account. Here is the regression without it; compare Adj R-squared and Root MSE for this fit with the ones that included year:

. regress GPA ACT

Source		SS	df	MS	Number of obs	=	705
Model		37.8888841	1	37.8888841	F(1, 703)	=	108.47
Residual		245.559522	703	.349302308	Prob > F	=	0.0000
Total		283.448406	704	.402625577	R-squared	=	0.1337
					Adj R-squared	=	0.1324
					Root MSE	=	.59102

	GPA	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
	ACT	.0578005	.0055498	10.41	0.000	.0469044	.0686966
	_cons	1.558702	.1380167	11.29	0.000	1.287728	1.829676