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              Number of obs =     705
-------------+------------------------------           F(  5,   699) =   22.12
Model |  38.7250909     5  7.74501817           Prob > F      =  0.0000
Residual |  244.723315   699  .350104886           R-squared     =  0.1366
-------------+------------------------------           Adj R-squared =  0.1304
Total |  283.448406   704  .402625577           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 |   .0343140 | .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:
```
F(  4,   699) =  0.60
Prob > F =  0.6648
```

Method 2: Use shorthand code that includes them all:
```
testparm _IYear*
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*ACT        _IYeaXACT_#         (coded as above)
```

Source |       SS       df       MS              Number of obs =     705
-------------+------------------------------           F(  9,   695) =   12.33
To test whether the interaction terms (separate slopes) are needed:
\texttt{testparm _IYeaXACT*}

(1) \_IYeaXACT\_1997 = 0
(2) \_IYeaXACT\_1998 = 0
(3) \_IYeaXACT\_1999 = 0
(4) \_IYeaXACT\_2000 = 0

\begin{align*}
F( 4, 695) &= 0.21 \\
\text{Prob} > F &= 0.9317
\end{align*}

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

\begin{align*}
F( 8, 695) &= 0.40 \\
\text{Prob} > F &= 0.9189
\end{align*}

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:
\texttt{. regress GPA ACT}

\begin{align*}
\text{Source} & | \quad \text{SS} \quad \text{df} \quad \text{MS} \\
\hline
\text{Model} & | \quad 37.8888841 \quad 1 \quad 37.8888841 \quad \text{F( 1, 703) = 108.47} \\
\text{Residual} & | \quad 245.559522 \quad 703 \quad .349302308 \quad \text{R-squared = 0.1337} \\
\text{Total} & | \quad 283.448406 \quad 704 \quad .402625577 \quad \text{Adj R-squared = 0.1324} \\
\end{align*}

\begin{align*}
\text{GPA} & | \quad \text{Coeff.} \quad \text{Std. Err.} \quad t \quad \text{P>|t|} \quad [95\% \text{ Conf. Interval}] \\
\hline
\text{ACT} & | \quad .0578005 \quad .0055498 \quad 10.41 \quad 0.000 \quad .0469044 \quad .0686966 \\
\_cons & | \quad 1.558702 \quad .1380167 \quad 11.29 \quad 0.000 \quad 1.287728 \quad 1.829676 \\
\end{align*}