New York University
Wilf Family Department of Politics

Quantitative Research in Political Science I

Professor Patrick Egan

Transforming X with Polynomials

When we have good reasons (theoretical or empirical) to presume that the ceteris paribus relationship between y and a certain x may be non-linear, we transform x to so as to make our model linear in its parameters. One way to do this that we discussed earlier is to substitute the *natural log* of x for x.

Here we focus on another common approach, which is to add one or more *polynomial* terms of x to the model as predictors. A polynomial regression model in which x of degree r with a vector of covariates $\mathbf{z}$ is written

$$y = \beta_1 x + \beta_2 x^2 + ... + \beta_r x^r + \mathbf{z}'\delta + u.$$

The quadratic model

One of the most commonly used polynomial models in practice is the *quadratic* model (dropping $\mathbf{z}$ for now):

$$y = \beta_1 x + \beta_2 x^2 + u.$$

The value of modeling the xy relationship this way is that it permits the effect of x on y to be non-monotonic in that it *changes signs at some point on the range of* x, creating a $\cap$-shaped pattern (if β_1 is positive and β_2 is negative) or a $\cup$-shaped pattern (if β_1 is negative and β_2 is positive) in the plot of residualized y on residualized x. (In practice, we typically find cases where β_1 and β_2 are of the same sign to be less interesting, as they are simply run-of-the-mill instances of x's effect on y being monotonic but non-linear.)

In the quadratic model, the effect of a one-unit change in x on y is of course

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x,$$

meaning that y attains an extremum w.r.t. x at $x = \frac{1}{2}\frac{\beta_1}{\beta_2}$. This is a maximum if $\beta_1 > 0$ and $\beta_2 < 0$ and a minimum if $\beta_1 < 0$ and $\beta_2 > 0$.

Tests of significance

Typically the significance tests we understake with quadratic models answer one of two questions:

1. **Does a quadratic model better fit the data than a linear model?**

The appropriate significance test to answer this question is straightforward: it is the test of the null that $\beta_2 = 0$, and thus the t-test and p-value associated with the estimate $\widehat{\beta}_2$. If we reject the null, then the quadratic model is a better fit; if we accept the null then the quadratic offers us no significant advantage over the linear model.

2. **At what values of x is the effect of x on y significantly distinct—and are these values of any substantive meaning?**

   This is often the much more important question. When x is a predictor of interest (as opposed to simply being a control variable) and we fit it with a quadratic, it is usually because we want to make the claim that the effect of x on y is significantly higher (or lower) *somewhere in the middle of the range of* x than at lower and higher values of x.

   Determining whether this is the case is an analytically distinct question from (1) for two reasons:

   (a) The quadratic may provide a superior fit, but the effect of x doesn't attain an extremum in the empirical range of x.

   (b) The quadratic may provide a superior fit and the effect of x attains an extremum in the empirical range of x, but $\frac{\partial y}{\partial x}$ at this extremum is not significantly different from $\frac{\partial y}{\partial x}$ at other meaningful values of x.

The upshot is that question 2 cannot be answered by simply looking at whether the t-statistic on $\widehat{\beta}_2$ is significant. Rather, it requires calculating predicted values of y across the empirical range of x and determining whether these values are significantly different from one another at values of x that are substantively interesting.

On the following pages, I illustrate this with a simple example from the General Social Survey Cumulative File:

```
set maxvar 7000
use GSS7212_R2.DTA, clear

An increasingly commonplace phenomenon in American politics is that those with low
and high levels of educational attainment are more likely to identify as Democrats
than those at middle levels.  Let's see if this is the case here.

. tab educ

    highest |
    year of |
     school |
  completed |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        151        0.27        0.27
          1 |         41        0.07        0.34
          2 |        142        0.25        0.59
          3 |        238        0.42        1.01
          4 |        309        0.54        1.55
          5 |        386        0.68        2.23
          6 |        752        1.32        3.55
          7 |        845        1.49        5.03
          8 |      2,598        4.57        9.60
          9 |      1,920        3.37       12.97
         10 |      2,635        4.63       17.61
         11 |      3,396        5.97       23.57
         12 |     17,493       30.75       54.32
         13 |      4,742        8.33       62.65
         14 |      6,170       10.84       73.50
         15 |      2,513        4.42       77.91
         16 |      6,988       12.28       90.20
         17 |      1,684        2.96       93.16
         18 |      1,977        3.47       96.63
         19 |        760        1.34       97.97
         20 |      1,157        2.03      100.00
------------+-----------------------------------
      Total |     56,897      100.00

. tab partyid

   political party |
       affiliation |      Freq.     Percent        Cum.
-------------------+-----------------------------------
   strong democrat |      9,117       16.07       16.07
  not str democrat |     12,040       21.22       37.29
      ind,near dem |      6,743       11.89       49.18
       independent |      8,499       14.98       64.16
      ind,near rep |      4,921        8.67       72.83
not str republican |      9,005       15.87       88.70
 strong republican |      5,548        9.78       98.48
       other party |        861        1.52      100.00
-------------------+-----------------------------------
             Total |     56,734      100.00
```

```
. tab partyid, nol

  political |
      party |
affiliation |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      9,117       16.07       16.07
          1 |     12,040       21.22       37.29
          2 |      6,743       11.89       49.18
          3 |      8,499       14.98       64.16
          4 |      4,921        8.67       72.83
          5 |      9,005       15.87       88.70
          6 |      5,548        9.78       98.48
          7 |        861        1.52      100.00
------------+-----------------------------------
      Total |     56,734      100.00
```

For now, we will treat both educ and partyid as continuous. Let's recode partyid to get rid of that pesky "other party" category:

```
. clonevar pid = partyid
(327 missing values generated)

. recode pid (7=.)
(pid: 861 changes made)
```

Now, using Stata's factor variables language, I type:

```
. reg pid c.educ##c.educ

      Source |       SS       df       MS              Number of obs =   55743
-------------+------------------------------           F(  2, 55740) =  365.24
       Model |  2861.19375     2  1430.59687           Prob > F      =  0.0000
    Residual |  218325.428 55740  3.91685375           R-squared     =  0.0129
-------------+------------------------------           Adj R-squared =  0.0129
       Total |  221186.622 55742  3.96804244           Root MSE      =  1.9791

------------------------------------------------------------------------------
         pid |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1693648   .0121507    13.94   0.000     .1455493    .1931802
             |
c.educ#c.educ |  -.0041802   .0004877    -8.57   0.000    -.0051361   -.0032244
             |
       _cons |   1.229714   .0754809    16.29   0.000     1.081771    1.377657
------------------------------------------------------------------------------
```

Sure enough, we get a nice significant coefficient on the quadratic term, which tells us that quadratic x is a superior predictor of y than linear x. If we were to—say—include educ as a control variable in a model of party id, we'd want to include it as a quadratic predictor.

But: At what values of educ is the effect of educ on y significantly distinct---and are these values of any substantive meaning? A first clue that there may be trouble

is to calculate the point at which the effect of educ on pid is estimated to attain a maximum. This is .5 *(.1693648/.0041802) = 20.25, falling just outside x's empirical range (look at previous page to see its max = 20).

We can see this problem graphically with the following commands:

```
. levelsof educ
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

. margins, at(educ=(`r(levels)'))
```

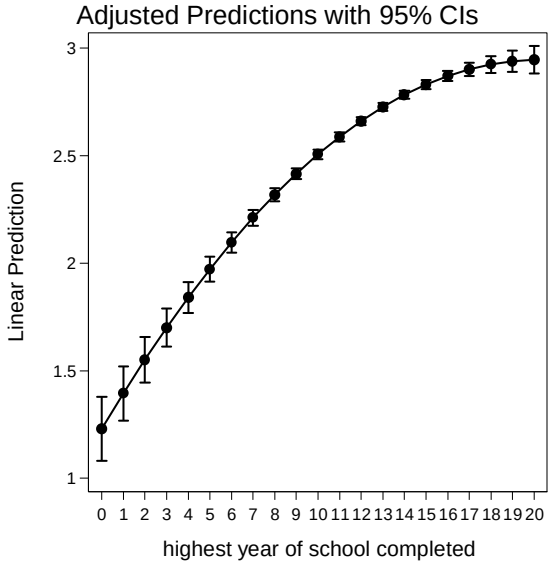
Before going to the output, let's pause for a moment to admire this trick. When we want predictions generated at every value of a continuous x, we first type "*levels of* x". This tells Stata to store all empirical levels of x in the macro `r(levels)'. We then include the macro as the levels of x at which we desire predictors in the *at* option of the *margins* command. Cool, right? Here's the output:

```
Adjusted predictions                              Number of obs   =      55743
Model VCE    : OLS

Expression   : Linear prediction, predict()

1._at        : educ            =           0
2._at        : educ            =           1
3._at        : educ            =           2
4._at        : educ            =           3
5._at        : educ            =           4
6._at        : educ            =           5
7._at        : educ            =           6
8._at        : educ            =           7
9._at        : educ            =           8
10._at       : educ            =           9
11._at       : educ            =          10
12._at       : educ            =          11
13._at       : educ            =          12
14._at       : educ            =          13
15._at       : educ            =          14
16._at       : educ            =          15
17._at       : educ            =          16
18._at       : educ            =          17
19._at       : educ            =          18
20._at       : educ            =          19
21._at       : educ            =          20
```

	Delta-method				
	Margin	Std. Err.	t	P>|t|	[95% Conf. Interval]
_at					
1	1.229714	.0754809	16.29	0.000	1.081771 1.377657
2	1.394899	.0642745	21.70	0.000	1.26892 1.520877
3	1.551723	.0540833	28.69	0.000	1.445719 1.657726
4	1.700186	.0449211	37.85	0.000	1.612141 1.788232
5	1.84029	.0368074	50.00	0.000	1.768147 1.912432
6	1.972032	.0297692	66.24	0.000	1.913684 2.03038
7	2.095414	.0238421	87.89	0.000	2.048684 2.142145
8	2.210436	.0190655	115.94	0.000	2.173068 2.247805
9	2.317098	.0154611	149.87	0.000	2.286794 2.347401
10	2.415398	.0129807	186.08	0.000	2.389956 2.440841
11	2.505339	.011444	218.92	0.000	2.482908 2.527769
12	2.586919	.0105481	245.25	0.000	2.566244 2.607593
13	2.660138	.0099957	266.13	0.000	2.640547 2.67973
14	2.724997	.0096515	282.34	0.000	2.70608 2.743914
15	2.781496	.0096392	288.56	0.000	2.762603 2.800389
16	2.829634	.010337	273.74	0.000	2.809373 2.849894
17	2.869412	.0121787	235.61	0.000	2.845541 2.893282
18	2.900829	.0153747	188.68	0.000	2.870694 2.930963
19	2.923885	.0198948	146.97	0.000	2.884892 2.962879
20	2.938582	.0256333	114.64	0.000	2.88834 2.988823
21	2.944918	.0325016	90.61	0.000	2.881214 3.008621

Now, let’s plot it:

```
. marginsplot

  Variables that uniquely identify margins: educ
```



Thus while we have shown that quadratic x is a better predictor of y than linear x, it is improper to say that the effect of x on y changes in a non-monotonic fashion over the empirical range of x here. It does not.

OK, but let's now try this:

```
. tab race

    race of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |     46,350       81.23       81.23
      black |      7,926       13.89       95.12
      other |      2,785        4.88      100.00
------------+-----------------------------------
      Total |     57,061      100.00

. tab race, nol

    race of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     46,350       81.23       81.23
          2 |      7,926       13.89       95.12
          3 |      2,785        4.88      100.00
------------+-----------------------------------
      Total |     57,061      100.00


. tab year

   gss year |
   for this |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
       1972 |      1,613        2.83        2.83
       1973 |      1,504        2.64        5.46
       1974 |      1,484        2.60        8.06
       1975 |      1,490        2.61       10.67
       1976 |      1,499        2.63       13.30
       1977 |      1,530        2.68       15.98
       1978 |      1,532        2.68       18.67
       1980 |      1,468        2.57       21.24
       1982 |      1,860        3.26       24.50
       1983 |      1,599        2.80       27.30
       1984 |      1,473        2.58       29.88
       1985 |      1,534        2.69       32.57
       1986 |      1,470        2.58       35.15
       1987 |      1,819        3.19       38.34
       1988 |      1,481        2.60       40.93
       1989 |      1,537        2.69       43.63
       1990 |      1,372        2.40       46.03
       1991 |      1,517        2.66       48.69
       1993 |      1,606        2.81       51.50
       1994 |      2,992        5.24       56.75
```

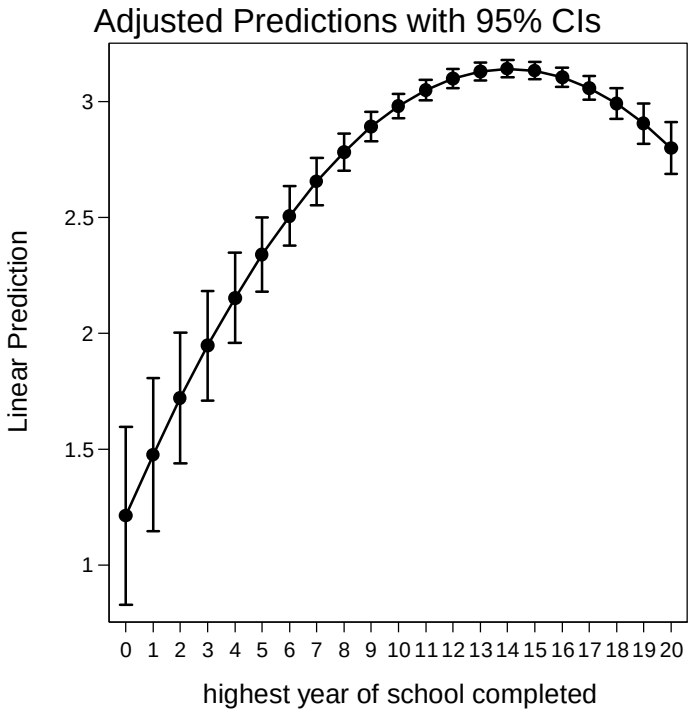
```
       1996 |      2,904        5.09       61.84
       1998 |      2,832        4.96       66.80
       2000 |      2,817        4.94       71.74
       2002 |      2,765        4.85       76.58
       2004 |      2,812        4.93       81.51
       2006 |      4,510        7.90       89.41
       2008 |      2,023        3.55       92.96
       2010 |      2,044        3.58       96.54
       2012 |      1,974        3.46      100.00
------------+-----------------------------------
      Total |     57,061      100.00
```

It wouldn't be surprising if we were to find this relationship to be stronger over time and also to be largely confined to whites (as blacks are pretty much uniformly Democratic regardless of educational attainment). So let's do this:

```
. reg pid7 c.educ##c.educ if year>=2000 & race==1

      Source |       SS       df       MS              Number of obs =   14112
-------------+------------------------------           F(  2, 14109) =   44.84
       Model |  351.015047     2  175.507523            Prob > F      =  0.0000
    Residual |  55226.0801 14109  3.91424482           R-squared     =  0.0063
-------------+------------------------------           Adj R-squared =  0.0062
       Total |  55577.0952 14111  3.93856532           Root MSE      =  1.9784

------------------------------------------------------------------------------
        pid7 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .2741085   .0292453     9.37   0.000     .2167839    .3314332
             |
c.educ#c.educ |   -.009739   .0010901    -8.93   0.000    -.0118757   -.0076023
             |
       _cons |   1.212135   .1957169     6.19   0.000      .828504    1.595766
------------------------------------------------------------------------------
```

The coefficient on the quadratic term remains significant in this restricted set of cases. Let's do the same routine again [omitting some output here]:

```
levelsof educ
margins, at(educ=(`r(levels)'))
marginsplot
```



There we go: among whites surveyed in the year 2000 or later, the effect of education on partisanship is curvilinear. It peaks at about 14 years of education, or among those who have completed some college but have not obtained a bachelor's degree.

Let's say we wanted to identify the levels of x at which education's effect is significantly lower than at educ=14.

We do this by typing the commands

```
levelsof educ
margins, at(educ=(`r(levels)')) pwcompare
```

This yields estimates of the differences in educ's effect on y between all possible pair-wise combinations of the empirically-observed levels of x. The (very long table of) output includes the estimates below. Careful with the interpretation here. Because education can take on the value zero, 14 years of education is actually the 15^{th} empirically observed value. Thus we are interested in the predictions that Stata scores as x=15:

	Contrast	Delta-method Std. Err.	Unadjusted [95% Conf.	Interval]
15 vs 13	.0417912	.0113363	.0195706	.0640118
15 vs 14	.0111567	.0057054	-.0000266	.0223399
16 vs 15	-.0083213	.0063662	-.0207998	.0041572
17 vs 15	-.0361204	.0138681	-.0633038	-.0089371
18 vs 15	-.0833975	.0228494	-.1281855	-.0386096
19 vs 15	-.1501526	.0335421	-.2158995	-.0844056
20 vs 15	-.2363855	.0460977	-.326743	-.1460279

Education's effect on y at 14 years of education is significantly higher than at all other values of x EXCEPT for educ=13 years and educ=15 years, where the differences are insignificant [note how the conf. interval contains zero for these two pairwise comparisons].

Look back at the marginplot. Note that this is a different conclusion than we would reach if we simply (and incorrectly) glanced at which CIs overlap on the plot. Why?