The Stata Journal (*yyyy*) *vv*, Number *ii*, pp. 1–5

Stata tip #: Leaps and bounds

Maarten L. Buis
University of Konstanz
Konstanz, Germany
maarten.buis@uni-konstanz.de

A simple way of adding a variable in a non-linear way to a model is to transform that variable. Common transformations are adding a quadratic term or taking a logarithm, but other transformations are possible as well, like taking the cube root Cox (2011) or adding splines [R] **mkspline**. The purpose of this tip is to discuss yet another under used alternative transformation: the combination of continuous variables and indicator (dummy) variables.

Sometimes a continuous variable consists of qualitatively different segments. A good example of such a variable is hours per week a respondent usually works. In many countries numbers less than 40 on such a variable represent respondents who work part-time, the number 40 represents respondents that have full-time work, and numbers above 40 represent respondents who routinely work over-time. Using the `nlsw88` dataset that comes with Stata, we could analyse how working hours influence average hourly wage, that is, the total earnings in a week divided by the hours worked that week. If we just add hours linearly, then we would conclude that an extra hour working is related to a 4 cents increase in hourly wage.

```
. sysuse nlsw88, clear
(NLSW, 1988 extract)

.
. reg wage hours i.union i.race grade i.south , noheader
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       hours |   .0425932   .0086062     4.95   0.000     .0257144     .059472
             |
  (output omitted)
------------------------------------------------------------------------------
. qui margins, at(union=0 race=1 grade=12 south=0) over(hours)
. marginsplot, noci plotopts(msymbol(i)) ///
>     ytitle("predicted hourly wage")    ///
>     title("")
  Variables that uniquely identify margins: hours
```

However, we might hypothesize that working ‘normal’ hours makes it easier for companies to standardize the allocation of tasks to the workers. As a consequence, companies might be willing to pay a premium for working full-time. This means that working more hours may increase hourly wage, but that there is an extra ‘jump’ at 40. To test that, we can add both the variable hours and an additional indicator variable for full-time workers to our model. Cox and Schechter (2019) wrote a useful tutorial on how to effectively create indicator variables. In this model an extra hour working is

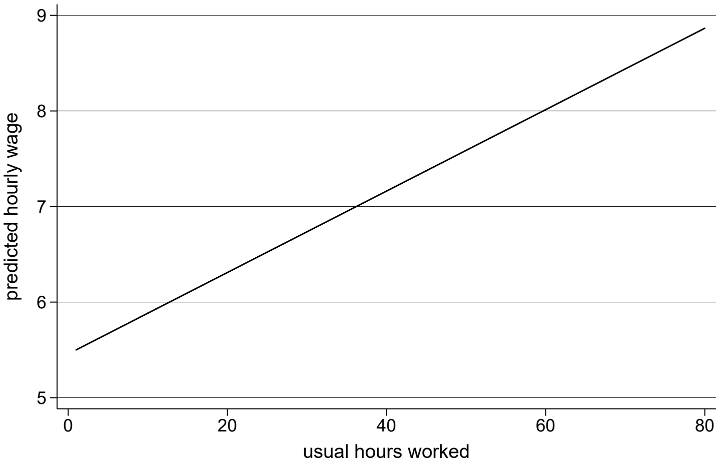

Figure 1: Linear effect of hours

still associated with a 4 cent increase in hourly wage, but those working full-time get a 35 cent ‘bonus’. In this case the indicator variable introduced a single spike at 40 hours per week.

```
. gen fulltime = hours == 40 if hours < .
(4 missing values generated)
. reg wage i.fulltime hours i.union i.race grade i.south , noheader
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  1.fulltime |   .3486994   .1775229     1.96   0.050     .0005353    .6968636
       hours |   .0385415   .0088436     4.36   0.000     .0211972    .0558858
             |
  (output omitted)
------------------------------------------------------------------------------
. qui margins, at(union=0 race=1 grade=12 south=0) over(hours)
. marginsplot, noci plotopts(msymbol(i)) ///
>     ytitle("predicted hourly wage")    ///
>     title("")
  Variables that uniquely identify margins: hours
```

Sometimes over-time is paid at a higher rate. So we might expect that working more hours generally increases the hourly wage, but that after 40 hours there is an extra jump that does not immediately disappear like before, but persists. To test that, we can introduce the variable hours and an indicator variable for those respondents that routinely work over-time to the model. However, the results show that working overtime leads to a persistent (non-significant) 11 cent decrease in hourly wage.

```
. gen overtime = hours >  40 if hours < .
```

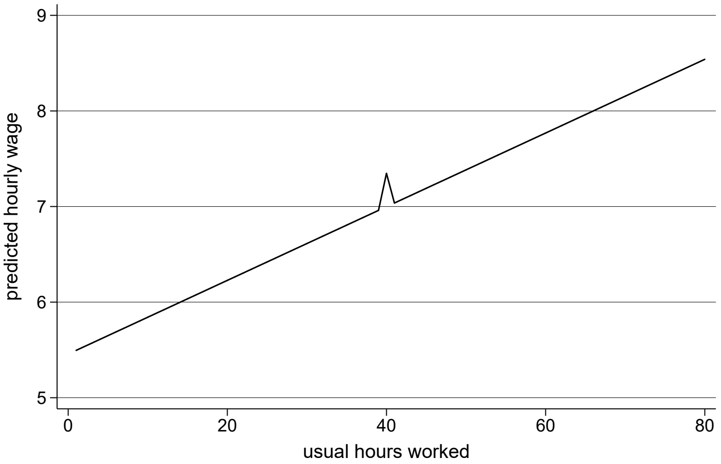

Figure 2: Linear effect of hours with a jump at working full time

```
(4 missing values generated)
. reg wage i.overtime hours i.union i.race grade i.south , noheader
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  1.overtime |  -.1088363   .2742177    -0.40   0.691    -.6466419    .4289693
       hours |   .0449326   .0104327     4.31   0.000     .0244716    .0653936
             |
  (output omitted)
------------------------------------------------------------------------------
. qui margins, at(union=0 race=1 grade=12 south=0) over(hours)
. marginsplot, noci plotopts(msymbol(i)) ///
>     ytitle("predicted hourly wage")    ///
>     title("")
  Variables that uniquely identify margins: hours
```

We forgot that not everybody gets their overtime paid. So for those who get paid for working overtime, overtime will increase their hourly wage. However, for those who are not paid for overtime, overtime will decrease their average hourly income. We might expect that unpaid overtime happens in professions where people are intrinsically motivated (for example, academics), so they may work long hours. Whereas paid overtime happens in occupations where people are less intrinsically motivated, in which case both the workers and the employer have an incentive to keep the amount of overtime within bounds. So we hypothesize that the group of respondents working small amounts of over-time mainly consists of people getting paid over-time, while the group of respondents working large amounts of overtime consists mainly of people who do not get (completely) paid for over-time. In that case, we would expect a sharp increase in hourly wage at 41 hours per week, but after that a decrease. This is implemented by

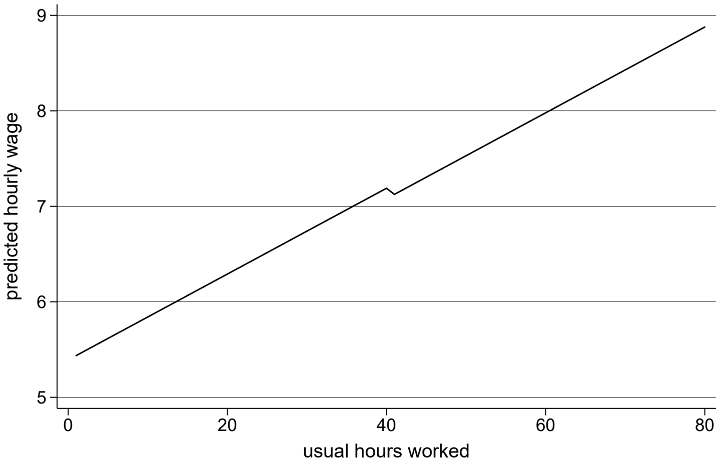

Figure 3: Linear effect of hours with a persistent jump for over-time

including an interaction between the overtime indicator variable and the hours variable. In this case it makes sense to center the hours variable at 41; that way the effect of the overtime indicator variable can be interpreted as the jump that occurs when one starts to work overtime. In this model working an extra hour increases the hourly wage by 6 cents if one works part-time. If one starts working over-time there is an immediate bonus of 1 dollar and 9 cents, but every extra hour decreases the hourly wage by $6-17=-11$ cents. This type of regression is sometimes called segmented, broken-stick, or piecewise regression. This type of model is also closely related to a regression discontinuity design (Calonico et al. 2014, 2017).

```
. gen hours_c = hours - 41
(4 missing values generated)

. reg wage i.overtime##c.hours_c i.union i.race grade i.south , noheader
------------------------------------------------------------------------------------
               wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------+---------------------------------------------------------------
         1.overtime |   1.090213   .3562397     3.06   0.002     .3915431    1.788883
            hours_c |   .0638511   .0109757     5.82   0.000     .0423253     .085377
                    |
 overtime#c.hours_c |
                 1  |  -.1728014   .0331008    -5.22   0.000    -.2377198    -.107883
                    |
  (output omitted)
------------------------------------------------------------------------------------

. lincom 1.overtime#c.hours_c + hours_c

 ( 1)  hours_c + 1.overtime#c.hours_c = 0

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1089503   .0312445    -3.49   0.000    -.1702281   -.0476725
------------------------------------------------------------------------------
```

```
. qui margins, at(union=0 race=1 grade=12 south=0) over(hours)
. marginsplot, noci plotopts(msymbol(i)) ///
>     ytitle("predicted hourly wage")    ///
>     title("")
  Variables that uniquely identify margins: hours
```

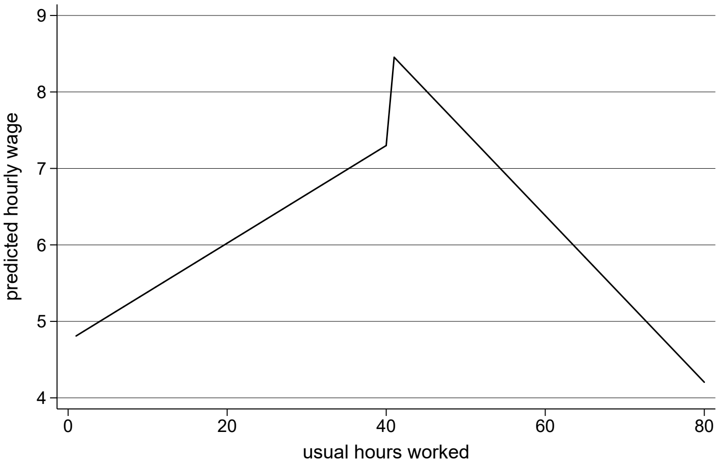

Figure 4: different linear effects of hours for respondents working over-time or not with a jump

So by combining continuous and indicator variables one can allow for non-linearity by adding spikes, persistent jumps, or complete breaks to the regression line. This flexibility allows one to tailor the kind on non-linearity in the model to the research question and what one knows about the variables involved with only a few parameters. Moreover, those parameters those parameters are easy to interpret.

References

Calonico, S., M. D. Cattaneo, M. H. Farrell, and R. Titiunik. 2017. rdrobust: Software for regression-discontinuity designs. *The Stata Journal* 17(2): 372–404.

Calonico, S., M. D. Cattaneo, and R. Titiunik. 2014. Robust data-driven inference in the regression-discontinuity design. *The Stata Journal* 14(4): 909–946.

Cox, N. J. 2011. Stata tip 96: Cube Roots. *The Stata Journal* 11(1): 149–154.

Cox, N. J., and C. B. Schechter. 2019. Speaking Stata: How best to generate indicator or dummy variables. *The Stata Journal* 19(1): 246–259.