Econ 452 Section 10 - STATA

Connor Cole

November 13, 2015

Squared Variables

As we discussed in class, sometimes we want to include squared values of variables, either because the theory seems to suggest that there should be squared values (for example, the Mincer earnings model directly implies that there should be an experience and experienced squared variable when modeling wages over the life cycle), or because we think it fits the data better. Including squared variables is straight forward. First we generate a new variable that is a square of the previous variable. Let's say we're thinking about a regression of `VariableName1` on `VariableName2` and `VariableName3` , and we'd like to include a square of `VariableName2`

```
generate VariableName2sq = VariableName2^2

regress VariableName1 VariableName2 VariableName2sq VariableName3
```

Note that even though $VariableName2sq$ is a function of $VariableName2$, $VariableName2sq$ is not perfectly collinear with $VariableName2$ as it is not a linear function of $VariableName2$.

Testing Squared Variables

If we wanted to test whether or not the relationship between a dependent variable and the independent variable should include a higher order term like a square, then we can simply include these higher order terms and test whether or not the coefficients on these variables is statistically distinguishable from 0. If we had estimated the previous model, then we could test this by simply typing:

```
test (VariableName2sq=0)
```

If we aren't sure whether or not $VariableName2$ belongs in the regression at all, we could test the joint significance of the linear and quadratic terms as follows:

```
test (VariableName2=0) (VariableName2sq=0)
```

We previously covered these F-tests in a different section.

Marginal Effects with Squares

Now, let's consider the marginal effects of a change in $VariableName2$ on $VariableName1$. Consider first what will happen if the 'true' model is:

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \beta_2 VariableName3 + \epsilon$$

Note that for any value of $VariableName2$, the marginal effect of $VariableName2$ on $VariableName1$ is: β_1, as:

$$\frac{\partial \mathbb{E}[VariableName1|X]}{\partial VariableName2}\Big|_{VariableName2=c_2, VariableName3=c_3} = \beta_1$$

for all values of $VariableName2$ and $VariableName3$. That is, our model has $VariableName2$ have a constant marginal effect on $VariableName2$.

What happens if we include a square? Then our true model looks like:

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \beta_2 VariableName2sq + \beta_3 VariableName3 + \epsilon$$

And the marginal effect of $VariableName2$ on $VariableName1$ is now:

$$\frac{\partial \mathbb{E}[VariableName1|X]}{\partial VariableName2}\Big|_{VariableName2=c_2, VariableName3=c_3} = \beta_1 + \beta_2 c_1$$

Unlike previously, the value of $VariableName2$ impacts the partial effect of $VariableName2$.

Now, let's think about marginal effects in STATA. Previously, we used the `margins` command to look at predicted values of the dependent variable, $\hat{y}$. Now, we will use the `margins` command to look at marginal effects of different covariates on the dependent variable.

Let's say we returned to our original regression without squares of covariates:

```
regress VariableName1 VariableName2 VariableName3
```

Then we could compute the marginal effects of $VariableName2$ on $VariableName1$ at specific values of $VariableName2$ and $VariableName3$ as follows:

```
margins, dydx(VariableName2) at(VariableName2=2 VariableName3=10)
```

In the language of the equations presented above, this command will estimate and give standard errors for:

$$\frac{\partial \mathbb{E}[\widehat{VariableName1}|X]}{\partial VariableName2}\Big|_{VariableName2=2,VariableName3=10} = \hat{\beta}_1$$

Since the model we estimated forces marginal effects to be constant across all values of $VariableName2$ and $VariableName3$, it doesn't matter in this case what values at which we specify the marginal effects to be evaluated. The results that we get typing in this command in terms of estimated marginal effects are the same as those we would see from the standard regression output table in STATA.

Now, let's add the squared term to our estimated regression. We could estimate this regression as we did previously:

```
regress VariableName1 VariableName2 VariableName2sq VariableName3
```

This regression returns the proper estimates of coefficients and all relevant information for hypothesis testing. However, note that if we tried to use the `margins` command to look at marginal effects, STATA would not know that we had made $VariableName2sq$ by squaring $VariableName2$, and hence would only compute marginal effects by looking at movement in $VariableName2$ alone. While estimating the regression this way returns us the right coefficients, if we want to look at marginal effects then we need to tell STATA that we are creating a variable that is the square of a previous. The following command estimates the same relationship as the regression above, but tells STATA that there is a variable that is a direct square of another variable included:

```
regress VariableName1 VariableName2 c.VariableName2#c.VariableName2
VariableName3
```

Then we could compute the marginal effects of $VariableName2$ on $VariableName1$ at specific values of $VariableName2$ and $VariableName3$ using the same command:

```
margins, dydx(VariableName2) at(VariableName2=2 VariableName3=10)
```

Once again, in the language of the equations above, this command estimates:

$$\frac{\partial \mathbb{E}[\widehat{VariableName1}|X]}{\partial VariableName2}\Big|_{VariableName2=2,VariableName3=10} = \hat{\beta}_1 + \hat{\beta}_2 \cdot 2$$

Note that, unlike previously, our choice of the value of $VariableName2$ that we evaluate at matters for our marginal effect here.

Interacted Variables

The same techniques apply to looking at interaction terms. Let's say that we're considering estimating the following model:

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \beta_2 VariableName3 + \epsilon$$

If we believed that there might exist some interaction between $VariableName2$ and $VariableName3$, we would then estimate the model:

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \beta_2 VariableName3 + \beta_3 VariableName2 \cdot VariableName3 + \epsilon$$

Note that the model without interaction terms as a basis precedes the model with interaction terms. We could estimate this model by producing an additional variable where we interact $VariableName2$ and $VariableName3$:

```
generate InteractTerm=VariableName2*VariableName3

regress VariableName1 VariableName2 VariableName3 interactTerm
```

Interpreting what happens with square variables can be a bit tricky. Again, remember that we want the model without interactions to precede the model with interactions. Let's say our underlying model was:

$$VariableName1 = \beta_0 + \beta_1 VariableName2 + \beta_2 VariableName3 + \beta_3 VariableName3^2 + \epsilon$$

Then, if we believed that there might exist interactions between $VariableName2$ and $VariableName3$, then we would estimate the following model:

$$\begin{aligned} VariableName1 = &\beta_0 + \beta_1 VariableName2 + \beta_2 VariableName3 + \beta_3 VariableName3^2 + \\ &\beta_4 VariableName3 \cdot VariableName2 + \beta_5 VariableName3^2 \cdot VariableName2 + \epsilon \end{aligned}$$

Note that we have interacted all terms involving $VariableName3$ with all terms involving $VariableName2$ from our original model. We could estimate this model by producing the additional interaction variables that we need:

```
generate VariableName3sq=VariableName3^2
```

```
generate InteractTerm=VariableName3*VariableName2
```

```
generate InteractTermsq=VariableName3sq*VariableName2
```

```
regress VariableName1 VariableName2 VariableName3 VariableName3sq InteractTerm
InteractTermsq
```

Testing Interactions

If we think two variables interact, then we can test whether or not a differential relationship with the dependent variable exists across observations by including the interacted term and testing whether or not the interacted term is significant. In the previous model without squares we estimated, we would do this by typing:

```
test (InteractTerm=0)
```

If, on the other hand, we had included squared terms that we had interacted, then if we were testing whether or not there is an interaction we would need to test whether or not the variables are jointly significant. So, in the model with squares, we would do this by typing:

```
test (InteractTerm=0) (InteractTermsq=0)
```

Marginal Effects with Interacted Terms

Consider the marginal effect of a change in $VariableName2$ in the model with an interacted term we described at the beginning of the section:

$$\frac{\partial \mathbb{E}[VariableName1|X]}{\partial VariableName2}\bigg|_{VariableName2=c_2, VariableName3=c_3} = \beta_1 + \beta_3 c_3$$

As previously, we could use the estimates we had before to estimate this partial effect, but if we wanted to use the `margins` command, STATA would not know that one of our variables was producing by interacting two other variables. So, if we wanted to estimate the model in a way that allows us to calculate marginal effects, we would type:

```
regress VariableName1 VariableName2 VariableName3
c.VariableName2#c.VariableName3
```

```
margins, dydx(VariableName2) at(VariableName2=2 VariableName3=10)
```

This command estimates:

$$\frac{\partial \mathbb{E}[Vari\widehat{ableN}ame1|X]}{\partial VariableName2}\Big|_{VariableName2=2, VariableName3=10} = \hat{\beta}_1 + \hat{\beta}_3 \cdot 10$$