ACADEMIC YEAR 2016/2017
Università degli Studi di Milano
GRADUATE SCHOOL IN SOCIAL, ECONOMIC AND POLITICAL SCIENCES
APPLIED MULTIVARIATE ANALYSIS
Luigi Curini
luigi.curini@unimi.it

Regression Diagnostics: Nonlinear Regression Functions – Interaction Models

When we do linear regression, we assume that the relationship between the response variable and the predictors is linear. This is the assumption of linearity. If this assumption is violated, the linear regression will try to fit a straight line to data that does not follow a straight relationship. This is a common cause of non-normally distributed residuals as well as heteroscedasticity. Therefore checking for linearity is important to be sure that no main assumption of OLS is violated (of course the non-linearity can implies these violation, but not necessary so!).

Moreover, by being linear OLS is **also an additive** technique: we assume that the effect of an IV on the DV is the same for all values of the other IV in the model, once they are fixed at a given value (that is, the slope of the population regression function is constant, so that the effect on Y of a unit change in X does not depend on the value of one or more IV). In this sense, non-linearity is important also for the interpretation you can give to the results of your model.

The regression function can be nonlinear in two different cases: 1) the effect on Y of a change in X1 depends on the value of X1 itself. For example, reducing class sizes by one student per teacher might have a greater effect on the math achievement of students if class sizes are already small than if they are so large that the teacher can do little more than keep the class under control. If so, then the test score (Y) is a nonlinear function of the student-teacher ratio (X1), where this function is steeper when X1 is small (L graph vs. just a straight line). 2) the effect on Y of a change in X1 depends on the value of another IV, say X2. For example, the effect on test scores of reducing the student-teacher ratio will be greater in schools where the parents of children have a higher level of education than in schools where the parents of children have a lower level of education (in a figure, the slope of the relationship between X1 and Y depends on X2, and it will be more or less steeper according to X2).

In both cases, the population regression function is a nonlinear function of the IV, that is the conditional expectation of Y is a nonlinear function of one or more of the X's. Although they are nonlinear in the X's, these models are linear functions of the unknown coefficients (or parameters) of the population regression model that, taken together, are able to express the non-linear relationship between Y and Xs. Therefore these coefficients can be estimated and tested using OLS.

Note that the term nonlinear regression applies to two different families of models. In the first family, the regression function is a nonlinear function of the X's but is a linear function of the unknown parameters (the Betas). In the second family, the regression function is a nonlinear function of the unknown parameters and may or may not be a nonlinear function of the X's (this is the case of a logit, probit, etc.). Here we consider just the first family of models.

You already discussed quadratic models. In this lecture we focus instead on…

Interaction models

If interaction is taking place, then OLS will not capture this effect, unless we model it. For example, let's analyze the relationship between the thermometer rating of the Democratic Party, and respondents' attitudes toward the role of government, controlling for respondent's political knowledge.

```
codebook dem_therm (goes from 0 to 100)
codebook progovmt (goes from 0 to 3)
codebook polknow3 (goes from 0 to 3)
recode polknow3 (0=0 "Low pol. knowledge") (1/2=1 "Medium-high
pol. knowledge"), gen (poldummy)
```

Let's do a preliminary analysis (making controlled comparison!)

```
tab poldummy progovmnt, sum(dem_therm) nost
```

The dem-therm progovmnt relationship does indeed strengthen as political knowledge increases. How would we use regression analysis to estimate the size and the statistical significance of these relationships (that is, interactive ones)?

If we estimate the following model:

Demo therm = a +b1*progov+b2*poldummy

the three parameters are the additive building blocks of the model. We can use these parameters to estimate the mean of the DV for different combinations of the IV. Still in this case, the relationship between progov and demo_therm (that is, the slope of the regression coefficient relating the IV and the DV) would be constant, no matter the value of the poldummy variable. That is, when poldummy=0, the regression becomes a+b1*progov, so the intercept is a and the slope is b1. When poldummy=1, now the regression is a+b1*progov+b2, so the slope remains b1, while the intercepts becomes a+b2. Stated in terms of the example, b1 is the effect of one unit increase in progov, holding the political knowledge constant, and b2 is the effect of political knowledge, holding the progov constant. Thus b2 is the difference between the intercepts of the two regression lines. In this specification, the effect of one unit increase in progov has the same impact on Demo therm regardless of the value of poldummy, that is the two lines have the same slope:

```
reg dem_therm progovmnt poldummy
predict yhat1
separate yhat1, by(poldummy)
```

Alternatively you could write:

```
line yhat10 yhat11 progovmnt, sort legend(cols(1))
```

However, we want to adjust this additive estimate, depending on where we are on the political knowledge variable. As the values of political knowledge chances, what about the relationship between progov and demo therm? Is it constant or does it change? To accomplish this goal, we need to include an interaction variable as an IV, that is we multiply our two IV.

```
gen interact = progovmnt * poldummy
```

Now, all respondents who are coded 0 on progovnm will of course have a value of 0 on the interaction variable. For respondents coded 1 or higher on progovnm however the magnitude of the interaction variable will increase as political knowledge increases. Now the coefficient b3 tells us how much to adjust our additive estimate for each one-unit increase in political knowledge.

So:

```
Demo therm = a +b1*progov+b2* poldummy +b3*interaction
```

Now the equation becomes:

```
Demo therm = 54.3 +3.33*progo-12.5* poldummy +5.35*(progovmnt*
poldummy)
```

The constant is the estimated mean of dem_therm for respondents who have values of 0 on all the IV. Now consider progovm. For a unit change in progovm, its impact on demo_therm will be equal to b2+b3*polknow3. When polknow3 is equals 0, then for every increase of one unit in progo, the demo_therm, controlling for polknow, increases by 3.33. However, if polknow is different from 0, now there is also an interaction effects. In particular, if now polknow is equals to 1, every increase by one unit of progo, controlling for polknow, is associated with an increase of almost 9 points in the demo_therm.

We can even use a visual representations to simplify and clarify our relationship.

```
predict zhat1
separate zhat1, by(poldummy)
line zhat10 zhat11 progovmnt, sort legend(cols(1))
```

```
twoway (lfit dem_therm progovmnt if(poldummy==0)) (lfit dem_therm
progovmnt if(poldummy==1)) , legend(order(1 "Low pol. knowledge" 2
"Medium-High pol. knowledge") cols(1))
```

In the figure, the two lines have different slopes and intercepts. The different slopes permit the effect of one unit increase of progovn to differ for low and high political knowledge respondents.

Demo therm = a +b1*progov+b2*poldummy+b3*(progov*poldummy)

In this case, when poldummy=0, the regression function is a+b1progov (intercept a and slope b1), whereas when poldummy=1, the regression function is (a+b2) + (b1+b3)progov (intercept a+b2 and slope b1+b3). The difference between the two intercepts is b2 and the difference between the two slopes is b3. Now we have two different regression functions relating Y to X, depending on the value of Z!. As we can see, the slopes of the linear relationship between progovm and the predicted values are different, according to the values of polknow. Moreover, also the intercept is different, being lower for the high polknow values (given the large and negative coefficient for poldummy).

We can repeat the entire analysis this time using the original polknow3 variable.
Let's do once again a preliminary analysis (making controlled comparison!)

```
tab polknow3 progovmnt, sum(dem_therm) nost
```

The dem-therm progovmnt relationship does indeed strengthen as political knowledge increases. How would we use regression analysis to estimate the size and the statistical significance of these relationships (that is, interactive ones)? We have to follow the same procedure employed earlier, that is:

```
gen interact = progovmnt * polknow3
```

Therefore:

Demo therm = a +b1*progov+b2*polknow3+b3*interaction

The coefficient b3 tells us how much to adjust our additive estimate for each one-unit increase in political knowledge.

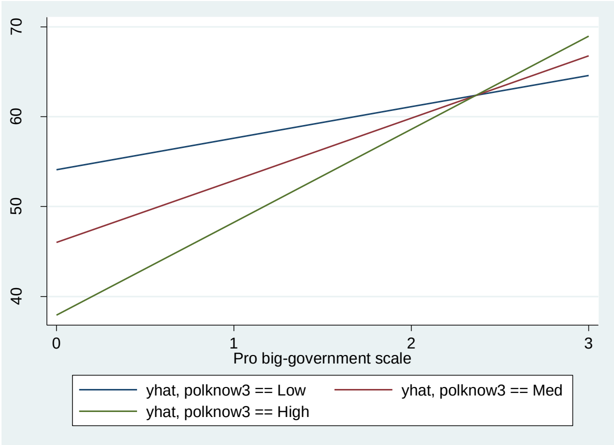
```
reg dem_therm progovmnt polknow3 interact
```

Now the estimated equation becomes:

Demo therm = 54.1 +3.5*progo-8.1*polknow3+3.42*(progovmnt*polknow3)

The constant is the estimated mean of dem_therm for respondents who have values of 0 on all the IV. Now consider progovm. For a unit change in progovm, its impact on demo_therm will be equal to b2+b3*polknow3. When polknow3 is equals 0, then for every increase of one unit in progo, the demo_therm, controlling for polknow, increases by 3.5. However, if polknow is different from 0, now there is also an interaction effects. In particular, if now polknow is equals to 1, every increase by one unit of progo, controlling for polknow, is associated with an increase of almost 7 points in the demo_therm. And so on.

We can even use a visual representations to simplify and clarify our relationship.



As we can see, the slopes of the linear relationship between progovm and the predicted values are different, according to the values of polknow. Moreover, also the intercept is different, being lower for the high polknow values (given the large and negative coefficient for polknow3). You can also use LINCOM ecc. ecc.

Interaction between two continuous variables or interval-level variables

Now suppose that both IV are continuous. For example, let's estimate the impact on `demo_therm` of `income_hh` and `kerry_therm`. Moreover let's suppose that there is an interaction relationship between these two variables. In particular we suspect a "liberal-bias hypothesis" that runs as follows: "*increasing income decreases the relative support for the Democratic Party but in much weaker (stronger) way among people that really like (dislike) the candidate John Kerry*". In this sense, we take the `kerry_therm` variable as a proxy for being a liberal citizen: the more you like Kerry, the more you are liberal.

If the relationship is additive, the effect of an increase in `income_hh` does not depend on `kerry_therm`. However, there might be an interaction between these two variables so that the effect on `demo_therm` of one unit increase in `income_hh` depends on `kerry_therm`. The interaction can be modeled by augmenting the linear regression model with an interaction term that is the product of X1 and X2.

Y=a+b1X1+b2X2+b3(X1*X2)

The interaction term allows the effect of a unit change in X1 to depend on X2. Indeed, if X1 change, we get: Y+deltaY=a+b1(X1+deltaX1)+b2X2+b3((X1+deltaX1)*X2)

Thus the effect on Y of a change in X1, holding X2 constant, is: deltaY/deltaX1=b1+b3X2, which depends on X2 (you can get the same result by *taking the derivative* of the equation Y=a+b1X1+b2X2+b3(X1*X2) in terms of X1, i.e.,$\partial Y/_{\partial X1}$). For example, in our example, if b3 is positive, then the effect on `demo_therm` of one point increase of `kerry_therm`, is greater, by the amount b3, as the income of the respondents increase. The deltaY/deltaX1 is called the *marginal effect* of X1.

In a similar way, the marginal effect of X2 is deltaY/deltaX2.

Putting the two effects together shows that the coefficient of b3 on the interaction term is the effect of a unit increase in X1 and X2 above and beyond the sum of the effects of a unit increase in X1 alone and a unit increase in X2 alone (this is true whether X1 and/or X2 are continuous or binary). That is, if X1 changes by deltaX1 and X2 changes by deltaX2, then the expected change in Y is deltaY=(b1+b3X2)deltaX1+(b2+b3X1)deltaX2+b3deltaX1deltaX2. The first term is the effect from changing X1 holding X2 constant; the second term is the effect from changing X2 holding X1 constant; and the final term is the extra effect from changing both X1 and X2.

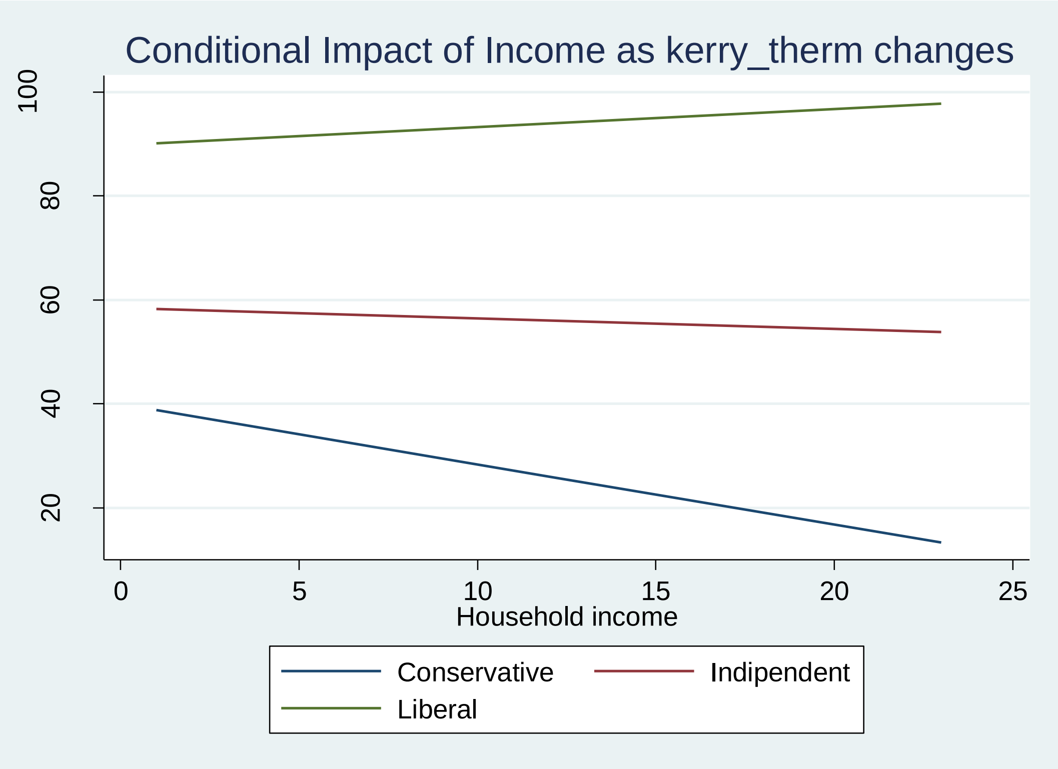
Back to our example about nes2004: according to our hypothesis we expect a significant and **negative** coefficient for `income_hh` and a significant and **positive** coefficient for the `interaction` term (why that? Go back to our previous hypothesis!!!).

```
reg dem_therm income_hh kerry_therm
gen interaction = kerry_therm * income_hh
reg dem_therm income_hh kerry_therm interaction
```

```
------------------------------------------------------------------------------
   dem_therm |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   income_hh |  -.7548556   .1951649    -3.87   0.000    -1.137821   -.3718899
 kerry_therm |   .5506042   .0524792    10.49   0.000     .4476259    .6535825
 interaction |   .0081741   .0032035     2.55   0.011     .0018881    .0144602
       _cons |   33.62795   3.306702    10.17   0.000     27.13932    40.11659
------------------------------------------------------------------------------
```

And indeed:

```
twoway (lfit dem_therm income_hh  if (kerry_therm ==0)) (lfit
dem_therm income_hh if (kerry_therm ==50)) (lfit dem_therm
income_hh if (kerry_therm ==100)), legend(order(1 "Conservative" 2
"Indipendent" 3 "Liberal"))
```



When `kerry_therm` is at 0 ("Conservative" type) the slope of the line (the marginal effect) relating `income_hh` and `dem_therm` is estimated to be (-0.754+0.008*0)=-0.754. When the `kerry_therm` is at 50 ("Indipendent type") the marginal effect is ((-0.754+0.008*50)=-0.354, that is the slope of the line is estimated to be steeper in the first case. Finally, when the `kerry_therm` is at 100 ("Liberal type") the marginal effect is ((-0.754+0.008*100)=+0.46. That is, the marginal impact of `income_hh` on `dem_therm` changes its sign! Note however that up to now we have estimated an average marginal impact, without considering the level of uncertainty connected to it (i.e. its standard error). So that we cannot be sure that for example +0.46 is statistically different from zero or no. How to do that?…More on this next lecture!

Summing up: a general approach to Modeling nonlinearities using OLS. Some Golden Rules!

1. identify a possible nonlinear relationship (from the theory!): that is analysts should use interaction models whenever the hypothesis they want to test is conditional in nature
2. specify a nonlinear function and estimate its parameters by OLS. It can be a quadratic or an interaction term. Scholars should include all constitutive terms in their interaction model specifications
3. determine whether the nonlinear model improves upon a linear model. Looking at what? At the R^2?
4. plot the estimated nonlinear regression function
5. estimate the effect on Y of a unit change in X. That is, scholars should not interpret constitutive terms as if they are unconditional marginal effects.
6. Analysts should calculate substantively meaningful marginal effects and standard errors. This is relevant. If the interaction term is significant and the two constitutive terms of such

interaction term are not significant, this is not a problem! That would mean that the impact of X on Y could be significant only for some values of Z for example. We will investigate better this issue in the next lecture.

Two good papers to read to understand the mistakes you can commit when you do not interpret correctly the non-linear relationship you are adding in your model! In particular, the note 4 in the second paper explains the mistakes of the first paper.

Budge, I. and R. I. Hofferbert. 1990. Mandates and Policy Outputs: U.S. Party Platforms and Federal Expenditures. *American Political Science Review* 84(1): 111-131.

King, G. and M. Laver. 1993. Party Platforms, Mandates, and Government Spending. *The American Political Science Review* 87(3): 744-750.