迴歸進階主題 3

Functional Form Misspecification (Quadratics)

(取自 Wooldridge Chap 3, 6, 9)

- Chap 3 Slides p.5-6

- **Example: Family income and family consumption**

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u$$

Family consumption | Family income | Family income squared | Other factors

  - Model has two explanatory variables: inome and income squared
  - Consumption is explained as a quadratic function of income
  - One has to be very careful when interpreting the coefficients:

By how much does consumption increase if income is increased by one unit?

$$\frac{\Delta cons}{\Delta inc} \approx \beta_1 + 2\beta_2 inc$$

Depends on how much income is already there

- **Example: CEO salary, sales, and CEO tenure**

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 ceoten + \beta_3 ceoten^2 + u$$

Log of CEO salary | Log sales | Quadratic function of CEO tenure with the firm

  - Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm
  - Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm

- **Meaning of "linear" regression**
  - The model has to be linear in the parameters (not in the variables)

- Chap 3 Exercise

C10 Use the data in HTV to answer this question. The data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

(i) What is the range of the *educ* variable in the sample? What percentage of men completed twelfth grade but no higher grade? Do the men or their parents have, on average, higher levels of education?

(ii) Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

by OLS and report the results in the usual form. How much sample variation in *educ* is explained by parents' education? Interpret the coefficient on *motheduc*.

(iii) Add the variable *abil* (a measure of cognitive ability) to the regression from part (ii), and report the results in equation form. Does "ability" help to explain variations in education, even after controlling for parents' education? Explain.

(iv) (Requires calculus) Now estimate an equation where *abil* appears in quadratic form:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u.$$

Using the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$, use calculus to find the value of *abil*, call it $abil^*$, where *educ* is minimized. (The other coefficients and values of parents' education variables have no effect; we are holding parents' education fixed.) Notice that *abil* is measured so that negative values are permissible. You might also verify that the second derivative is positive so that you do indeed have a minimum.

(v) Argue that only a small fraction of men in the sample have "ability" less than the value calculated in part (iv). Why is this important?

(vi) If you have access to a statistical program that includes graphing capabilities, use the estimates in part (iv) to graph the relationship between the predicted education and *abil*. Set *motheduc* and *fatheduc* at their average values in the sample, 12.18 and 12.45, respectively.

C3.10 (i) The variable *educ* ranges from 6 to 20. Out of 1,230 men, 512, or 41.63%, completed 12^{th} grade, but no higher. The average level of education for the men in the sample is about 13.04, which is higher than the average of *motheduc* (12.18) and *fatheduc* (12.45).

(ii) The regression results are

$$\widehat{educ} = 6.96 + .304\, motheduc + .190\, fatheduc$$
$$n = 1{,}230 \quad R^2 = .249.$$

About 25% of the variation in *educ* is explained by parents' education. The coefficient on *motheduc* means that, holding father's education fixed, another year of the mother's education is associated with about .304 additional years of education for the child, or slightly less than one-third of a year.

(iii) When *abil* is added to the regression, we get

$$\widehat{educ} = 8.45 + .189\, motheduc + .111\, fatheduc + .502\, abil$$
$$n = 1{,}230 \quad R^2 = .428$$

The three explanatory variables together explain almost 43% of the variation in *educ*. This is much more than in part (ii); clearly *abil* is helping.

(iv) When $abil^2$ is added to the regression, the estimated equation is

$$\widehat{educ} = 8.24 + .190\, motheduc + .109\, fatheduc + .401\, abil + .051\, abil^2$$
$$n = 1{,}230 \quad R^2 = .444$$

The derivative with respect to *abil* is $.401 + .102\, abil$. Setting equal to zero and solving gives

$$abil^* = -.\frac{401}{.102} \approx -3.93,$$

so about -4. The second derivative is $.102$, and so we know we have found the global minimum.

(v) Out of 1,230 men, only 15 have $abil < -3.93$, or only about 1.2 percent of the sample. This is reassuring because it means we can effectively ignore what is happening to the left of -3.93. The important story is that the level of education increases with ability at an increasing rate.

(vi) I used the equation from part (iv) and plugged in the mean values for *motheduc* and *fatheduc*. Thus, I used the equation

$$\widehat{educ} = 8.24 + .190(12.18) + .109(12.45) + .401\, abil + .051\, abil^2$$

which has an intercept of about 11.9. I generated 2,000 values of *abil*, equally spaced between -5 and 6, to generate the following graph:

- Chap 6, Slides p.3-4

- **Using quadratic functional forms**
- **Example: Wage equation**

Concave experience profile

$$\widehat{wage} = \underset{(.35)}{3.73} + \underset{(.041)}{.298}\, exper - \underset{(.0009)}{.0061}\, exper^2$$

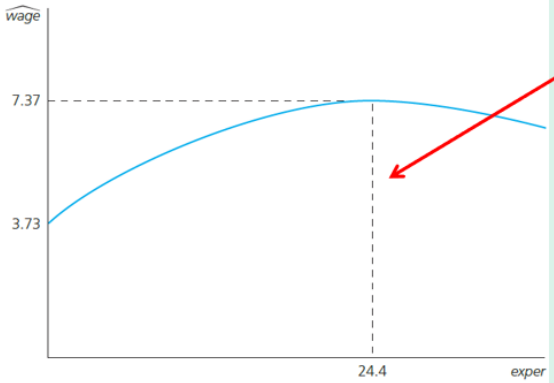
$$n = 526, R^2 = .093$$

- **Marginal effect of experience**

The first year of experience increases the wage by some \$.30, the second year by .298 - 2(.0061)(1) = \$.29 etc.

$$\frac{\Delta wage}{\Delta exper} = .298 - 2(.0061)exper$$

- **Wage maximum with respect to work experience**



$$x^* = \left|\frac{\hat{\beta}_1}{2\hat{\beta}_2}\right| = \left|\frac{.298}{2(.0061)}\right| \approx 24.4$$

Does this mean the return to experience becomes negative after 24.4 years?

Not necessarily. It depends on how many observations in the sample lie right of the turnaround point.

In the given example, these are about 28% of the observations. There may be a specification problem (e.g. omitted variables).

EXAMPLE 6.2 **Effects of Pollution on Housing Prices**

We modify the housing price model from Example 4.5 to include a quadratic term in *rooms*:

$$\log(price) = \beta_0 + \beta_1\log(nox) + \beta_2\log(dist) + \beta_3 rooms + \beta_4 rooms^2 + \beta_5 stratio + u. \quad [6.14]$$

The model estimated using the data in HPRICE2 is

$$\widehat{\log(price)} = \underset{(.57)}{13.39} - \underset{(.115)}{.902}\log(nox) - \underset{(.043)}{.087}\log(dist) - \underset{(.165)}{.545}\, rooms + \underset{(.013)}{.062}\, rooms^2 - \underset{(.006)}{.048}\, stratio$$

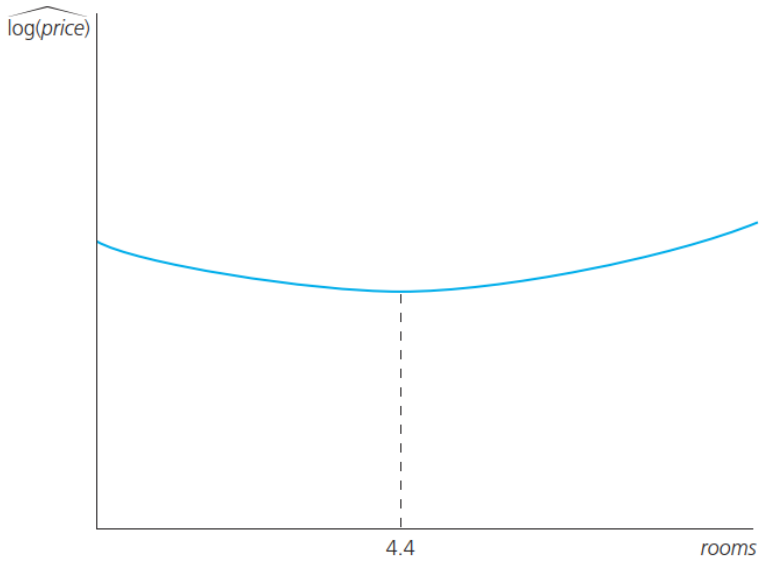
$$n = 506, R^2 = .603.$$

The quadratic term $rooms^2$ has a t statistic of about 4.77, and so it is very statistically significant. But what about interpreting the effect of *rooms* on log(*price*)? Initially, the effect appears to be strange. Because the coefficient on *rooms* is negative and the coefficient on $rooms^2$ is positive, this equation literally implies that, at low values of *rooms*, an additional room has a *negative* effect on log(*price*). At some point, the effect becomes positive, and the quadratic shape means that the semi-elasticity of *price* with respect to *rooms* is increasing as *rooms* increases. This situation is shown in Figure 6.2.

We obtain the turnaround value of *rooms* using equation (6.13) (even though $\hat{\beta}_1$ is negative and $\hat{\beta}_2$ is positive). The absolute value of the coefficient on *rooms*, .545, divided by twice the coefficient on $rooms^2$, .062, gives $rooms^* = .545/[2(.062)] \approx 4.4$; this point is labeled in Figure 6.2.

Do we really believe that starting at three rooms and increasing to four rooms actually reduces a house's expected value? Probably not. It turns out that only five of the 506 communities in the sample

FIGURE 6.2 $\widehat{\log(price)}$ as a quadratic function of *rooms*.



have houses averaging 4.4 rooms or less, about 1% of the sample. This is so small that the quadratic to the left of 4.4 can, for practical purposes, be ignored. To the right of 4.4, we see that adding another room has an increasing effect on the percentage change in price:

$$\Delta\widehat{\log(price)} \approx \{[-.545 + 2(.062)]rooms\}\Delta rooms$$

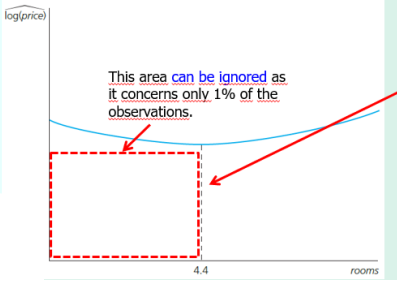
and so

$$\begin{aligned}\%\Delta\widehat{price} &\approx 100\{[-.545 + 2(.062)]rooms\}\Delta rooms \\ &= (-54.5 + 12.4\ rooms)\Delta rooms.\end{aligned}$$

Thus, an increase in *rooms* from, say, five to six increases price by about $-54.5 + 12.4(5) = 7.5\%$; the increase from six to seven increases price by roughly $-54.5 + 12.4(6) = 19.9\%$. This is a very strong increasing effect.

The strong increasing effect of rooms on log(*price*) in this example illustrates an important lesson: one cannot simply look at the coefficient on the quadratic term—in this case, .062—and declare that it is too small to bother with, based only on its magnitude. In many applications with quadratics the coefficient on the squared variable has one or more zeros after the decimal point: after all, this coefficient measures how the slope is changing as x (*rooms*) changes. A seemingly small coefficient can have practically important consequences, as we just saw. As a general rule, one must compute the partial effect and see how it varies with x to determine if the quadratic term is practically important. In doing so, it is useful to compare the changing slope implied by the quadratic model with the constant slope obtained from the model with only a linear term. If we drop $rooms^2$ from the equation, the coefficient on *rooms* becomes about .255, which implies that each additional room— starting from any number of rooms—increases median price by about 25.5%. This is very different from the quadratic model, where the effect becomes 25.5% at *rooms* = 6.45 but changes rapidly as *rooms* gets smaller or larger. For example, at *rooms* = 7, the return to the next room is about 32.3%.

- **Calculation of the turnaround point**



Turnaround point:

$$x^* = \left|\frac{-.545}{2(.062)}\right| \approx 4.4$$

Increase rooms from 5 to 6:

$-.545 + .124(5) = +7.5\%$ *price*

Increase rooms from 6 to 7:

$-.545 + .124(6) = +19.9\%$ *price*

- Chap 6, C2

C2 Use the data in WAGE1 for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

C6.2 (i) The estimated equation is

$$\widehat{\log(wage)} = \underset{(.106)}{.128} + \underset{(.0075)}{.0904}\, educ + \underset{(.0052)}{.0410}\, exper - \underset{(.000116)}{.000714}\, exper^2$$

$n = 526,\ R^2 = .300,\ \overline{R}^2 = .296.$

(ii) The t statistic on $exper^2$ is about –6.16, which has a p-value of essentially zero. So $exper^2$ is significant at the 1% level (and much smaller significance levels).

(iii) To estimate the return to the fifth year of experience, we start at *exper* = 4 and increase *exper* by one, so $\Delta exper = 1$:

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)4] \approx 3.53\%.$$

Similarly, for the 20th year of experience,

$$\%\Delta\widehat{wage} \approx 100[.0410 - 2(.000714)19] \approx 1.39\%.$$

(iv) The turnaround point is about .041/[2(.000714)] ≈ 28.7 years of experience. In the sample, there are 121 people with at least 29 years of experience. This is a fairly sizeable fraction of the sample.

- Chap 6, C4

C4 Use the data in GPA2 for this exercise.

(i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

where *hsize* is the size of the graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

(iii) Is this analysis representative of the academic performance of *all* high school seniors? Explain.

(iv) Find the estimated optimal high school size, using log(*sat*) as the dependent variable. Is it much different from what you obtained in part (ii)?

C6.4 (i) The estimated equation is

$$\widehat{sat} = \underset{}{997.98} + \underset{(3.99)}{19.81}\, hsize - \underset{(0.55)}{2.13}\, hsize^2$$
$$\underset{(6.20)}{}$$

$n = 4{,}137, \quad R^2 = .0076.$

The quadratic term is very statistically significant, with t statistic ≈ -3.87.

(ii) We want the value of *hsize*, say *hsize**, where $\widehat{sat}$ reaches its maximum. This is the turning point in the parabola, which we calculate as *hsize** = 19.81/[2(2.13)] ≈ 4.65. Since *hsize* is in 100s, this means 465 students is the "optimal" class size. Of course, the very small R-squared shows that class size explains only a tiny amount of the variation in SAT score.

- 討論重點:

(1) 如何判斷是否該有二次項?

(2) 有了二次項該如何解讀?

- Chap 9, Slides p.2-3 (RESET)

- **Tests for functional form misspecification**
  - One can always test whether explanatory should appear as squares or higher order terms by testing whether such terms can be excluded
  - Otherwise, one can use general specification tests such as RESET
- **Regression specification error test (RESET)**
  - The idea of RESET is to include squares and possibly higher order fitted values in the regression (similarly to the reduced White test)

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

Test for the exclusion of these terms. If they cannot be excluded, this is evidence for omitted higher order terms and interactions, i.e. for misspecification of functional form.

- **Example: Housing price equation**

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u$$

$$\Rightarrow \; F_{2,(88-3-2-1)} = 4.67, p - value = .012$$

Evidence for misspecification

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

$$\Rightarrow \; F_{2,(88-3-1-2)} = 2.56, p - value = .084$$

Less evidence for misspecification

- **Discussion**
  - One may also include higher order terms, which implies complicated interactions and higher order terms of all explanatory variables
  - RESET provides little guidance as to where misspecification comes from

EXAMPLE 9.2 **Housing Price Equation**

We estimate two models for housing prices. The first one has all variables in level form:

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqrft + \beta_3 bdrms + u. \quad [9.4]$$

The second one uses the logarithms of all variables except *bdrms*:

$$lprice = \beta_0 + \beta_1 llotsize + \beta_2 lsqrft + \beta_3 bdrms + u. \quad [9.5]$$

Using $n = 88$ houses in HPRICE1, the RESET statistic for equation (9.4) turns out to be 4.67; this is the value of an $F_{2,82}$ random variable ($n = 88, k = 3$), and the associated p-value is .012. This is evidence of functional form misspecification in (9.4).

The RESET statistic in (9.5) is 2.56, with p-value = .084. Thus, we do not reject (9.5) at the 5% significance level (although we would at the 10% level). On the basis of RESET, the log-log model in (9.5) is preferred.

討論重點:

(1) RESET 可以告訴我們什麼?

(2) 它如何操作? 如何解讀 F 統計量?