Notes: Lecture 3

So far we assumed that the relationship between the response variable and the predictors is linear. Sometimes we risk to fit a straight line to data that do not follow a straight relationship (this can also be cause of non-normally distributed residuals as well as heteroscedasticity). Then let us start to move beyond the assumption of linearity, first investigating cases in which the regression function is a nonlinear function of the X's but is a linear function of the unknown parameters (the Betas).

1) The slope of X depends on the value of X: Quadratic model

use "\...caschool.dta"

corr testscr avginc

twoway (scatter testscr avginc) (lfit testscr avginc)

reg testscr avginc

predict r, resid

kdensity r, norm

rvfplot, yline(0)

avplots

avplots, mlabel(county)

lvr2plot

lvr2plot, mlabel(county)

estat imtest, white → this test detects non-linear heteroscedasticity

estat hottest → this test detects linear heteroscedasticity

acprplot avginc, lowess lsopts(bwidth(1))

twoway (scatter testscr avginc) (lfit testscr avginc) (qfit testscr avginc) (lowess testscr avginc)

The quadratic regression function seems to fit the data better than the linear one. Imagine to draw a curve that fits the points of the scatterplot. This curve would be steeper for low values of district income, then would be flatter as district income gets higher.

gen avginc2 = avginc^2

reg testscr avginc avginc2

predict rr, resid

kdensity rr, norm

The significant coefficient for income^2 formally rejects the hypothesis that the relationship between income and test scores is linear

rvfplot, yline(0)

avplots

lvr2plot

estat imtest, white

estat hettest

acprplot avginc2, lowess lsopts(bwidth(1))

*using margins

reg testscr c.avginc##c.avginc

margins, dydx(avginc) at (avginc = (10 40 46 55))

margins, dydx(avginc) at (avginc = (5 (10) 55))

margins, at (avginc = (5 (1) 55))

marginsplot

2) linear-log model

An alternative to use a quadratic relationship is using the natural logarithm of X. This is sometimes called a linear-log model (given that the X is logged): Y=ln(X): the logarithmic function is steeper for small than for large values of X, it is only defined for X>0, it is positive for X>1 (equal to 0 when X=1) and has slope 1/X.

scatter testscr avginc

*to show the effect of using a log scale for avginc:

scatter testscr avginc, xscale(log)

gen lnavginc = log(avginc)

reg testscr lnavginc

A 1% change in X is associated with a change in Y of 0.01*Beta. In our case, then, a 1% increase in income is associated with an increase in test scores of 0.01*36.42=0.36 points.

di log(10)

2.3025851

di log(11)

2.3978953

di log(40)

3.6888795

di log(41)

3.7135721

lincom (_b[_cons]+ _b[lnavginc]*2.3978953)-(_b[_cons]+ _b[lnavginc]*2.3025851)

lincom (_b[_cons]+ _b[lnavginc]*3.7135721)-(_b[_cons]+ _b[lnavginc]*3.6888795)

3) Log-linear model

use "\…\nations.dta"

graph matrix gnpcap school2 school1, half

We want to explain the gnpcap using the school1 and school2 variables → the relationship seems non-linear (especially for school2)

reg gnpcap school2 school1

acprplot school2 , lowess

acprplot school1, lowess

Two options: A) add a quadratic term for school2 (and for school1); B) transform the dependent variable (which is skewed to the right)

kdensity gnpcap, normal

sktest gnpcap

Let us try to transform gnpcap to make it more normally distributed. Potential transformations include taking the log, the square root or raising the variable to a power. Selecting the appropriate transformation is somewhat of an art.

ladder gnpcap (look for the transformation with the smallest chi-square)

gladder gnpcap

generate lggnp=log(gnpcap)

label variable lggnp "log of gnpcap"

kdensity lggnp, normal

hist lggnp, normal

sktest lggnp

graph matrix lggnp school2 school1, half

regress lggnp school2 school1

In the log-linear model, a one-unit change in X (deltaX=1) is associated with a 100*Beta% change in Y. Translated into percentages, a unit change in X is associated with a 100*b1% change in Y.

acprplot school2 , lowess

acprplot school1, lowess

less deviation from nonlinearity than before

lincom (_b[_cons]+ _b[school2]*60)-(_b[_cons]+ _b[school2]*40)

lincom (_b[_cons]+ _b[school2]*40)-(_b[_cons]+ _b[school2]*20)

In this case the relationship between the dependent variable and the independent variables (after the transformation) is linear.