**Exercise Chapter 3**

The Stata data file ‘Labour\_Force\_SA\_SALDRU\_1993’ has the data used in Chapters 2 and 3.

1. Using this data run the earnings function as a linear, semi-log, and double-log function and derive the Mincerian return to education for each of the specification.

|  |
| --- |
| . /\*  > Exercise Chapter 3  > Using this data run the earnings function as a linear, semi-log, and double-log function and derive the Mincerian return to education for each of the specifications.  > \*/  .  . gen ln\_educ=log(educ)  (4073 missing values generated)  . egen educ\_mean=mean(educ)  . egen wphy\_mean=mean(wphy)  .  . /\*The Mincerian return to education is defined as the percentage change of earning for one additional year of education\*/  .  . reg wphy educ  Source | SS df MS Number of obs = 6968  -------------+------------------------------ F( 1, 6966) = 393.86  Model | 151135.02 1 151135.02 Prob > F = 0.0000  Residual | 2673037.95 6966 383.726379 R-squared = 0.0535  -------------+------------------------------ Adj R-squared = 0.0534  Total | 2824172.97 6967 405.364285 Root MSE = 19.589  ------------------------------------------------------------------------------  wphy | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  educ | 1.081474 .0544934 19.85 0.000 .97465 1.188297  \_cons | .2264038 .4974665 0.46 0.649 -.7487821 1.20159  ------------------------------------------------------------------------------  . predict wphy\_res, resid  (20464 missing values generated)  . histogram wphy\_res, normal  (bin=38, start=-14.215448, width=20.249003)  . gen ror\_linear=\_b[educ]/wphy\_mean if e(sample)==1  (20464 missing values generated)  . sum ror\_linear  Variable | Obs Mean Std. Dev. Min Max  -------------+--------------------------------------------------------  ror\_linear | 6968 .1210081 0 .1210081 .1210081  . reg logwphy educ  Source | SS df MS Number of obs = 6968  -------------+------------------------------ F( 1, 6966) = 2680.34  Model | 2368.42412 1 2368.42412 Prob > F = 0.0000  Residual | 6155.34858 6966 .883627416 R-squared = 0.2779  -------------+------------------------------ Adj R-squared = 0.2778  Total | 8523.77271 6967 1.22344951 Root MSE = .94001  ------------------------------------------------------------------------------  logwphy | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  educ | .1353827 .002615 51.77 0.000 .1302565 .1405088  \_cons | .4581331 .0238719 19.19 0.000 .4113368 .5049294  ------------------------------------------------------------------------------  . predict logwphy\_res, resid  (20464 missing values generated)  . histogram logwphy\_res, normal  (bin=38, start=-4.4470341, width=.23780323)  . gen ror\_semi=\_b[educ] if e(sample)==1  (20464 missing values generated)  . reg logwphy ln\_educ  Source | SS df MS Number of obs = 6142  -------------+------------------------------ F( 1, 6140) = 1980.53  Model | 1711.50261 1 1711.50261 Prob > F = 0.0000  Residual | 5305.97068 6140 .864164606 R-squared = 0.2439  -------------+------------------------------ Adj R-squared = 0.2438  Total | 7017.47329 6141 1.14272485 Root MSE = .9296  ------------------------------------------------------------------------------  logwphy | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  ln\_educ | 1.123521 .0252459 44.50 0.000 1.07403 1.173011  \_cons | -.7386552 .0548523 -13.47 0.000 -.846185 -.6311254  ------------------------------------------------------------------------------  . predict ln\_ln\_wphy\_res, resid  (21290 missing values generated)  . histogram ln\_ln\_wphy\_res,normal  (bin=37, start=-4.4174975, width=.24086408)  . exit |

|  |  |  |
| --- | --- | --- |
| Residuals from linear | Residuals from semi-log | Residuals from double log |
|  |  |  |

2. Which do you prefer and why?

Viewing the residuals from these specifications is the first aspect of the answer. It is clear the linear specification is very far from producing residuals remotely close to being normal. Both the semi-logarithmic and double logarithmic specification are much closer to a normal distribution. Next you need to consider the economic meaning of the different specifications. Will an additional year of education increase earnings by the same absolute amount whatever the level of education? That does not seem likely. Alternatively, will an additional year of education have the same percentage effect of earnings? That is what the semi-logarithmic specification says and that seems much more plausible (and is consistent with the residual diagnostic reported above). To test for the double logarithmic specification your need to run a more general specification (see below). The double logarithmic imposes concavity of the relationship between education and its return. It is clear that the relationship in the data is convex, so we do not prefer the double logarithmic specification.

|  |
| --- |
| .  . reg logwphy educ educ\_sq  Source | SS df MS Number of obs = 6968  -------------+------------------------------ F( 2, 6965) = 1640.15  Model | 2729.10459 2 1364.5523 Prob > F = 0.0000  Residual | 5794.66811 6965 .831969578 R-squared = 0.3202  -------------+------------------------------ Adj R-squared = 0.3200  Total | 8523.77271 6967 1.22344951 Root MSE = .91212  ------------------------------------------------------------------------------  logwphy | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  educ | -.0308544 .0083775 -3.68 0.000 -.0472769 -.0144319  educ\_sq | .0117721 .0005654 20.82 0.000 .0106637 .0128804  \_cons | .8151844 .0288205 28.28 0.000 .7586875 .8716814  ------------------------------------------------------------------------------ |

The Stata data file ‘Macro\_1980\_2000\_PENN61.dta’ has the macro data used in this chapter. Using this data estimate the following models and answer the questions below:

|  |
| --- |
|  |
| .  . use "$book\_2015\Macro\_1980\_2000\_PENN61.dta", clear  .  . gen time=0 if year==1980  (82 missing values generated)  . replace time=1 if year==2000  (82 real changes made)  .  . gen rgdpch\_tot=rgdpch\*pop  . gen ln\_rgdpch\_tot=log(rgdpch\_tot)  .  . gen ln\_kap\_ss=log(kap\_ss)  . gen ln\_pop=log(pop)  .  . reg ln\_rgdpch\_tot ln\_kap\_ss ln\_pop ,robust  Linear regression Number of obs = 164  F( 2, 161) = 2051.52  Prob > F = 0.0000  R-squared = 0.9681  Root MSE = .3435  ------------------------------------------------------------------------------  | Robust  ln\_rgdpch\_~t | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  ln\_kap\_ss | .7026651 .0203784 34.48 0.000 .6624217 .7429085  ln\_pop | .2885614 .0236724 12.19 0.000 .2418129 .3353099  \_cons | 2.102042 .3768596 5.58 0.000 1.357817 2.846267  ------------------------------------------------------------------------------  . test ln\_kap\_ss +ln\_pop=1  ( 1) ln\_kap\_ss + ln\_pop = 1  F( 1, 161) = 0.28  Prob > F = 0.5983  . reg lrgdpch lkp ,robust  Linear regression Number of obs = 164  F( 1, 162) = 1192.40  Prob > F = 0.0000  R-squared = 0.9089  Root MSE = .34271  ------------------------------------------------------------------------------  | Robust  lrgdpch | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  lkp | .7026283 .0203476 34.53 0.000 .6624475 .7428091  \_cons | 1.959049 .2070911 9.46 0.000 1.550103 2.367995  ------------------------------------------------------------------------------  . reg lrgdpch lkp ln\_pop,robust  Linear regression Number of obs = 164  F( 2, 161) = 621.42  Prob > F = 0.0000  R-squared = 0.9090  Root MSE = .3435  ------------------------------------------------------------------------------  | Robust  lrgdpch | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  lkp | .7026652 .0203784 34.48 0.000 .6624218 .7429085  ln\_pop | -.0087735 .0166211 -0.53 0.598 -.0415969 .0240499  \_cons | 2.102042 .3768596 5.58 0.000 1.357816 2.846267  ------------------------------------------------------------------------------ |

3. Are the data consistent with constant returns to scale?

The data are clearly consistent with constant returns to scale. In the boxed Table above we carried out two equivalent tests for constant returns to scale.

4. Is the share of capital consistent with the national accounts in any of these regressions?

According to the national accounts data, the share of capital in GDP is around . Using the macro data, the coefficients on and suggest that . Meanwhile, Hall and Jones data yields . Both values are substantially higher than that from the national accounts.

5. If not suggest reasons for the parameter estimate you observe.

It could be due to the exclusion of human capital (measured by years of education in this case). That you can test. In fact, inclusion of years of education does not greatly affect the parameter estimate on physical capital. One possibility is that unobservables in the residuals are positively correlated with capital. The next question enables you to test if that is at least part of the explanation.

6. Using the data for 1980 and 2000, create a cross section of differenced variables and re-estimate this equation in differences

|  |
| --- |
| .  . tsset nwbcode year  panel variable: nwbcode (strongly balanced)  time variable: year, 1980 to 2000, but with gaps  delta: 1 unit  . reg d20\_lrgdpch d20\_lkp  Source | SS df MS Number of obs = 82  -------------+------------------------------ F( 1, 80) = 96.77  Model | 5.40953702 1 5.40953702 Prob > F = 0.0000  Residual | 4.47227136 80 .055903392 R-squared = 0.5474  -------------+------------------------------ Adj R-squared = 0.5418  Total | 9.88180838 81 .121997634 Root MSE = .23644  ------------------------------------------------------------------------------  d20\_lrgdpch | Coef. Std. Err. t P>|t| [95% Conf. Interval]  -------------+----------------------------------------------------------------  d20\_lkp | .5443869 .055341 9.84 0.000 .4342548 .6545189  \_cons | .0587519 .0339368 1.73 0.087 -.0087845 .1262883  ------------------------------------------------------------------------------ |

7. Comment on the new point estimates and why they differ from those used in your answers to questions 3 and 4.

What this specification does is to remove the time invariant unobservables from the data (see below for details). If the unobserved is time-invariant we are allowing for its possible correlation with the capital variable and we do obtain a lower parameter estimate of . This is because of the positive relationship between capital per labour and the unobserved term, . Note that the point estimate remains far above the value of 0.3 assumed in the Hall and Jones decomposition.

Note well that the problem with all these specifications are from the zero conditional mean assumption. If does not hold, the parameter estimates will be biased and we cannot interpret the results as being causal.

More detailed notes:

Cobb-Douglas human capital augmented production function

(3.1)

(3.2)

The level equation

(3.3)

If the unobserved total factor productivity, , is correlated with the capital-labour ratio, the parameter estimate will be biased in the same direction as .

Write out the two level equations for periods 1 and 0

(3.4)

(3.5)

Difference equations [ (3.4) – (3.5) ]

If the unobserved is *time-invariant* (i.e. ), we can remove this unobserved factor by simply differencing the equations. Then, the parameter estimate from this difference equation is unbiased. The estimating equation becomes,

(3.6)