CAU Kiel
Chair of Econometrics

PC Tutorial - Econometrics I
Winter term 2023/2024

Problem Set I: Linear Regression

Suppose you have to quantify the causal effect of the number of children on the wage of married women. To this end, use the data set **mroz.dta** in STATA.

1. Start by estimating the simple linear regression

$$\log(wage) = \beta_0 + \beta_1 kidstotal + u,$$

   where *kidstotal* is the total number of (non-adult) kids. Use heteroscedasticity consistent standard errors. (Note: you have to define this variable as the sum of $kidslt6$ and $kidsge6$.)

   (a) What is the estimated effect of another child on the wage? Why is it preferable to have $\log(wage)$ on the LHS instead of *wage*?

   (b) Is the effect quantitatively relevant and statistically significant?

   (c) Report a 90% confidence interval for β_1. Interpret.

   (d) Is *kidstotal* an important driver of the female wage?

   (e) Which model deficiencies may invalidate the interpretation of β_1 as a causal effect?

2. Now augment the regression with the following control variables: *exper*, $exper^2$, *educ*, and *age*. Re-estimate.

   (a) Explain for each of the control variables why it makes sense to include it.

   (b) What is the estimated effect of another child on the wage? What has changed? Why?

   (c) Is the effect quantitatively relevant and statistically significant?

   (d) Can you now interpret β_1 as a causal effect?

3. In the next step split *kidstotal* into $kidslt6$ and $kidsge6$ and estimate

$$\log(wage) = \beta_0 + \beta_1 kidslt6 + \beta_2 kidsge6 + \beta_3 exper + \beta_4 exper^2 + \beta_5 educ + \beta_6 age + u.$$

   (a) Explain why it may make sense to include $kidslt6$ and $kidsge6$ separately instead of just *kidstotal*.

(b) What is the estimated effect of another (young or old) child on the wage?

(c) Are the effects quantitatively relevant and statistically significant?

(d) Perform an F-test of the hypothesis of joint significance for kidslt6 and kidsge6.

(e) Perform an LM-test of the hypothesis of joint significance for kidslt6 and kidsge6. Here (and only here) assume the disturbances are homoscedastic.

(f) Perform an F-test of the hypothesis that the age of the kids does not play a role, i.e., $\beta_1 = \beta_2$.

4. (*For self study.*) Use the data in CORNWELL.dta (from Cornwell and Trumball, 1994) to estimate a model of county-level crime rates, using the year 1987 only.

   (a) Using logarithms of all variables, estimate a model relating the crime rate (*crmrte*) to the deterrent variables probability of arrest (*prbarr*), probability of conviction (*prbconv*), probability of prison sentence (*prbpris*), and average sentence (*avgsen*, in days). Interpret the coefficients. Are the effects quantitatively relevant and statistically significant?

   (b) Add log(*crmrte*) for 1986 as an additional explanatory variable, and comment on how estimates elasticities differ from part a.

   (c) Add the various wage variables as regressors (in logs). Why could they be relevant? Compute the F statistic for joint significance of all the wage variables.

   (d) Redo part c, but make the test robust to heteroscedasticity of unknown form.

5. (*For self study.*) Use the data in NLS80.dta (from Blackburn and Neumark, 1992). Assume the model

$$\begin{aligned} \log(wage) &= \beta_0 + \beta_1 exper + \beta_2 tenure + \beta_3 married + \beta_4 south + \ldots \\ &\quad \ldots + \beta_5 urban + \beta_6 black + \beta_7 educ + \gamma abil + \nu. \end{aligned}$$

   (a) Use either kww or iq (two different test scores) as proxies for ability. Compare the estimated returns to education without a proxy for ability, with kww as proxy, with iq as proxy, and with both proxies. Interpret.

   (b) Include both kww and iq and test for joint significance.

   (c) Compare the equation without proxy for ability with the equation that includes both kww and iq. How does the estimated wage differential between nonblacks and blacks change? Try to explain.

   (d) Compute the new variables $kww0 = kww - \overline{kww}$, where $\overline{kww}$ is the average kww score in the sample. Include $kww0$, $educ$ and the interaction $educ \times kww0$ as regressors. Calculate the partial effect of another year in school for people with $kww = 20, 35, 50$. Also calculate the partial effect predicted for each individual and display its distribution with the help of a histogram (Stata command `histogram >varname<`). Interpret your results.