Data Science Practical
Case 6 - Academic year 2024-2025

What factors explain wages?

In the field of economics, it is very common that studies are employed that try to explain wages, which are sometimes called "return to education". Do they only depend on the (type of) job you are doing? Or are there other factors important? In particular, you could imagine that parents with a high education might be more likely to raise a child that receives a high education. This would in turn lead to better chances on the labor market and thus a larger wage. You could test whether such a relationship exists by performing a linear regression. At this point, you note that you only have mother's and father's years of education at your disposal. You could combine this information in one variable and run a simple linear regression between wage (or log wage?) and parents' years of education. If you do so, you make a certain assumption. Could you think of an easy way to test whether this assumption is plausible?

You believe (and hope!) that it is not just your parents' education that determines the wage you receive. You recognize some variables in the data set that express a certain degree of unobserved ability or expertness, and you are convinced that at least some of them should be included in the regression model. However, you remember the concept of collinearity and you want to make sure that your regression is not suffering from it. You extend your model with some new explanatory variables (neglect the binary variables for the moment) and use some measures and tests to show that the newly built model "outperforms" the previous regression(s) you did.

You realize you are still not completely satisfied with the model you built. At the beginning of this case, you namely hypothesized that your parents' education might directly influence the child's education which in turn might affect wages. This can be modeled in a rather simple way and you decide to do it. Make sure you clearly explain what you did and why you did it this way. It would be beneficial to the reader of your work if you could characterize the interpretation of the newly created coefficient by some simple examples. In addition, you wonder whether there are any non-linear relationships that should be taken into account. For example, does it make sense to include a variable such as tenure in squared form in the model?

Lastly, you start including your binary variables. In a lot of research, you see that especially these variables are of interest to the researcher. The goal of the research is then (for example) to find a gender differential: do men systematically earn more than women? You notice that such a variable "gender" is not present in your data set, but that there are several others that might be interesting to study. Is it enough to simply include the binary variables? Should we think of including interaction terms? Should we worry about entering a dummy trap? Make sure you explain every step you take in solid statistical and economic terms. A careful explanation of (the interpretation of) estimation results is vital. This also means that it is important that you check whether the underlying assumptions of the linear regression model are met.