Experiment Project

Due Date: June 18 (Wednesday in Class)

1. Please refer to the attached data set `chip2002.dta`, which is the data collected by the Chinese Household Income Project (CHIP) in year 2002. please use the techniques we have learned in class to complete following tasks:

   (a) Please refer to the `haven` package in `R` and find which `R` function you can use to read `chip2002.dta` into memory as `data.frame` object for further analysis.

   (b) Please formally establish and estimate a linear regression model to investigate the relationship between logarithm wage and years of schooling. Formally report your result. Besides, please provide the confidence interval for the estimated coefficients.

   (c) If we want to check whether there exist significant gender differences in wages, how can we formally establish econometric model? Please formally establish an econometric model and estimate it. Report your results and briefly discuss the results you obtained.

2. Pleas refer to the attached data set `HTV.dta`. The data set includes information on wages, education, parents' education, and several other variables for 1230 working men in 1991. Please complete following questions and formally demonstrate your discussion.

   (a) What is the range of the *educ* variable in the sample? What is the percentage of men who completed 12th grade but no higher grade?

   (b) Estimate the regression model

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

   and report the estimation results to discuss whether the OLS estimation $\hat{\beta}_1$ and $\hat{\beta}_2$ are significant or not.

   (c) Estimate the regression model

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

   by OLS. Test the null hypothesis that *educ* is linearly related to *abil* against the alternative that the relationship is quadratic.

   (d) For the regression model in (c), test $H_0 : \beta_1 = \beta_2$.

(e) Add the two college tuition variables ($tuit17$ and $tuit18$) to linear regression model in (c). Does it make any sense to make inference of the corresponding estimated coefficients? Please briefly provide your rationale and propose an approach if we still wish to investigate the relationship between college tuition and the highest grade ($educ$) completed by working men.