Second homework assignment

The file “synthetic_cps.RData” contains the original data (*cps*) and a synthetic version of the Current Population Survey (*syn.data*) which was also used in the video lectures. However, the synthetic datasets in this exercise contain only *m=5* replicates and the variables tax, race and social security payments (*ss*) are synthesized.

1) Compute the point and variance estimate for the average amount of taxes payed based on the original data and on the synthetic data (use the appropriate “combining rules” for your synthetic data for this and all the following exercises, i.e. analyze each replicate separately first and combine the results with the appropriate formulae as discussed in the course). Use your own code, i.e., don’t use the synthpop package to compute the results.

2.) Compute the point and variance estimate for the proportion of blacks (*race*==2) based on the original data and on the synthetic data (remember that the estimated variance for the estimated proportion $\hat{p}$ based on a sample of size *n* is $\hat{p}(1-\hat{p})/n$)

3.) Check whether you find indications for a gender wage gap by running a linear regression of income (on a logarithmic scale) on all other variables in the data and looking at the regression coefficient of the sex variable (for simplicity only include the main effects of all variables in your regression), i.e., run the following regression

$$\log(income) \sim \beta_0 + \beta_1 \cdot tax + \beta_2 \cdot csp + \beta_3 \cdot age + \beta_4 \cdot educ + \beta_5 \cdot marital + \beta_6 \cdot race + \beta_7 \cdot sex + \beta_8 \cdot ss + \varepsilon,$$

where all categorical variables should be turned into dummy variables (note that R will do this automatically, if your categorical variables are coded as *factors*).

Again, compute the regression coefficient based on the original data and based on the synthetic data.

4.) Is the effect of *sex* significant for both datasets? What is the p-value of the underlying t-test for both datasets?

5.) Bonus question: Do you think it is appropriate to conclude that there is a gender wage gap based on your results, i.e. that your estimated regression coefficient can be interpreted as the causal effect of being female on income?