Laboratory 1 AED

Practical Part (8 points)

The dataset you will be working with is used to find out which predictors (independent variables) have a statistically significant impact on the labor-force participation of married women (lfp).

Variables Description

Variable(s)	Description	Type
lfp	Labor-force participation of the married white woman	Categorical: 0/1
k5	Number of children younger than 6 years old	Positive integer
k618	Number of children aged 6-18	Positive integer
age	Age in years	Positive integer
wc	Wife`s college attendance	Categorical: 0/1
hc	Husband`s college attendance	Categorical: 0/1
lwg	Log expected wage rate for women in the labor force	Numerical
inc	Family income without the wife`s income	Numerical

1. Load and print the first 5 observations of the dataset. Discuss if you have to encode any predictors and carry out the encoding. See if you may need to drop some columns. **(0.5 points)**
2. Discuss what effects of given predictors, you would expect to see on labor-force participation of married women (lfp). Make a hypothesis for each predictor. **(0.5 points)**
3. Check the types of your data. Change the types as appropriate (if any categorical variable present change its type to category). **(0.5 points)**
4. Compile the table with summary statistics (min, max, med, etc.). The Variance, Skewness and Kurtosis are not computed by default with the *describe()* function in Python, therefore compute separately these three measure of variability and add them to the table. **(0.5 points)**
5. Examine and comment on the table and report if you see anything unusual in the statistics of your variables. If some observation values appear unusual or wrong explain your approach to dealing with these observations. For example dropping the observations with such values is one possibility, substitution is another, one can also just leave them as they are. Your task is to identify the correct approach and explain your decision. **(0.5 points)**
6. Check for any NaN values in the dataset. Substitute the NaN values with appropriate measures of central tendency (mean, median, or mode). **(0.5 points)**
7. Produce the histograms of all variables (except wc and hc) and comment on their distributions (for each variable separately). Notice any outliers, or fat tails. Put this into context knowing what your variables mean. So for example if you see a fat right tale for age then explain it something like “the fat tale indicates that our dataset contains older women disproportionately, this could have the following consequences ... ” **(1 point)**

1. Create box plots for all variables where you split by the lfp variable (make sure to adjust the number of axes). You will end up with a pair of box plots for each variable, make sure that all box plot pairs are presented in one and the same graph for comparison ease. Comment on each box plot pair and explain if you think the variable in question will have a statistically significant impact on lfp given the distributions presented in the box plots. noticing if the distributions are located higher for properties on the river versus those not on the river. **(1 point)**

2. Create the heatmap and add the correlation coefficients to it. What are the 5 strongest correlations that you see? Are their signs as you would have expected? Do you see any sign of multicollinearity? (remember a correlation coefficient between two predictors larger than 0.8 is a reason for concern, anything below is not). **(0.5 points)**

3. Run a Logit regression with all of the given predictors and comment on the effects of predictors and their coefficient signs, plus their statistical significance and whether they confirm the hypotheses you made in (2). **(1.5 points)**

4. Now run a Probit model. Are the coefficients very different from those found by Logit? Did you expect this? Why if yes and why if not? **(0.5 points)**

5. Comment on the goodness of fit of the two models. Which one seems to be the better model for these data? **(0.5 points)**

Theoretical Part (2 points)

By using the minimization technique of least squares, get to the values of β_0 and β_1 by solving the system of equations (1) and (2). **Show all the steps!**

To find $\hat{\beta}_0$ and $\hat{\beta}_1$ we need to minimize the sum of squared residuals:

$$\min \rightarrow \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$\min f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \quad \text{w.r.t.} \quad \hat{\beta}_0, \hat{\beta}_1$$

A function reached its minimum when its first derivative is equal to zero

F.O.C.

1. $\frac{\delta f}{\delta \hat{\beta}_0} = \frac{\delta f[\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_0} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

2. $\frac{\delta f}{\delta \hat{\beta}_1} = \frac{\delta f[\sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2]}{\delta \hat{\beta}_1} = 0 \quad \rightarrow \quad -2 \sum_{i=1}^{n} X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{[\sum_{i=1}^{n} [(X_i - \bar{X})(Y_i - \bar{Y})]}{\sum_{i=1}^{n} [(X_i - \bar{X})^2}$$