Homework 5

Introduction to Data Science (QR2)

due: **10/15/25**

Instructions: Please submit solutions on Populi. Only a knitted pdf of an `Rmarkdown` file will be accepted.

Across industrialized countries, it is a well-studied phenomenon that childless women are paid more on average than mothers. In this exercise, we use survey data to investigate how the structural aspects of jobs affect the wages of mothers relative to the wages of childless women.[1]

In this paper, the authors examine the association between the so-called *mother wage penalty* (i.e., the pay gap between mothers and non-mothers) and occupational characteristics. Three prominent explanations for the motherhood wage penalty–"stressing work-family conflict and job performance," "compensating differentials," and "employer discrimination"–provide hypotheses about the relationship between penalty changes and occupational characteristics. The authors use data from 16 waves of the National Longitudinal Survey of Youth to estimate the effects of five occupational characteristics on the mother wage penalty and to test these hypotheses.

This paper uses a type of data known as "panel data." Panel data consist of observations on the same people over time. In this example, we are going to analyze the same women over multiple years. When analyzing panel data, each time period is referred to as a *wave*, so here each year is a wave. **The most general form of model for working with panel data is the *two-way fixed effects model*, in which there is a fixed effect for each woman and for each wave.**

The data file is `yu2017sample.csv`. The names and descriptions of variables are:

Variable	*Description*
`PUBID`	ID of woman
`year`	Year of observation
`wage`	Hourly wage, in cents
`numChildren`	Number of children that the woman has (in this wave)
`age`	Age in years
`region`	Name of region (North East = 1, North Central = 2, South = 3, West = 4)
`urban`	Geographical classification (urban = 1, otherwise = 0)
`marstat`	Marital status
`educ`	Level of education
`school`	School enrollment (enrolled = `TRUE`, otherwise = `FALSE`)
`experience`	Experience since 14 years old, in days
`tenure`	Current job tenure, in years
`tenure2`	Current job tenure in years, squared
`fullTime`	Employment status (employed full-time = `TRUE`, otherwise = `FALSE`)
`firmSize`	Size of the firm
`multipleLocations`	Multiple locations indicator (firm with multiple locations = 1, otherwise = 0)
`unionized`	Job unionization status (job is unionized = 1, otherwise = 1)
`industry`	Job's industry type
`hazardous`	Hazard measure for the job (between 1 and 5)
`regularity`	Regularity measure for the job (between 1 and 5)

[1] The exercise is based on: Wei-hsin Yu and J anet Chen-Lan Kuo (2017) "The Motherhood Wage Penalty by Work Conditions: How Do Occupational Characteristics Hinder or Empower Mothers?" *American Sociological Review* 82(4): 744-769.

a. How many different women are in the data? How many observations per year? We will refer to each row as a "person-year observation" since the row contains data on a given person in a particular year. In a few sentences, describe one advantage and one disadvantage of using a contemporary cohort of women rather than an older cohort in estimating the predictors of the mother wage gap.

b. `numChildren` is the variable representing how many children the woman had at the time of an observation. Please provide a table that shows the proportion of observations by the number of children. Provide a brief substantive interpretation of the results.

c. Create a new indicator variable `isMother` that takes a value of 1 if the woman has at least one child in that year and a value of 0 otherwise. Tabulate the new variable. Briefly comment on the results.

d. Create a new variable called `logwage` that is the log of `wage`. Make two boxplots, one for `wage` and the other for `logwage`, as a function of educational level (`educ`). Compare the two boxplots and discuss the purpose of the log transformation.

e. In the same graph, plot the mean `logwage` against year for mothers, then for non-mothers in a different color or line type. Include a legend and a proper title. Make sure the figure and axes are clearly labeled. Give a brief interpretation of the results.

f. Run a regression using fixed effects for both *woman* and *year*. You should be sure to include variables for number of children (`numChildren`) and job characteristics (`fullTime`, `multipleLocations`, `unionized`, `industry`). Note: that you should *not* use the `isMother` variable you created earlier in this model. Report the coefficient of `numChildren`. Provide a brief substantive interpretation of this coefficient and the coefficients for any two other variables. (*Hint*: fixed effects means including the relevant factor variables in the regression model – see the bolded statement in the problem introduction).

g. **OPTIONAL stretch problem:** Add interactions between `numChildren` and `regularity` and between `numChildren` and `hazardous` to the model in the previous question. Report the five coefficients involving these variables. Interpret the interaction term for `numChildren` and `hazardous`. Can we interpret the effect of occupation characteristics on motherhood wage penalty as "causal"? Why or why not?