Mother Wage Penalty

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.

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?](https://doi.org/10.1177/0003122417712729)" *American Sociological Review* 82(4): 744-769.

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 each woman and for each wave.

The data file is yu2017sample.csv, which is a CSV file. The names and descriptions of variables are:

|  |  |
| --- | --- |
| Name | 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 2) |
| regularity | Regularity measure for the job (between 1 and 5) |
| competitiveness | Competitiveness measure for the job (between 1 and 5) |
| autonomy | Autonomy measure for the job (between 1 and 5) |
| teamwork | Teamwork requirements measure for the job (between 1 and 5) |

## Question 1

What years are included in the data? How many women are included, and how many person-years are included? Briefly comment on the results.

## Answer 1

Panel.data<-read.csv("yu2017sample.csv")  
##Number of study years  
waves<-sort(unique(Panel.data$year))  
length(waves)

## [1] 16

##Number of women in study  
cohort<-subset(Panel.data,subset = (Panel.data$year==waves[1]))  
length(cohort$X)

## [1] 374

##Person-years  
length(cohort$X)\*length(waves)

## [1] 5984

## Question 2

Create a new variable called logwage that is the log of wage. Explain why it might be important to use such a variable in the study at hand. Calculate and comment on the correlations between logwage, numChildren, experience, and at least two other continuous variables in the data, e.g., competitiveness, autonomy, teamwork.

## Answer 2

Panel.data$logwage<-log(Panel.data$wage)

## Apllication of log reduces the variation in the data for meaningful statistical inference to be made.

data<-subset(Panel.data,select = c(logwage,numChildren,experience,competitiveness,hazardous,regularity))  
cor(data,use="pairwise.complete.obs")

## logwage numChildren experience competitiveness hazardous  
## logwage 1.00000000 0.10316379 0.54428981 0.19060845 0.04252011  
## numChildren 0.10316379 1.00000000 0.28173364 -0.05432589 0.07090866  
## experience 0.54428981 0.28173364 1.00000000 0.10943345 0.04208356  
## competitiveness 0.19060845 -0.05432589 0.10943345 1.00000000 0.03969617  
## hazardous 0.04252011 0.07090866 0.04208356 0.03969617 1.00000000  
## regularity -0.17809811 -0.02625733 -0.11896347 0.17959709 0.23896171  
## regularity  
## logwage -0.17809811  
## numChildren -0.02625733  
## experience -0.11896347  
## competitiveness 0.17959709  
## hazardous 0.23896171  
## regularity 1.00000000

## The correlation between wages and expereience appears to be the most positive, which means among all the variables included in the study the more experience the person the higher the wages and vice versa.

## Question 3

Run a two-way fixed effects (2FE) regression (linear regression with unit and time fixed effects) where the outcome is the log wage and the predictor is the number of children that a woman has. Include fixed effects for each *woman* and each *year* (i.e., include factorized variables for year and PUBID in the regression model). What is the estimated coefficient on the variable numChildren? Provide a brief substantive interpretation of the coefficient.

## Answer 3

fit.fixed<-summary(lm(logwage~numChildren +factor(PUBID) +factor(year),data=Panel.data))  
fit.fixed$coefficients[1,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## 5.734167e+00 1.682148e-01 3.408836e+01 2.852790e-246

## The coefficient of the number of children for each woman in a given year is 5.734 and it is the increase in the number of wages by 5.734 cent when the number of children increase by one.

## Question 4

What is the standard error of the coefficient for numChildren? What is the value of the estimate divided by the standard error and what does that mean? If you have the null hypothesis that this coefficient is equal to zero, at what level of would you still be able to reject the null hypothesis? What is the meaning of rejecting the null hypothesis in this exercise? Based upon Question 3, are there other predictors we might want to include in this model? For what reasons might we want to include them?

## Answer 4

##The standard error is 0.1682

fit.fixed$coefficients[1,1]/ fit.fixed$coefficients[1,2]

## [1] 34.08836

##This is the value of the test statistics of the null hypothesis of that coefficient of the number of children is zero. ## At alpha level of 0.001, the null hypothesis. ## To reject the null hypothesis means that the coefficient of the number children is statistically significant or not equal to zero. ## There may be other predictors which when included in the model may affect the coefficients (confounding variables)

## Question 5

Keeping in fixed effects for *woman* and *year*, introduce the following variables in the regression model:

* Location (region, urban)
* Marital Status (marstat)
* Human Capital (educ, school, experience, tenure, tenure2)
* Job Characteristics (fullTime, firmSize, multipleLocations, unionized, industry)

Report the coefficient and standard error of numChildren now. Is the coefficient statistically significant? Provide a brief substantive interpretation of this coefficient and the coefficients for any two other variables.

## Answer 5

fit.fixed.location<-summary(lm(logwage~numChildren +region+urban+factor(PUBID) +factor(year),data=Panel.data))  
fit.fixed.location$coefficients[1,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## 5.716819e+00 1.673526e-01 3.416033e+01 5.852771e-247

## The coefficient is statistically significant since the p value is less than 0.05, implying that we reject the null hypothesis that the coefficient is not significant.

## The amount of hourly wages increases by 5.7168 cent when the number of children increase by 1.

fit.fixed.location$coefficients[2,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## -7.581207e-02 6.536442e-03 -1.159837e+01 5.580375e-31

## The coefficient is statistically significant since the p value is less than 0.05, implying that we reject the null hypothesis that the coefficient is not significant.

## The amount of hourly wages decreases by 7.5812 cents when they move from one region to another.

fit.fixed.location$coefficients[3,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## -0.009955987 0.010533682 -0.945157315 0.344592781

## The coefficient is not statistically significant since the p value is less than 0.05, implying that we accept the null hypothesis that the coefficient is not significant.

## The amount of wages had decrease by 0.01 cents when the mothers move from the urban regions.

fit.fixed.Maital<-summary(lm(logwage~numChildren +marstat +factor(PUBID) +factor(year),data=Panel.data))  
fit.fixed.Maital$coefficients[1,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## 5.842337e+00 1.665126e-01 3.508645e+01 9.375728e-260

## The coefficient is statistically significant since p value is less than 0.05

fit.fixed.human<-summary(lm(logwage~educ+school+experience+tenure+tenure2 +factor(PUBID) +factor(year),data=Panel.data))  
fit.fixed.human$coefficients[1,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## 6.3163675 0.1572658 40.1636333 0.0000000

## The coefficient is statistically significant since p value is less than 0.05

fit.fixed.job<-summary(lm(logwage~fullTime +firmSize+multipleLocations+unionized+industry +factor(PUBID) +factor(year),data=Panel.data))  
fit.fixed.job$coefficients[1,c(1:4)]

## Estimate Std. Error t value Pr(>|t|)   
## 5.737033e+00 2.085167e-01 2.751355e+01 4.214030e-161

## The coefficient is statistically significant since p value is less than 0.05

## Question 6

In order to replicate the original study, add interactions between numChildren and regularity and between numChildren and hazardous. Report the five coefficients involving these variables and their standard errors. Is it statistically significant? Interpret the interaction term for numChldren and hazardous. Based on this model, what change would you expect to see in the log wages of two otherwise identical mothers, if the first had one more child than the second? Can we interpret the effect of occupation characteristics on motherhood wage penalty as causal? Why or why not?

## Answer 6

fit.fixed.interact<-summary(lm(logwage~numChildren\*regularity+numChildren\*hazardous,data=Panel.data))  
fit.fixed.interact$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 7.45939829 0.038629333 193.101917 0.000000e+00  
## numChildren 0.14070767 0.034827824 4.040094 5.365162e-05  
## regularity -0.67428427 0.031687897 -21.278922 2.965419e-99  
## hazardous 0.10689508 0.010680100 10.008809 1.606333e-23  
## numChildren:regularity -0.05271296 0.028774542 -1.831930 6.697831e-02  
## numChildren:hazardous -0.01269134 0.008974541 -1.414149 1.573353e-01

## Base on the p-values, all of the five coefficients are statistically significant with the exception of the interaction between numchildren and hazardou whose p-value is greater than 0.05.