Empirical Project in Introductory Econometrics Course

Edwin Leuven

2023-01-31

Introduction

The aim of the project is to study the differences in real wages in 2019 US$ (rw) between immigrants and non-immigrants, and draw some conclusions as to what explains these differences. The issue of migrant wages features prominently in the United States, a country that receives a significant proportion of migrants from around the world. A major issue is that migrants contribute significantly to wage inequality in the United States. While some researchers argue that migrants do exacerbate wage inequality, other researchers opine that the observation only holds for unskilled migrants (Lin and Weiss 2019).

In light of these concerns, the objectives of this analysis are as follows;

a. To quantify the wage gap between immigrants and non-immigrants in the United States.
b. To investigate whether and how the wage gap varies by time since immigrants entered the US.
c. To identify possible drivers of the migrant wage gap using regression analysis.

Data

The data comes from the 2019 US current population survey (CPS). The data is available on this site, https://ceprdata.s3.amazonaws.com/data/cps/data/cepr_org_2019.zip

The data file comes in `Stata` format. I imported the data into R using the read_dta() function of the "haven" package.

```
my_df <- read_dta("cepr_org_2019.dta") %>%
  ## Remove columns that have no data
  janitor::remove_empty() %>%
  ## Remove columns that have constant values
  janitor::remove_constant() %>%

  tibble() %>%

  ## -1 indicates missing value
  mutate(

    across(

      .cols = everything(),

      .fns = ~ case_when(

        . == -1 ~ NA,


        TRUE ~ .
```

In the appendix below you will find a list of variables and description. Categorical variables are imported as factor variables with descriptive labels, and the file should be self explanatory.

The dataset contains many demographic and employment related variables that may explain wages. You should carefully go over the variable list to see what information is available and tabulate variables to see how they are coded.

Empirical Approach

I first do exploratory analysis to quantify the wage gap between migrant and non-migrant workers in the United States. specifically, I construct plots and calculate summary statistics.

Next, I run regressions to examine the factors that have a relationship with real wages between migran and non-migrant workers. In this analysis, I use nationality and the length that migrants have been in the United States as controls, over and above personal and household variables that may affect income. From theory and past empirical analysis, I find that the following factors have a significant relationship with wage differentials between migrants and non-immigrants.

I then run the following regression model.

$rw = \beta_0 + \beta_1 citizen + \beta_2 female + \beta_3 wbhaom + \beta_4 age + \beta_5 educ + \beta_6 cert + \beta_7 hourslw + \beta_8 faminc + \beta_9 docc03 + \beta_{10} arrived$

Results

Quantifying the Wage Gap Between Migrants and Non_Immigrants

I first start by running summary statistics regarding the earnings of migrants versus non-immigrants.

Table 1: Summary of Wages for US Citizens and Migrants

skim_type	forborn	Mean	SD	Q1	Median	Q3	Max
numeric	US	25.94918	19.59816	13.95	20	31.25000	392.3050
numeric	Foreign	25.31543	20.56011	13.00	18	29.81375	288.3333

Table 1 shows a notable disparity in wages between migrants and citizens. We can visualise this disparity. The Figure below shows that we fail to reject the null hypothesis that there is no wage disparity and accept the alternative hypothesis that there is indeed a significant wage disparity between migrants and US born residents at 1% significance level.

Which factors Have a Notable relationship with the Wage Gap Between Migrants and Non-Immigrants?

In this section, I use data visualization to explore factors that may have a significant relationship with wages for migrants and non-immigrants.

I start by doing a pairs plot for all the variables that I hypothesize have a relationship with the wages gap.

The data shows that there is a notable difference in the incomes of citizens and immigrants. Citizens have a higher mean and median income compared to immigrants.

The income of immigrants varies by the time of arrival. New arrivals receive substantially less income compared to other people. However, the income improves markedly after the second year. After the 5th year, there is a gradual decline in income.

The significant drivers of wage differential between migrants and non-migrants are;

- Foreign Born (forborn).
- The level of education (educ).

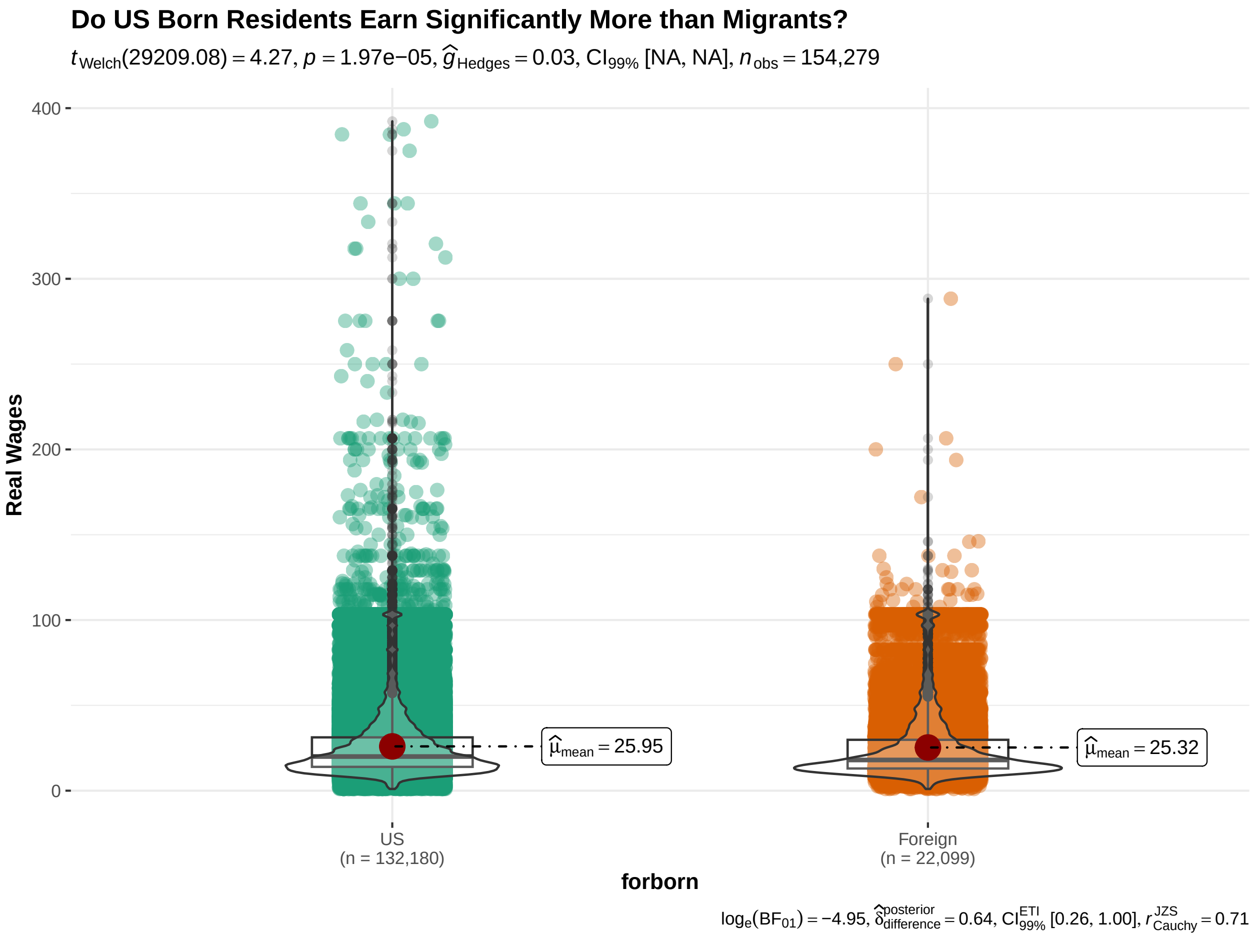
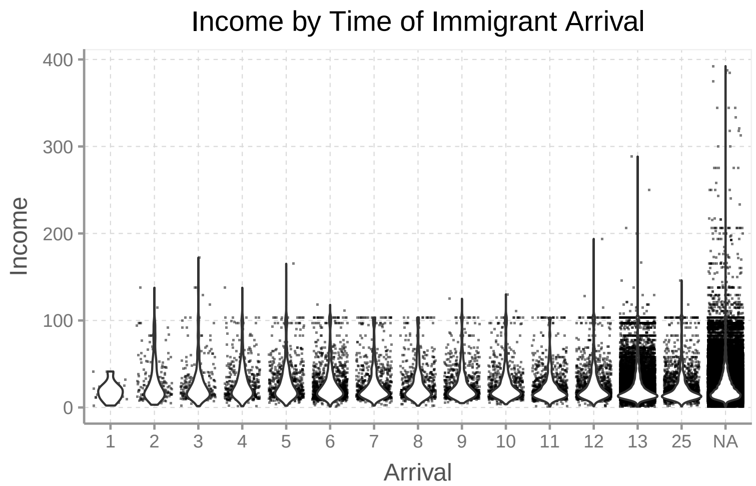
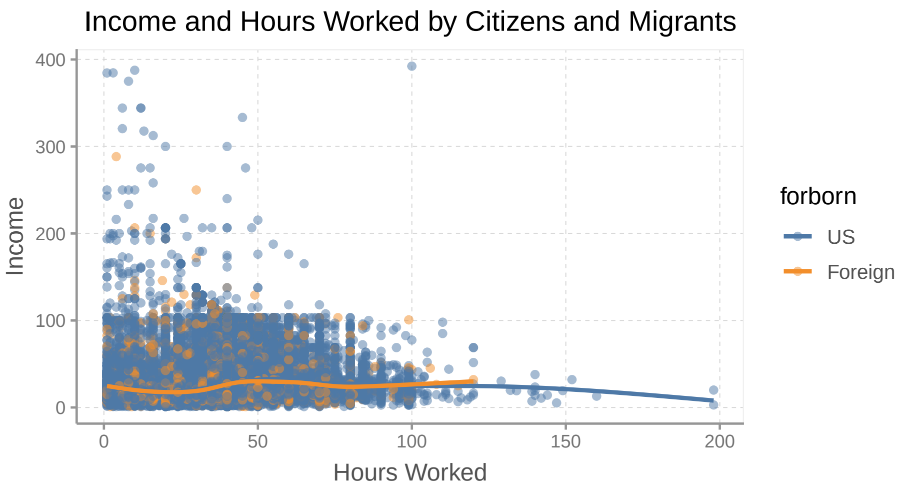


Figure 1: Wage Differential Between Migrants and Non-Immigrants

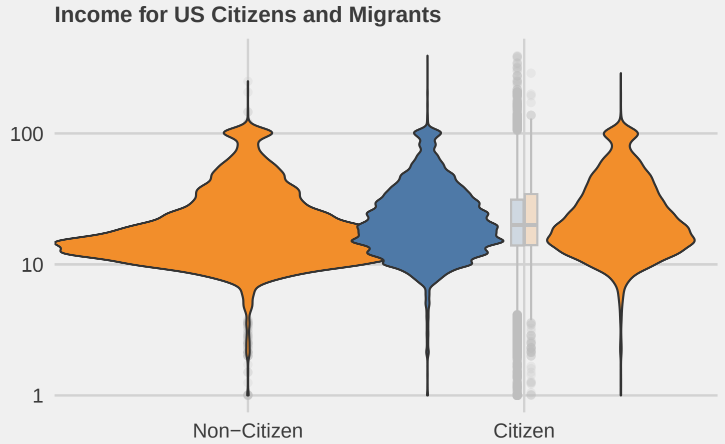
- Gender (female).
- Occupation, which may have high correlation with the level of education (docc03).
- Age, which may proxy experience (age).
- Race (wbhaom).
- Certification (wbhaom).
- Hours worked (hourslw).
- Family income (faminc).
- Time of arrival in the United States (arrived).


Income by Time of Immigrant Arrival

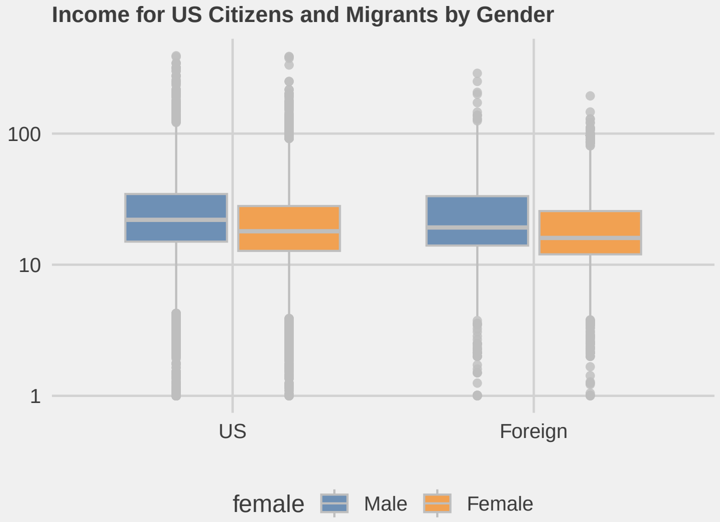
Data Source: 2019 US current population survey (CPS)


Income and Hours Worked by Citizens and Migrants

Data Source: 2019 US current population survey (CPS)


Income for US Citizens and Migrants

Data Source: 2019 US current population survey (CPS)


Income for US Citizens and Migrants by Gender

Data Source: 2019 US current population survey (CPS)

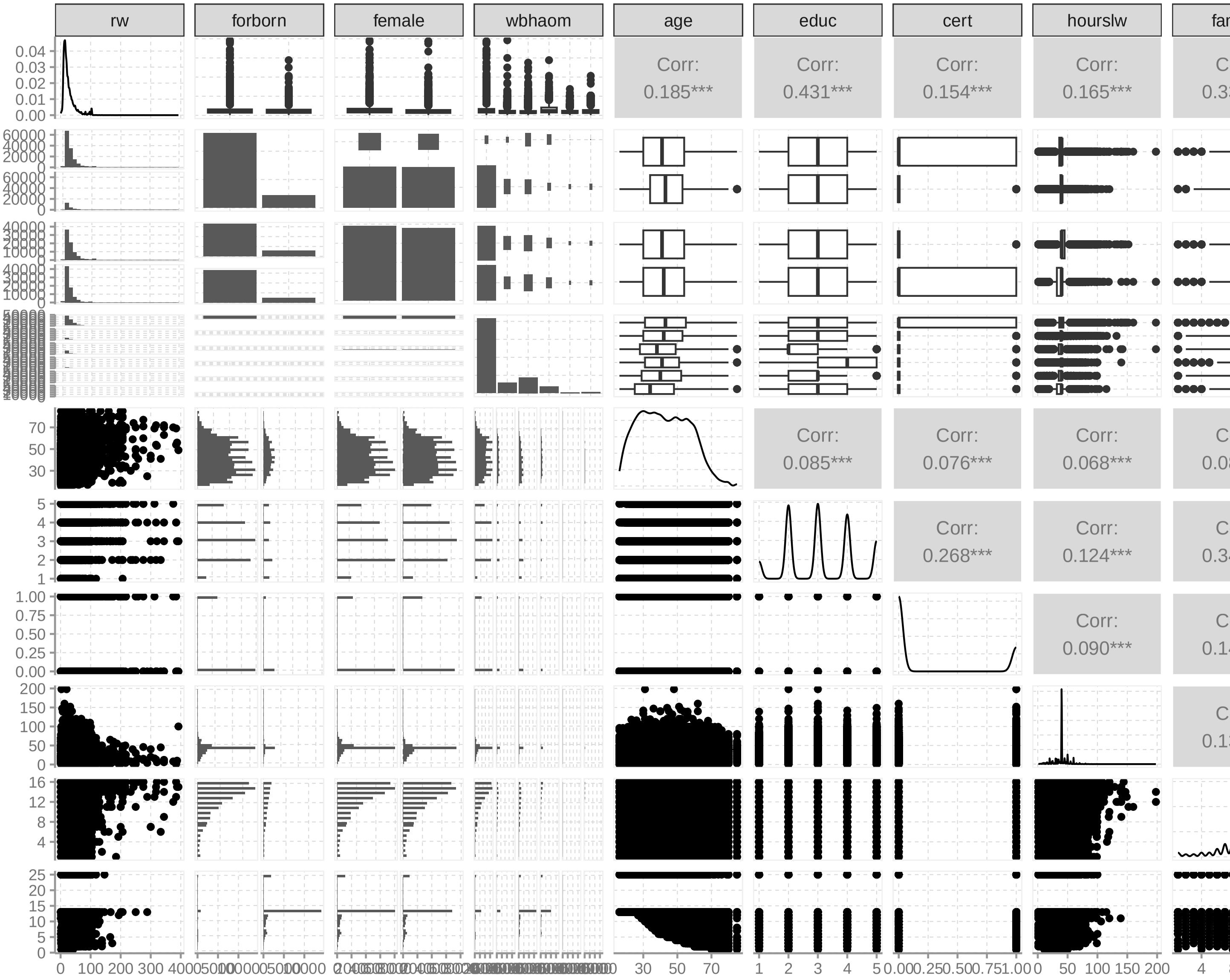
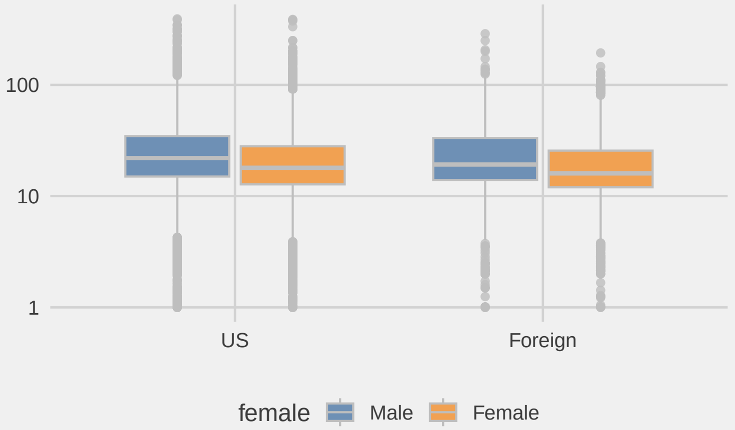


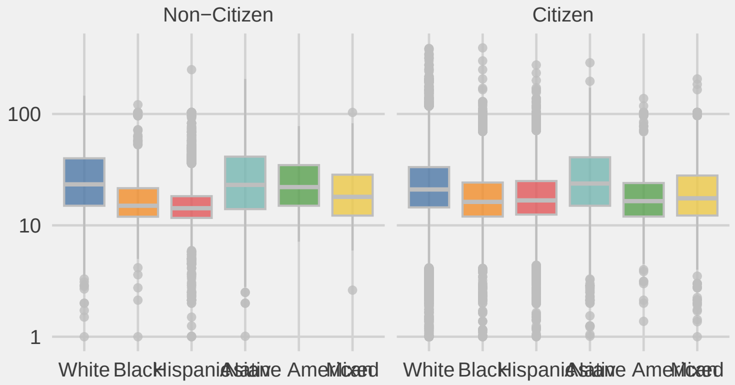
Figure 2: Pairs Plots

Income for US Citizens and Migrants by Gender



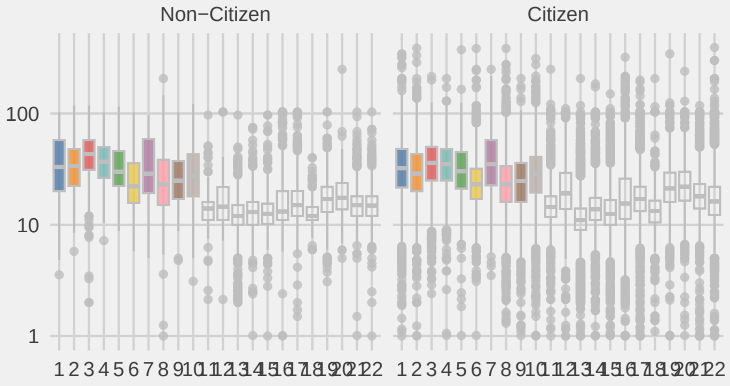
Data Source: 2019 US current population survey (CPS)

Income by Race for Citizens vs Immigrants



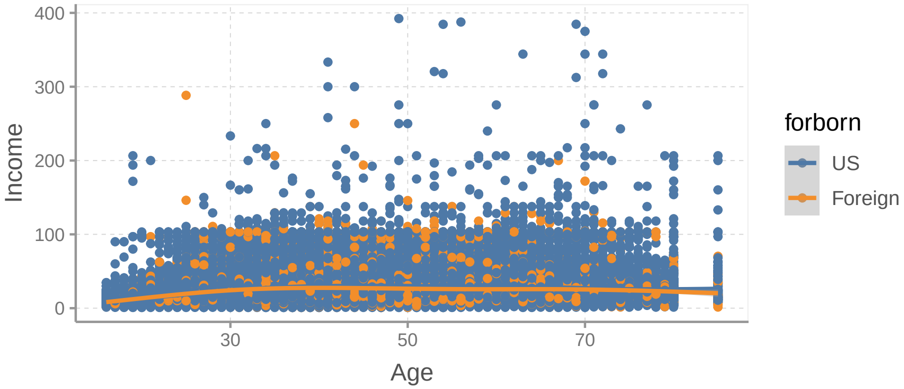
Data Source: 2019 US current population survey (CPS)

Income by Occupation: Citizens vs Immigrants



Data Source: 2019 US current population survey (CPS)

Income for Citizens vs Immigrants by Age Group



Data Source: 2019 US current population survey (CPS)

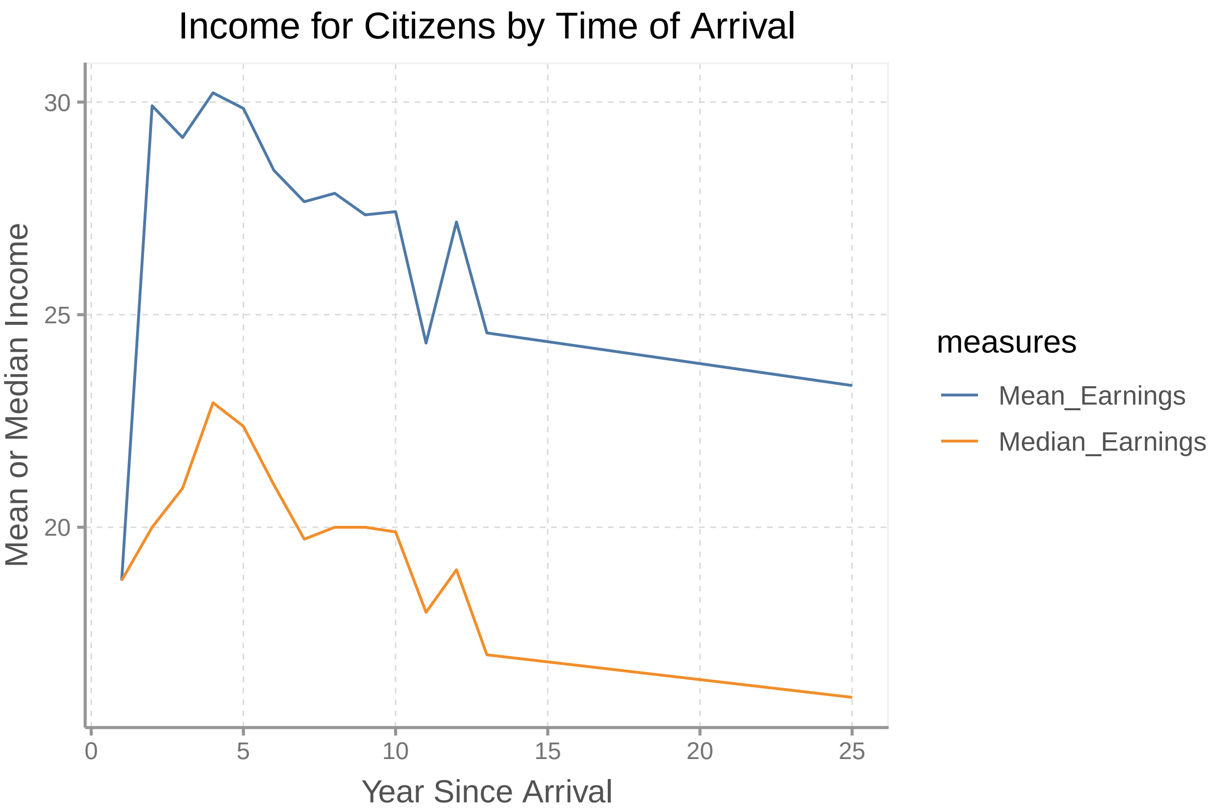
Regression Analysis

This section presents the findings from my *regression models* which look at

a. The immigrant wage gap and possible explanations
b. Whether and how the wage gap varies by time since immigrants entered the US.

In this analysis, I regress the income (rw) of individuals and several variables. The summary of regression results is in the table below. The analysis shows the following variables to be the significant drivers of wages.

- **Age**: This variable has a positive relationship with income. Meaning that older employees with greater experience tend to have more pay, ceteris paribus.
- **Education**: More educated employees have better pay compared to less educated employees, all else remaining the same.
- **Certification**: Employees with a professional certification eran more on average, all else remaing the same.
- **Race**: White people tend to have better pay compared to employees from other races hodling other factors constant.
- **Hours worked (hourslw)**: The more hours an employee puts in , the higher the average pay holding other factors constant.
- **Rural**: Employees in rural locations tend to get less pay.
- **Occupation**: Some occupations are more lucrative than others. Given that immigrants are likely to occupy unskilled occupations, it is likely that they have lower average salaries.
- **Year of arrival** in the united states: Very new arrivals in the US have notably lower pay compared to those that arrived earlier. The income rises after the first year and then declines gradually.



Data Source: 2019 US current population survey (CPS)

We start by running a simple model of real wages against the `forborn`, a variable that captures whether or not the person was born outside the United States.

```
##
## Call:
## lm(formula = rw ~ forborn, data = my_df)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24.95 -12.20  -6.08   5.30 366.36
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    25.94918    0.05429 477.953  < 2e-16 ***
## forbornForeign -0.63374    0.14345  -4.418 9.98e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.74 on 154277 degrees of freedom
## Multiple R-squared:  0.0001265,  Adjusted R-squared:  0.00012
## F-statistic: 19.52 on 1 and 154277 DF,  p-value: 9.977e-06
```

This simple model shows that foreign born US residents earn significantly less than the people born in the United States. However, the explanatory power of the model is too low. I add more variables in the next model.

```
##
## Call:
## lm(formula = rw ~ forborn + female + wbhaom + age + educ + cert +
##     hourslw + factor(arrived) + rural + docc03, data = my_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -49.749  -8.258  -1.701   4.377 250.282
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            11.106680   3.735473   2.973  0.00295 **
## forbornForeign          1.263709   0.392948   3.216  0.00130 **
## femaleFemale           -5.033763   0.238853 -21.075  < 2e-16 ***
## wbhaomBlack            -4.714249   0.449386 -10.490  < 2e-16 ***
## wbhaomHispanic         -2.627405   0.319123  -8.233  < 2e-16 ***
## wbhaomAsian            -0.589022   0.326496  -1.804  0.07123 .
## wbhaomNative American  -3.029283   3.174137  -0.954  0.33991
## wbhaomMixed             0.025230   1.304296   0.019  0.98457
## age                     0.123819   0.009665  12.811  < 2e-16 ***
## educ                    4.313494   0.112111  38.475  < 2e-16 ***
## cert                    2.014265   0.320078   6.293 3.17e-10 ***
## hourslw                 0.054342   0.010266   5.294 1.21e-07 ***
## factor(arrived)2        6.752708   3.805394   1.775  0.07599 .
## factor(arrived)3        8.141132   3.758629   2.166  0.03032 *
## factor(arrived)4        9.243603   3.703946   2.496  0.01258 *
## factor(arrived)5        9.874985   3.675745   2.687  0.00722 **
## factor(arrived)6        9.122777   3.657184   2.494  0.01262 *
## factor(arrived)7       10.007707   3.673214   2.725  0.00644 **
## factor(arrived)8        9.619454   3.693595   2.604  0.00921 **
```

```
## factor(arrived)9        9.699314   3.678598   2.637  0.00838 **
## factor(arrived)10      10.470279   3.682172   2.844  0.00447 **
## factor(arrived)11       8.252274   3.666461   2.251  0.02441 *
## factor(arrived)12      10.353807   3.661677   2.828  0.00469 **
## factor(arrived)13       8.419809   3.639619   2.313  0.02071 *
## factor(arrived)25       6.386043   3.660521   1.745  0.08107 .
## rural                  -1.959596   0.462569  -4.236 2.28e-05 ***
## docc032                -2.716895   0.649032  -4.186 2.85e-05 ***
## docc033                 4.475820   0.605044   7.398 1.43e-13 ***
## docc034                 0.151492   0.776126   0.195  0.84525
## docc035                -2.722559   0.941790  -2.891  0.00385 **
## docc036               -13.890049   1.009938 -13.753  < 2e-16 ***
## docc037                -3.852083   1.349398  -2.855  0.00431 **
## docc038               -11.398484   0.631396 -18.053  < 2e-16 ***
## docc039                -8.606414   1.008861  -8.531  < 2e-16 ***
## docc0310               -4.352475   0.618810  -7.034 2.07e-12 ***
## docc0311              -17.737402   0.763192 -23.241  < 2e-16 ***
## docc0312              -16.024314   1.095216 -14.631  < 2e-16 ***
## docc0313              -18.807314   0.571155 -32.929  < 2e-16 ***
## docc0314              -16.604144   0.584202 -28.422  < 2e-16 ***
## docc0315              -19.209417   0.657200 -29.229  < 2e-16 ***
## docc0316              -14.617494   0.558406 -26.177  < 2e-16 ***
## docc0317              -15.638795   0.528939 -29.566  < 2e-16 ***
## docc0318              -16.882817   0.873254 -19.333  < 2e-16 ***
## docc0319              -13.235749   0.587667 -22.523  < 2e-16 ***
## docc0320              -14.801654   0.776688 -19.057  < 2e-16 ***
## docc0321              -15.856510   0.558803 -28.376  < 2e-16 ***
## docc0322              -17.439170   0.575222 -30.317  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 16.12 on 23539 degrees of freedom
##   (130693 observations deleted due to missingness)
## Multiple R-squared:  0.386,  Adjusted R-squared:  0.3848
## F-statistic: 321.7 on 46 and 23539 DF,  p-value: < 2.2e-16
```

Summary and conclusion

In this analysis, I examined the drivers of the differential in earnings between US born and migrant workers in the United States. The data shows that migrants have the lowest earnings in their first year of arrival. Earnings increase steeply in the first five years, followed by a gradual decline. The drivers of income differentials include gender, age, race, education, certification, hours worked, job location (rural vs urban), and occupation. The data has a severe case of missing data making analysis challenging. There is potential for omitted variable bias.

References

Appendix - Variable list [Delete this section in your hand-in]

Amo-Agyei, Silas, and International Labour Office. 2020. *The Migrant Pay Gap: Understanding Wage Differences Between Migrants and Nationals*. International Labour Organisation (ILO).

Lin, Ken-Hou, and Inbar Weiss. 2019. “Immigration and the Wage Distribution in the United States.” *Demography* 56 (6): 2229–52.