Empirical Project in Introductory Econometrics Course

Noé NOTTER

2023-03-15

Introduction

In many countries, migrant workers, both men and women, make up a large share of the workforce and make important contributions to societies and economies. Despite the positive migration experiences of many, there remains a strong link between immigration and failure to respect fundamental rights at work. Migrant workers often face unequal treatment in the labor market, particularly with regard to wages. One way to measure inequalities between migrants and non-migrants is to compare the differences in real wages between these two categories. By using datas and economic tools we are going to provide a clear explanation to explain first the immigrant wage gap and second whether and how the wage gap varies by time since immigrants entered the US. Most of the analysis will be done using econometric tools, which, if used well, will provide us with a good estimate of reality and allow us to draw meaningful conclusions.

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric

##         year month minsamp            hhid hhid2 hrsample hrsersuf hhnum
##      1: 2019     1       8 003960201671209 07011       07       01     1
##      2: 2019     1       8 003960201671209 07011       07       01     1
##      3: 2019     1       8 013764301391203 07011       07       01     1
##      4: 2019     1       4 020310137808156 09111       09       11     1
##      5: 2019     1       4 020310137808156 09111       09       11     1
##     ---
## 291386: 2019    12       4 951285190502136 10011       10       01     1
## 291387: 2019    12       4 951606170801615 10111       10       11     1
## 291388: 2019    12       4 951606170801615 10111       10       11     1
## 291389: 2019    12       8 981558015110655 09111       09       11     1
## 291390: 2019    12       8 981558015110655 09111       09       11     1
##           hrlonglk lineno    fnlwgt   orgwgt   lonwgt    famwgt age female
##      1: Continuing      1 1100.6982 4469.187 1594.958 1100.6982  74      1
##      2: Continuing      2 1240.0336 4907.492 1797.918 1100.6982  79      0
##      3: Continuing      1 1232.6945 4683.562 1786.226 1232.6945  42      1
##      4: Continuing      1 1264.3044 5145.019 1833.109 1300.2472  61      0
##      5: Continuing      2 1300.2472 5195.926 1884.113 1300.2472  56      1
##     ---
```

```
## 291386: Continuing      4  717.1595 2836.486 1023.637  717.1595  17      0
## 291387: Continuing      1  733.4708 2554.982 1046.919  733.4708  21      0
## 291388: Continuing      2  866.8010 3545.230 1240.495  866.8010  50      1
## 291389: Continuing      1  833.8034 3244.061 1193.272  833.8034  43      1
## 291390: Continuing      2  833.8034 3244.061 1193.272  833.8034  41      1
##              wbho    wbhao    wbhom   wbhaom racehpia racehpi racea forborn
##      1:    White    White    White    White        0       0    -1       0
##      2:    White    White    White    White        0       0    -1       0
##      3:    White    White    White    White        0       0    -1       0
##      4:    White    White    White    White        0       0    -1       0
##      5:    White    White    White    White        0       0    -1       0
##     ---
## 291386: Hispanic Hispanic Hispanic Hispanic        0       0    -1       0
## 291387:    Other    Asian    Other    Asian        0       0     5       1
## 291388:    Other    Asian    Other    Asian        0       0     5       1
## 291389:    Other    Asian    Other    Asian        1       1    -1       0
## 291390:    Other    Asian    Other    Asian        1       1    -1       0
##          citizen     prcitshp arrived prinusyr      penatvty      pemntvty
##      1:       1   Born in US      NA       NA United States United States
##      2:       1   Born in US      NA       NA United States United States
##      3:       1   Born in US      NA       NA United States United States
##      4:       1   Born in US      NA       NA United States United States
##      5:       1   Born in US      NA       NA United States United States
##     ---
## 291386:       1   Born in US      NA       NA United States United States
## 291387:       0 Foreign born      13       21         Korea         Korea
## 291388:       0 Foreign born      13       21         Korea         Korea
## 291389:       1   Born in US      NA       NA United States United States
## 291390:       1   Born in US      NA       NA United States       Germany
##               pefntvty vet married       marstat ownchild ch02 ch05 ch35
ch613
##      1: United States   0       1       Married        0    0    0    0
0
##      2: United States   1       1       Married        0    0    0    0
0
##      3: United States   0       0      Divorced        3    0    0    0
1
##      4: United States   0       1       Married        0    0    0    0
0
##      5: United States   0       1       Married        0    0    0    0
0
##     ---
## 291386: United States   0       0 Never Married       NA   NA   NA   NA
NA
## 291387:         Korea   0       0 Never Married        0    0    0    0
0
## 291388:         Korea   0       1       Married       NA   NA   NA   NA
NA
## 291389: United States   0       0       Widowed       NA   NA   NA   NA
NA
```

```
## 291390: United States   0       0       Widowed       NA   NA   NA   NA
NA
##         ch1417 famrel84              famrel94
##      1:      0     <NA>      Reference person
##      2:      0     <NA>                Spouse
##      3:      0     <NA>      Reference person
##      4:      0     <NA>      Reference person
##      5:      0     <NA>                Spouse
##    ---
## 291386:     NA     <NA>             Own child
## 291387:      0     <NA>      Reference person
## 291388:     NA     <NA>        Other relative
## 291389:     NA     <NA> Not in primary family
## 291390:     NA     <NA> Not in primary family
##                                          famrel relahh hoh79 refper
##      1: Head, spouse, or unmarried reference person     NA     1      1
##      2: Head, spouse, or unmarried reference person     NA     1      0
##      3: Head, spouse, or unmarried reference person     NA     1      1
##      4: Head, spouse, or unmarried reference person     NA     1      1
##      5: Head, spouse, or unmarried reference person     NA     1      0
##    ---
## 291386:                                   Own child     NA     0      0
## 291387: Head, spouse, or unmarried reference person     NA     1      1
## 291388:                              Other relative     NA     0      0
## 291389:                       Not in primary family     NA     0      0
## 291390:                       Not in primary family     NA     0      0
##                      faminc   lfstat empl unem nilf selfemp selfinc
pubsect
##      1:           40000-49999     NILF    0    0    1      NA      NA
NA
##      2:           40000-49999     NILF    0    0    1      NA      NA
NA
##      3:           20000-24999 Employed    1    0    0       0       0
0
##      4: 75000+ / 75000-99,999 Employed    1    0    0       0       0
0
##      5: 75000+ / 75000-99,999 Employed    1    0    0       0       0
0
##    ---
## 291386:         100000-149999     NILF    0    0    1      NA      NA
NA
## 291387:           20000-24999     NILF    0    0    1      NA      NA
NA
## 291388:           20000-24999 Employed    1    0    0       0       0
0
## 291389:         100000-149999 Employed    1    0    0       0       0
1
## 291390:         100000-149999 Employed    1    0    0       0       0
1
##         pubfed pubst publoc                cow1               cow2 unemdur
```

```
##      1:     NA    NA     NA                <NA>               <NA>      NA
##      2:     NA    NA     NA                <NA>               <NA>      NA
##      3:      0     0      0 PRIVATE, FOR PROFIT               <NA>      NA
##      4:      0     0      0  PRIVATE, NONPROFIT               <NA>      NA
##      5:      0     0      0  PRIVATE, NONPROFIT               <NA>      NA
##     ---
## 291386:     NA    NA     NA                <NA>               <NA>      NA
## 291387:     NA    NA     NA                <NA>               <NA>      NA
## 291388:      0     0      0 PRIVATE, FOR PROFIT               <NA>      NA
## 291389:      0     1      0  GOVERNMENT - STATE PRIVATE, NONPROFIT      NA
## 291390:      0     1      0  GOVERNMENT - STATE GOVERNMENT - STATE      NA
##         jobloser jobleaver entrant unmem uncov union cert certgov schenrl
schhs
##      1:       NA        NA      NA    NA    NA    NA    0      NA      NA
NA
##      2:       NA        NA      NA    NA    NA    NA    0      NA      NA
NA
##      3:       NA        NA      NA     0     0     0    0      NA       0
NA
##      4:       NA        NA      NA     0     0     0    0      NA      NA
NA
##      5:       NA        NA      NA     0     0     0    0      NA      NA
NA
##     ---
## 291386:       NA        NA      NA    NA    NA    NA    0      NA       1
1
## 291387:       NA        NA      NA    NA    NA    NA    0      NA       1
0
## 291388:       NA        NA      NA     0     0     0    0      NA       0
NA
## 291389:       NA        NA      NA     1    NA     1    1       1       0
NA
## 291390:       NA        NA      NA     1    NA     1    1       1       0
NA
##         schcol schft schpt multjob multjobn pdemp1 pdemp2 nmemp1 nmemp2
state
##      1:     NA    NA    NA      NA       NA     NA     NA     NA     NA
Maine
##      2:     NA    NA    NA      NA       NA     NA     NA     NA     NA
Maine
##      3:     NA    NA    NA       0        1     NA     NA     NA     NA
Maine
##      4:     NA    NA    NA       0        1     NA     NA     NA     NA
Maine
##      5:     NA    NA    NA       0        1     NA     NA     NA     NA
Maine
##     ---
## 291386:      0     1     0      NA       NA     NA     NA     NA     NA
Hawaii
## 291387:      1     1     0      NA       NA     NA     NA     NA     NA
```

```
Hawaii
## 291388:     NA    NA    NA       0        1     NA     NA     NA     NA
Hawaii
## 291389:     NA    NA    NA       1        2     NA     NA     NA     NA
Hawaii
## 291390:     NA    NA    NA       1        2     NA     NA     NA     NA
Hawaii
##         metro centcity suburb rural cmsacode05 cmsacode14 fipscounty
##      1:     0        0      0     1       <NA>       <NA>          0
##      2:     0        0      0     1       <NA>       <NA>          0
##      3:     1        0      1     0       <NA>       <NA>          0
##      4:     1        0      0     0       <NA>       <NA>         19
##      5:     1        0      0     0       <NA>       <NA>         19
##     ---
## 291386:     1        0      1     0       <NA>       <NA>          3
## 291387:     1        1      0     0       <NA>       <NA>          3
## 291388:     1        1      0     0       <NA>       <NA>          3
## 291389:     1        0      1     0       <NA>       <NA>          3
## 291390:     1        0      1     0       <NA>       <NA>          3
##         principalcty smsastat05 smsastat14 cbsasz nyc la         educ
##      1:            0       <NA>          0   <NA>  NA NA           HS
##      2:            0       <NA>          0   <NA>  NA NA      College
##      3:            0       <NA>      38860   <NA>   0  0           HS
##      4:            0       <NA>      12620   <NA>   0  0     Advanced
##      5:            0       <NA>      12620   <NA>   0  0     Advanced
##     ---
## 291386:            0       <NA>      46520   <NA>   0  0           HS
## 291387:            0       <NA>      46520   <NA>   0  0 Some college
## 291388:            0       <NA>      46520   <NA>   0  0           HS
## 291389:            0       <NA>      46520   <NA>   0  0      College
## 291390:            0       <NA>      46520   <NA>   0  0     Advanced
##                            educ92 ind_nber
##      1:          HS graduate, GED     <NA>
##      2:         Bachelor's degree     <NA>
##      3:          HS graduate, GED     <NA>
##      4:           Master's degree     <NA>
##      5:                 Doctorate     <NA>
##     ---
## 291386:      12th grade-no diploma     <NA>
## 291387: Some college but no degree     <NA>
## 291388:          HS graduate, GED     <NA>
## 291389:         Bachelor's degree     <NA>
## 291390:           Master's degree     <NA>
##                                               ind_2d ind70 ind80 ind03
ind09
##      1:                                         <NA>  <NA>    NA  <NA>
<NA>
##      2:                                         <NA>  <NA>    NA  <NA>
<NA>
##      3:                               Retail trade 4670-5790  <NA>    NA  <NA>
```

```
<NA>
##      4:              Educational services 7860-7890  <NA>    NA  <NA>
<NA>
##      5:              Educational services 7860-7890  <NA>    NA  <NA>
<NA>
##     ---
## 291386:                                        <NA>  <NA>    NA  <NA>
<NA>
## 291387:                                        <NA>  <NA>    NA  <NA>
<NA>
## 291388: Food services and drinking places 8680,8690  <NA>    NA  <NA>
<NA>
## 291389:              Educational services 7860-7890  <NA>    NA  <NA>
<NA>
## 291390:              Educational services 7860-7890  <NA>    NA  <NA>
<NA>
##         ind12
##      1:  <NA>
##      2:  <NA>
##      3:  <NA>
##      4:  <NA>
##      5:  <NA>
##     ---
## 291386:  <NA>
## 291387:  <NA>
## 291388:  <NA>
## 291389:  <NA>
## 291390:  <NA>
##
ind14
##      1:
<NA>
##      2:
<NA>
##      3:                                          Miscellaneous general
merchandise stores 4529
##      4: Colleges, universities, and professional schools, including junior
colleges 6112, 6113
##      5: Colleges, universities, and professional schools, including junior
colleges 6112, 6113
##     ---
## 291386:
<NA>
## 291387:
<NA>
## 291388:                                      Restaurants and other food
services 722 exc. 7224
## 291389: Colleges, universities, and professional schools, including junior
colleges 6112, 6113
## 291390:                                                     Elementary and
```

```
secondary schools 6111
##                                  ind_m03 agric manuf servs docc70 docc80
##      1:                             <NA>    NA    NA    NA     NA   <NA>
##      2:                             <NA>    NA    NA    NA     NA   <NA>
##      3:      Wholesale and retail trade     0     0    NA     NA   <NA>
##      4: Educational and health services     0     0    NA     NA   <NA>
##      5: Educational and health services     0     0    NA     NA   <NA>
##     ---
## 291386:                             <NA>    NA    NA    NA     NA   <NA>
## 291387:                             <NA>    NA    NA    NA     NA   <NA>
## 291388:         Leisure and hospitality     0     0    NA     NA   <NA>
## 291389: Educational and health services     0     0    NA     NA   <NA>
## 291390: Educational and health services     0     0    NA     NA   <NA>
##                                                            docc03 occ70
occ80
##      1:                                                      <NA>    NA
NA
##      2:                                                      <NA>    NA
NA
##      3: Office and administrative support occupations 5000-5930    NA
NA
##      4:  Education, training, and library occupations 2200-2550    NA
NA
##      5:  Education, training, and library occupations 2200-2550    NA
NA
##     ---
## 291386:                                                      <NA>    NA
NA
## 291387:                                                      <NA>    NA
NA
## 291388:                        Management occupations 0010-0430    NA
NA
## 291389:  Education, training, and library occupations 2200-2550    NA
NA
## 291390:  Education, training, and library occupations 2200-2550    NA
NA
##          occ03 occ11                                                 occ12
##      1:  <NA>  <NA>                                                   <NA>
##      2:  <NA>  <NA>                                                   <NA>
##      3:  <NA>  <NA>                 Stock clerks and order fillers 43-5081
##      4:  <NA>  <NA>                         Postsecondary teachers 25-1000
##      5:  <NA>  <NA>                         Postsecondary teachers 25-1000
##     ---
## 291386:  <NA>  <NA>                                                   <NA>
## 291387:  <NA>  <NA>                                                   <NA>
## 291388:  <NA>  <NA>                          Food service managers 11-9051
## 291389:  <NA>  <NA> Other education, training, and library workers 25-90XX
## 291390:  <NA>  <NA>                     Secondary school teachers 25-2030
##                                                   occ_m03 manag83 manag03
##      1:                                              <NA>      NA      NA
```

```
##      2:                                          <NA>      NA      NA
##      3:   Office and administrative support occupations      NA       0
##      4:            Professional and related occupations      NA       0
##      5:            Professional and related occupations      NA       0
##     ---
## 291386:                                          <NA>      NA      NA
## 291387:                                          <NA>      NA      NA
## 291388: Management, business, and financial occupations      NA       1
## 291389:            Professional and related occupations      NA       0
## 291390:            Professional and related occupations      NA       0
##          hourslwa hourslw hourslwm reason79 reason94 absent79 absent94
abpaid
##      1:       NA      NA       NA       NA     <NA>     <NA>     <NA>
NA
##      2:       NA      NA       NA       NA     <NA>     <NA>     <NA>
NA
##      3:       NA      40       40       NA     <NA>     <NA>     <NA>
NA
##      4:       NA      35       35       NA     <NA>     <NA>     <NA>
NA
##      5:       NA      60       60       NA     <NA>     <NA>     <NA>
NA
##     ---
## 291386:       NA      NA       NA       NA     <NA>     <NA>     <NA>
NA
## 291387:       NA      NA       NA       NA     <NA>     <NA>     <NA>
NA
## 291388:       NA      40       40       NA     <NA>     <NA>     <NA>
NA
## 291389:       NA      44       40       NA     <NA>     <NA>     <NA>
NA
## 291390:       NA      44       32       NA     <NA>     <NA>     <NA>
NA
##          uhours uhourse why3579 why3594 ptecon unempt        prhrusl
hrsvary
##      1:     NA      NA    <NA>    <NA>     NA     NA           <NA>
NA
##      2:     NA      NA    <NA>    <NA>     NA     NA           <NA>
NA
##      3:     NA      NA    <NA>    <NA>      0     NA             40
0
##      4:     NA      35    <NA>    <NA>      0     NA          35-39
0
##      5:     NA      60    <NA>    <NA>      0     NA 50 or more hrs
0
##     ---
## 291386:     NA      NA    <NA>    <NA>     NA     NA           <NA>
NA
## 291387:     NA      NA    <NA>    <NA>     NA     NA           <NA>
NA
```

```
## 291388:      NA      40    <NA>    <NA>      0     NA            40
0
## 291389:      NA      40    <NA>    <NA>      0     NA            40
0
## 291390:      NA      40    <NA>    <NA>      0     NA            40
0
##         pehrusl1 pehrusl2 pehruslt peernhro imphrs hrsimptd uhoursi
blsimpt
##      1:       -1       -1       -1       NA     NA       NA      NA
NA
##      2:       -1       -1       -1       NA     NA       NA      NA
NA
##      3:       40       -1       40       NA     NA        0      40
0
##      4:       35       -1       35       NA     NA        0      35
0
##      5:       60       -1       60       NA     NA        0      60
0
##    ---
## 291386:       -1       -1       -1       NA     NA       NA      NA
NA
## 291387:       -1       -1       -1       NA     NA       NA      NA
NA
## 291388:       40       -1       40       NA     NA        0      40
0
## 291389:       40       10       50       NA     NA        0      40
0
## 291390:       40       12       52       NA     NA        0      40
0
##          blsimph blsimpw paidhre weekpay uearnwke uearnwk earnwke peernuot
##      1:       NA      NA      NA      NA       NA      NA      NA       -1
##      2:       NA      NA      NA      NA       NA      NA      NA       -1
##      3:        0       0       1  461.53       NA      NA      NA        2
##      4:       NA       0       0  623.07       NA      NA      NA        2
##      5:       NA       0       0 1153.84       NA      NA      NA        2
##    ---
## 291386:       NA      NA      NA      NA       NA      NA      NA       -1
## 291387:       NA      NA      NA      NA       NA      NA      NA       -1
## 291388:       NA       1       0 1096.00       NA      NA      NA        2
## 291389:       NA       0       0  923.07       NA      NA      NA        2
## 291390:       NA       0       0  923.07       NA      NA      NA        2
##         otcrec otcamt wage1    wage2    wage3    wage4       rw    rw_ot
tc
##      1:     NA     NA    NA       NA       NA       NA       NA       NA
NA
##      2:     NA     NA    NA       NA       NA       NA       NA       NA
NA
##      3:      0     NA    13       NA 13.00000 13.00000 13.00000 13.00000
NA
##      4:     NA     NA    NA 17.80200 17.80200 17.80200 17.80200 17.80200
```

```
0
##      5:     NA     NA    NA 19.23067 19.23067 19.23067 19.23067 19.23067
0
##     ---
## 291386:     NA     NA    NA       NA       NA       NA       NA       NA
NA
## 291387:     NA     NA    NA       NA       NA       NA       NA       NA
NA
## 291388:     NA     NA    NA 27.40000 27.40000 27.40000 27.40000 27.40000
0
## 291389:     NA     NA    NA 23.07675 23.07675 23.07675 23.07675 23.07675
0
## 291390:     NA     NA    NA 23.07675 23.07675 23.07675 23.07675 23.07675
0
##         proxy wholine                        reltoref
##      1: Proxy       2   Reference person w/ relatives
##      2:  Self       2                          Spouse
##      3:  Self       1   Reference person w/ relatives
##      4:  Self       1   Reference person w/ relatives
##      5: Proxy       1                          Spouse
##     ---
## 291386:  <NA>       2                           Child
## 291387:  Self       1   Reference person w/ relatives
## 291388: Proxy       1                          Parent
## 291389:  Self       1  Reference person w/o relatives
## 291390: Proxy       1 Unmarried partner w/o relatives

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.table':
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

##                       mean           sd       min       max      N
## year          2.019000e+03 0.000000e+00 2019.0000  2019.000 291390
## month         6.485034e+00 3.463702e+00    1.0000    12.000 291390
## minsamp       6.020715e+00 1.999896e+00    4.0000     8.000 291390
## lineno        1.740379e+00 1.011881e+00    1.0000    16.000 291390
## fnlwgt        2.679116e+03 1.443752e+03  150.7482 18675.318 290217
## orgwgt        1.071646e+04 5.812689e+03  591.1690 63257.219 290217
## lonwgt        3.852872e+03 2.068384e+03  215.9119 26764.607 273220
## famwgt        2.664173e+03 1.433756e+03  150.7482 18675.318 291390
```

```
## age          4.819712e+01 1.883334e+01  16.0000   85.000 291390
## female       5.202272e-01 4.995916e-01   0.0000    1.000 291390
## racehpia     6.108652e-03 7.791891e-02   0.0000    1.000 291390
## racehpi      4.245170e-03 6.501663e-02   0.0000    1.000 291390
## racea       -7.476887e-01 1.162069e+00  -1.0000    7.000 291390
## forborn      1.356807e-01 3.424498e-01   0.0000    1.000 291390
## citizen      9.343251e-01 2.477133e-01   0.0000    1.000 291390
## arrived      1.192379e+01 5.219307e+00   1.0000   25.000  43670
## prinusyr     1.495738e+01 6.883517e+00   1.0000   25.000  43670
## vet          7.828532e-02 2.686205e-01   0.0000    1.000 285558
## married      5.254161e-01 4.993545e-01   0.0000    1.000 291390
## ownchild     7.306842e-01 1.102770e+00   0.0000   11.000 175452
## ch02         9.268632e-02 2.899932e-01   0.0000    1.000 175452
## ch05         1.643184e-01 3.705653e-01   0.0000    1.000 175452
## ch35         1.059264e-01 3.077443e-01   0.0000    1.000 175452
## ch613        2.260846e-01 4.182958e-01   0.0000    1.000 175452
## ch1417       1.371600e-01 3.440172e-01   0.0000    1.000 175452
## relahh                NaN           NA      Inf     -Inf      0
## hoh79        6.021209e-01 4.894611e-01   0.0000    1.000 291390
## refper       3.466660e-01 4.759091e-01   0.0000    1.000 291390
## empl         5.966087e-01 4.905788e-01   0.0000    1.000 290217
## unem         2.095673e-02 1.432397e-01   0.0000    1.000 290217
## nilf         3.824345e-01 4.859827e-01   0.0000    1.000 290217
## selfemp      6.451509e-02 2.456689e-01   0.0000    1.000 186344
## selfinc      3.936268e-02 1.944568e-01   0.0000    1.000 186344
## pubsect      1.425858e-01 3.496508e-01   0.0000    1.000 186344
## pubfed       2.771755e-02 1.641628e-01   0.0000    1.000 186344
## pubst        4.923690e-02 2.163628e-01   0.0000    1.000 186344
## publoc       6.563131e-02 2.476372e-01   0.0000    1.000 186344
## unemdur      1.997534e+01 2.690272e+01   0.0000  119.000   6082
## jobloser     4.829004e-01 4.997486e-01   0.0000    1.000   6082
## jobleaver    1.312068e-01 3.376542e-01   0.0000    1.000   6082
## entrant      3.858928e-01 4.868455e-01   0.0000    1.000   6082
## unmem        1.014497e-01 3.019242e-01   0.0000    1.000 154727
## uncov        1.452205e-02 1.196297e-01   0.0000    1.000 139030
## union        1.144984e-01 3.184167e-01   0.0000    1.000 154727
## cert         1.738837e-01 3.790101e-01   0.0000    1.000 290217
## certgov      9.084100e-01 2.884492e-01   0.0000    1.000  50464
## schenrl      1.497933e-01 3.568697e-01   0.0000    1.000 173172
## schhs        3.991133e-01 4.897256e-01   0.0000    1.000  25940
## schcol       6.008867e-01 4.897256e-01   0.0000    1.000  25940
## schft        8.633770e-01 3.434555e-01   0.0000    1.000  25940
## schpt        1.366230e-01 3.434555e-01   0.0000    1.000  25940
## multjob      5.073175e-02 2.194500e-01   0.0000    1.000 173146
## multjobn     1.056057e+00 2.550584e-01   1.0000    4.000 173146
## pdemp1       2.379340e-01 4.258302e-01   0.0000    1.000  18316
## pdemp2       9.752650e-02 2.967258e-01   0.0000    1.000   2830
## nmemp1       8.677145e+00 1.374537e+01   1.0000   75.000   4358
## nmemp2       4.289855e+00 3.282356e+00   1.0000   10.000    276
## metro        8.096827e-01 3.925521e-01   0.0000    1.000 288555
```

```
## centcity      2.440338e-01 4.295136e-01    0.0000     1.000 291390
## suburb        3.934315e-01 4.885120e-01    0.0000     1.000 291390
## rural         1.884656e-01 3.910842e-01    0.0000     1.000 291390
## fipscounty    2.576813e+01 6.199081e+01    0.0000   810.000 291390
## principalcty  1.761213e-01 5.744177e-01    0.0000     7.000 291390
## smsastat14    2.270724e+04 1.650649e+04    0.0000 49740.000 291390
## nyc           5.836666e-02 2.344360e-01    0.0000     1.000 216048
## la            4.099089e-02 1.982696e-01    0.0000     1.000 216048
## ind80                  NaN           NA       Inf      -Inf      0
## agric         1.984502e-02 1.394679e-01    0.0000     1.000 186344
## manuf         1.030889e-01 3.040758e-01    0.0000     1.000 186344
## servs                  NaN           NA       Inf      -Inf      0
## docc70                 NaN           NA       Inf      -Inf      0
## occ70                  NaN           NA       Inf      -Inf      0
## occ80                  NaN           NA       Inf      -Inf      0
## manag83                NaN           NA       Inf      -Inf      0
## manag03       1.207766e-01 3.258684e-01    0.0000     1.000 186344
## hourslwa               NaN           NA       Inf      -Inf      0
## hourslw       3.884764e+01 1.288176e+01    1.0000   198.000 167107
## hourslwm      3.814880e+01 1.241466e+01    0.0000    99.000 167107
## reason79               NaN           NA       Inf      -Inf      0
## abpaid        5.004140e-01 5.000412e-01    0.0000     1.000   6039
## uhours                 NaN           NA       Inf      -Inf      0
## uhourse       3.861439e+01 1.116414e+01    0.0000    99.000 144257
## ptecon        2.668268e-02 1.611548e-01    0.0000     1.000 173146
## unempt        2.053601e-01 4.039977e-01    0.0000     1.000   6082
## hrsvary       6.528072e-02 2.470212e-01    0.0000     1.000 173022
## pehrusl1      2.099928e+01 2.151773e+01   -4.0000    99.000 291390
## pehrusl2     -6.079618e-01 2.975391e+00   -4.0000    99.000 291390
## pehruslt      2.125066e+01 2.198688e+01   -4.0000   198.000 291390
## peernhro      3.542637e+01 1.044374e+01    0.0000    99.000  67153
## imphrs        3.679363e+01 9.648855e+00   12.0000    46.000  11295
## hrsimptd      6.523396e-02 2.469390e-01    0.0000     1.000 173146
## uhoursi       3.862666e+01 1.094794e+01    1.0000    99.000 173022
## blsimpt       2.523304e-02 1.568326e-01    0.0000     1.000 173146
## blsimph       4.232725e-01 4.940806e-01    0.0000     1.000  89883
## blsimpw       3.825577e-01 4.860132e-01    0.0000     1.000 154727
## paidhre       5.809135e-01 4.934112e-01    0.0000     1.000 154727
## weekpay       1.006898e+03 7.089790e+02    0.0000  2884.610 154727
## uearnwke               NaN           NA       Inf      -Inf      0
## uearnwk                NaN           NA       Inf      -Inf      0
## earnwke                NaN           NA       Inf      -Inf      0
## peernuot      5.163115e-01 1.447890e+00   -1.0000     2.000 291390
## otcrec        1.787880e-01 3.831768e-01    0.0000     1.000  89883
## otcamt        2.292141e+02 2.858728e+02    0.0000  2884.610  10948
## wage1         1.864817e+01 1.042861e+01    0.0000    99.990  89883
## wage2         3.345966e+01 2.292930e+01    0.0000  1442.305  61666
## wage3         2.467504e+01 1.820373e+01    0.0000  1442.305 151549
## wage4         2.592125e+01 1.970482e+01    0.0000  1545.710 151549
## rw            2.585840e+01 1.974001e+01    1.0000   392.305 154279
```

```
## rw_ot          2.699360e+01 2.018339e+01    1.0000   392.305 154292
## tc             9.977793e-02 2.997060e-01    0.0000     1.000  64844
## wholine        1.318129e+00 6.912712e-01    0.0000    13.000 291341
```

Data

Selection of the relevant variables for future regressions

The data are provided by the 2019 US current population survey (CPS). They were collected on 291390 people at one moment of time, it is thus cross-sectional data. The data set includes 162 variables, and as we are not going to use them all, we need to analyze them to select to keep only the revelant variable. A variable might be revelant for many reasons. Here we are going through all the variables.

Since this project is an economic study, it makes sense that we would like to add variable having an economic interpretation. So we can say that some variables like “proxy”, “reltoref” will not be used.

Then some variables record a number of missing values (NA) too important to be part of the model. The dataset is composed by 25 totally empty variables, such as “reason79”, “cmsacode05”. Some variables have a very high ratio of NA values, we must be careful to decide whether or not we keep them. The variable “nmemp2” counts 99,9% of values missing, therefore we do not use it.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA  291390
```

```
##                                Appleton-Oshkosh-Neenah, WI
##                                                          0
##        Chicago-Naperville-Michigan City, IL-IN-WI (part)
##                                                          0
##        Cincinnati-Middletown-Wilmington, OH-KY-IN (part)
##                                                          0
##                         Cleveland-Akron-Elyria, OH (part)
##                                                          0
##                             Dallas-Fort Worth, TX (part)
##                                                          0
##                  Dayton-Springfield-Greenville, OH (part)
##                                                          0
##                                  Denver-Aurora-Boulder, CO
##                                                          0
##                                  Detroit-Warren-Flint, MI
##                                                          0
##                                         Fresno-Madera, CA
##                                                          0
##                  Grand Rapids-Muskegon-Holland, MI (part)
##                                                          0
##             Greensboro-Winston-Salem-High Point, NC (part)
##                                                          0
```

```
##                        Greenville-Anderson-Seneca, SC (part) 
##                                                            0 
##                         Houston-Baytown-Huntsville, TX (part) 
##                                                            0 
##                                         Huntsville-Decatur, AL 
##                                                            0 
##                   Indianapolis-Anderson-Columbus, IN (part) 
##                                                            0 
##                Johnson City-Kingsport-Bristol, TN-VA (part) 
##                                                            0 
##                         Los Angeles-Long Beach-Riverside, CA 
##                                                            0 
##                  Macon-Warner-Robins-Fort Valley, GA (part) 
##                                                            0 
##                                Milwaukee-Racine-Waukesha, WI 
##                                                            0 
##                Minneapolis-St. Paul-St. Cloud, MN-WI (part) 
##                                                            0 
##              New York-Newark-Bridgeport, NY-NJ-CT-PA (part) 
##                                                            0 
##            Philadelphia-Camden-Vineland, PA-NJ-DE-MD (part) 
##                                                            0 
##                               Raleigh-Durham-Cary, NC (part) 
##                                                            0 
##                 Sacramento-Arden-Arcade-Truckee, CA-NV (part) 
##                                                            0 
##                  Salt Lake City-Ogden-Clearfield, UT (part) 
##                                                            0 
##                          San Jose-San Francisco-Oakland, CA 
##                                                            0 
##                           Seattle-Tacoma-Olympia, WA (part) 
##                                                            0 
## Washington-Baltimore-Northern Virginia, DC-MD-VA-WV (part) 
##                                                            0 
##            Boston-Worcester-Manchester, MA-NH-CT-ME (part) 
##                                                            0 
##                           Bridgeport-New Haven-Stamford, CT 
##                                                            0 
##                                                         NA's 
##                                                       291390

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00    2.00    3.00    4.29    6.00   10.00  291114

## [1] 99.90528
```

We should also take into account the redundancy between variables. For example “empl” and “unem” are dummy variables for employment and unemployment. Some who answered 1 to employment will automatically answer 0 to unemployment, thus if we include both in the regression, perfect multicolinearity will emerge. The same case appears with geographical variables : “metro”, “centcity”, “suburbs”, “rural”. This is the reason why we will not include them all, mathematically :

$$Empl_i + Unemp_i = 1; Metro_i + Centcity_i + Suburb_i + Rural_i = 1$$

It can be hypothesized that an immigrant with an American mother or father might be more easily accepted into American society. An explanation for this could be because an immigrant in this case would certainly be more fluent in English and a softer accent compared to an immigrant without American parents. He might be considered more of an American than an immigrant. The problem here is that every immigrant in this case would have citizenship, and so it would be redundant with the “citizen” variable. So I decided not to include “pemntvty” and “pefntvty” in my regression model

Lastly, even if some variables are interesting, they should have been restated to be usable. This concerns for example the variable “reason94” describing the reason why the respondent worked less than 35 hours the previous week. These variables will be excluded from the regressions.

Overview of the final data set

We end up with 9 variables between the 162 proposed in the data set.

- wr : real wage. It is the dependent variable of our regressions.
- age: age of the respondent. If the immigrants are in general younger than the non-immigrants, it can create a bias when interpreting the causal effects. We know that salary is positively correlated with age since on average the older you are, the more experience you have.
- female: dummy, is equal to 1 if the respondent is a female. As the american society is still not equal on wage between genders, it is important to keep this variable. “Women in the United States are paid 77 cents for every dollar paid to men, and that gap is widest for women of color”, source U.S. Census Bureau, American Community Survey 1-Year Estimates, 2021.
- wbhao: ethnicity of the respondent. 5 categories possible: White, Black, Hispanic, Asian or Other. I believe this variable is better than “wbho” as it includes the Asian ethnicity, “wbhom” as it is more detailed, “wbhaom” as very few people answer “Native American” (only 3044 over 291390 people), probably because they answer “White” instead, “racea” which is too detailed and only for Asian people, “racehpia” and “racehpi” that are too specific. This variable will tell us the wage disparities between ethnic groups. Indeed, immigrating as a white people (major ethnic group in the US) or immigrating and belonging to a minority has probably an impact on the salary.
- forborn: dummy, is equal to 1 if the person is an immigrant. This will be the major regressor of our regressions since it tells us if the respondent is an immigrant or not.

- prinusyr: year entered the US. It will be helpful for the second question of this assignment. I preferred this variable to “arrived” since it is a little more detailed, with 25 categories instead of 13.
- empl: dummy, is equal to 1 is the respondent is employed. If the immigrants move to the US because they already have a promise of a work, we would interpret wrong the fact that immigrants are better paid.
- rural: dummy, is equal to 1 is the respondent lives in a rural area. As a person living in the city is statistically better paid than a person living in the countryside, in the event that all immigrants choose to settle in the city or in the countryside, this would distort the interpretation.
- educ: variables that has 4 categories to describe the education of the respondent: LTHS (Less Than High School), HS (High School), College, Advanced. As before, if immigrants are less educated when they settle in the United States compared to non-immigrants, it would be logical that on average immigrants have a job with fewer qualifications required and therefore a lower salary.
- manag03: dummy, is equal to 1 if the respondent is a manager.If non-immigrants monopolize management positions, they will statistically earn a higher salary than immigrants, and will therefore distort our interpretation.

I believe that using these variables in our regressions will minimized the bias associated with some omitted variables. I used minimized because I think the model could be improved with other variables such as a dummy variable for students. In a case where the immigrants are mostly students, they have no salary or a low salary due to a small job as a waiter, babysitter next to their study. The dataset provides this variable as “schft”, but too much data is missing especially after coupling with missing data in the salary variable,we would be left with only 10,043 observations, of which only 871 data are immigrants.

Without the variable “schft” we have 22099 observations about immigrants. Only 3.9% of immigrants answered the question “are you a full-time student?”, compared to 6.9% for non-immigrants.As we have seen, this variable is economically relevant in our case, but statistically irrelevant to be part of our regressions.

```
## [1] 10043

## [1] 871

## [1] 22099

## [1] 132180

## [1] 3.941355

## [1] 6.939023
```

As we are not being told of any relevant change in behavior between immigrants and non immigrants, we will therefore assume that there is no measurement errors.

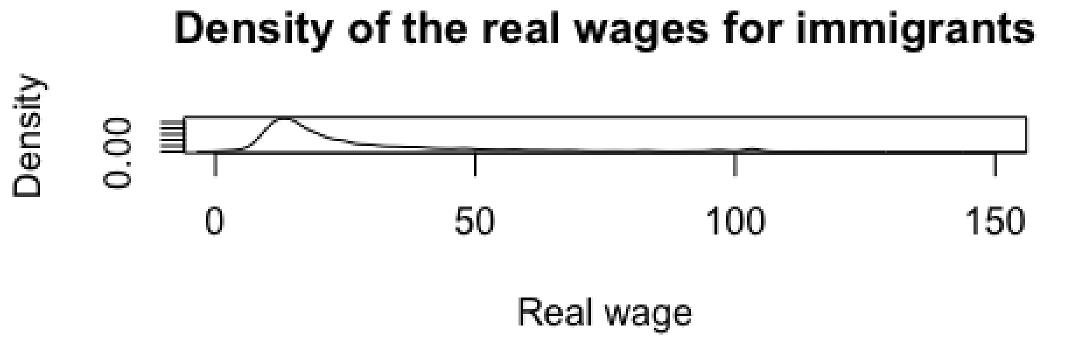
For the question b),the same variables will be used but just adding the time as a regressor variable “prinusyr” (Year entered US) and focusing only on the immigrants by setting forborn = 1.

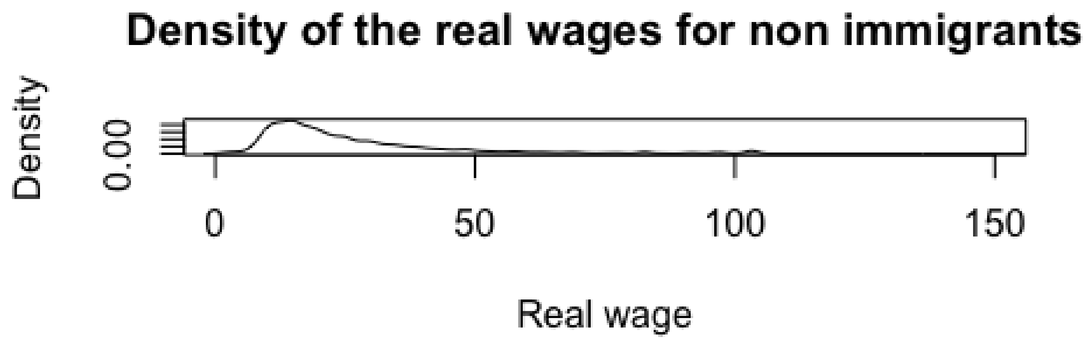
```
##         age female    wbhao forborn prinusyr empl rural         educ
manag03
##      1:  74      1    White       0       NA    0     1           HS
NA
##      2:  79      0    White       0       NA    0     1      College
NA
##      3:  42      1    White       0       NA    1     0           HS
0
##      4:  61      0    White       0       NA    1     0     Advanced
0
##      5:  56      1    White       0       NA    1     0     Advanced
0
##     ---
## 291386:  17      0 Hispanic       0       NA    0     0           HS
NA
## 291387:  21      0    Asian       1       21    0     0 Some college
NA
## 291388:  50      1    Asian       1       21    1     0           HS
1
## 291389:  43      1    Asian       0       NA    1     0      College
0
## 291390:  41      1    Asian       0       NA    1     0     Advanced
0
##               rw
##      1:       NA
##      2:       NA
##      3: 13.00000
##      4: 17.80200
##      5: 19.23067
##     ---
## 291386:       NA
## 291387:       NA
## 291388: 27.40000
## 291389: 23.07675
## 291390: 23.07675
```

Differences between immigrants and non-immigrants

```
##   Number of non immigrants Number of immigrants
##                     251854                39536
```

The full sample consists of 251854 non-immigrants and 39536 immigrants.






```
## [1] "Summary of the variables for immigrants"

##                  mean         sd min      max     N
## age       47.59221975 16.5204847  16  85.0000 39536
## female     0.52891036  0.4991698   0   1.0000 39536
## forborn    1.00000000  0.0000000   1   1.0000 39536
## prinusyr  15.42945670  6.6653351   1  25.0000 39536
## empl       0.63358914  0.4818297   0   1.0000 39483
## rural      0.05334379  0.2247210   0   1.0000 39536
## manag03    0.08913869  0.2849492   0   1.0000 26599
## rw        25.31543115 20.5601132   1 288.3333 22099

## [1] "Summary of the variables for non immigrants"

##                 mean         sd min     max      N
## age       48.2920779 19.1693904  16  85.000 251854
## female     0.5188641  0.4996450   0   1.000 251854
## forborn    0.0000000  0.0000000   0   0.000 251854
## prinusyr  10.4426705  7.2905534   1  25.000   4134
## empl       0.5907855  0.4916899   0   1.000 250734
## rural      0.2096770  0.4070789   0   1.000 251854
## manag03    0.1260446  0.3319007   0   1.000 159745
## rw        25.9491757 19.5981614   1 392.305 132180
```

```
## [1] "% of non-immigrants reporting a value in prinusyr"

## [1] 1.641427
```

Interpretation of these summaries: - the average real wage (25,3 vs 25,9) and their standard deviations (20,6 vs 19,6) for the non-immigrants and immigrants are approximately the same. - wage density is almost the same. The minimum wage is the same for both. The maximum salary is higher for the non-immigrants 392$ and 288 for the immigrants. - The age varies between 16 and 85 years old, with an average age around 48 years old for immigrants and immigrants. - We have as many men as women (≃50%) among the non-immigrants and immigrants. - There is 4 times more non-immigrants living in the rural zone compared to the immigrants (5,3% versus 21%). Employment and educational opportunities are more attractive in the city, so it makes sense that the immigrants choose to settle there rather than in rural areas. - More managers is present for the non-immigrants 12,6% compared to 8,9% for the immigrants. This difference could result from many explanations like a managers visa is harder to get, the immigrants people are less good at English.

The variable prinusyr used for question b) reveals something wrong. We see that we have some statistics (mean, sd, min, max) for this variable for 4134 non-immigrants. Recall that this variable means “years of arrival in the United States”, so it would be wrong to keep this data on people born in the country. Even if only 1.64% of non-immigrants are affected by the prinusyr variable, this could lead to misunderstanding. To avoid this, we will put 0 for the variable prinusyr concerning all non-immigrants.

```
## [1] "Ethnic groups of immigrants"

##    White    Black Hispanic    Asian    Other 
##     8037     3293    17350    10800       56 

## [1] "Ethnic groups of non-immigrants"

##    White    Black Hispanic    Asian    Other 
##   192221    26646    21421     7272     4294 

## [1] "% of each ethnic group between the immigrants people"

##        White      Black   Hispanic      Asian      Other 
##   20.3283084  8.3291178 43.8840550 27.3168758  0.1416431 

## [1] "% of each ethnic group between the non immigrants people"

##       White     Black  Hispanic     Asian     Other 
##   76.322393 10.579939  8.505325  2.887387  1.704956 

## [1] "Summary of the variable rw"

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    1.00   13.50   19.75   25.86   31.25  392.30  137111 

## [1] "% of missing data for the wage"
```

```
## [1] 47.05412
```

Here we notice that the majority of immigrants are Hispanic 43,9%, and the majority of non-immigrants are White 76,3%. Asian is the biggest ethnic difference between them, 10 times more Asian in the immigrants sample.

Since we have a lot of data missing on the salary, 47% (rw variable), I am now going to look closer if a certain population didn't answer the survey. If the immigrants for example have a very low rate of answer on this question, this could lead to misinterpretation when coming to the regressions.

```
##                                              Non immigrants White immigrants
## Total people in survey                               251854             8037
## Number of people answering to their salary           132180             3858
## Ratio per group (in %)                                   52               48
##                                              Black immigrants Hispanic
immigrants
## Total people in survey                                   3293
17350
## Number of people answering to their salary               1990
10107
## Ratio per group (in %)                                     60
58
##                                              Asian immigrants Other
immigrants
## Total people in survey                                  10800
56
## Number of people answering to their salary               6121
23
## Ratio per group (in %)                                     56
41
```

We can see that 52% of the non-immigrants reported their wage in the survey, as well as 48% of the White immigrants,and so on. The missing value on the variable seems to appears randomly distributed between all the ethnic groups. We can use the variables "rw" and "wbhao" safely.

Empirical Approach

First question: Quantify the immigrant wage gap and explore possible explanations.

I use log transform for the dependant variable "rw".I believe the increase to be relevant proportionally (+1% income) rather than linearly (+1$ income). Since I think a dollar is not the same for a millionaire and for a pauper, I do not choose linear in this case. A dollar does nothing for a millionaire but a lot for a pauper, so I choose ln(). The estimates will thus be interpreted such that adding one number of the regressor has an impact in percentages of the wage.

Question a) quantify the immigrant wage gap and explore possible explanations. To start I am going to analyze a simple regression where the logarithm of the real wage is the dependant variable and the dummy variable "forborn" is the regressor. This is supposed to give us a first look about the immigrant wage gap.

$$log(wr_i) = \beta_0 + \beta_1 forborn_i + u_i$$

As this first simple regression will suffer from omitted variables bias, I am going to add some regressors to switch on a multiple log-linear regression.

$$\begin{aligned} &log(wr_i) \\ &= \beta_0 + \beta_1 forborn_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 female_i + \beta_5 empl_i + \beta_6 rural_i + \beta_7 educ_i \\ &\quad + \beta_8 manag03_i + \beta_9 whbao_i + u_i \end{aligned}$$

Since there are five categories of education and five categories of ethnic groups, the relation will actually look like:

$$\begin{aligned} &log(wr_i) \\ &= \beta_0 + \beta_1 forborn_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 female_i + \beta_5 empl_i + \beta_6 rural_i + \beta_7 HS_i \\ &\quad + \beta_8 Somecollege_i + \beta_9 College_i + \beta 10 Advanced_i + \beta 11 manag03_i + \beta 12 Black_i \\ &\quad + \beta_{13} Hispanic_i + \beta_{14} Asian_i + \beta_{15} Other_i + u_i \end{aligned}$$

We know that the relation between wage and age is non-linear. Actually the age-wage curve is concave, our first years of work we learn way more experiences than during our last year of work. The relation is supposed quadratic and thus we use a quadratic multiple regression model by including the variable "age^2". This will eliminate a bias of functional form misspecification.

I will end the question by computing the 95% confidence interval of the estimate of the dummy "forborn" and computing the Wald test to check the consistency of the estimates.

Second question: Investigate whether and how the wage gap varies by time since immigrants entered the US.

As we did for the question 1, the first regression is going to be a simple regression, of the date of arrival "prinusyr" on the log of wages.

All the immigrants reported their arrival date in the United-States with a code. The higher it is the sooner they came. The code ranges from 1 to 25. For example 1 means the person arrived before 1950, 25 means after 2017, 14 means between 1994-1995.

$$log(wr_i) = \beta_0 + \beta_1 prinusyr_i + u_i$$

This first regression will suffer from omitted variables bias, so we are going to switch to the quadratic multiple regression model while maintaining the regressor "prinusyr". Apart from this addition, the multiple regression is the same. The dependent variable is the logarithm of the real wage. We remove the variable of employment since it was colinear with another regressor (see results from question a).

The regression will be:

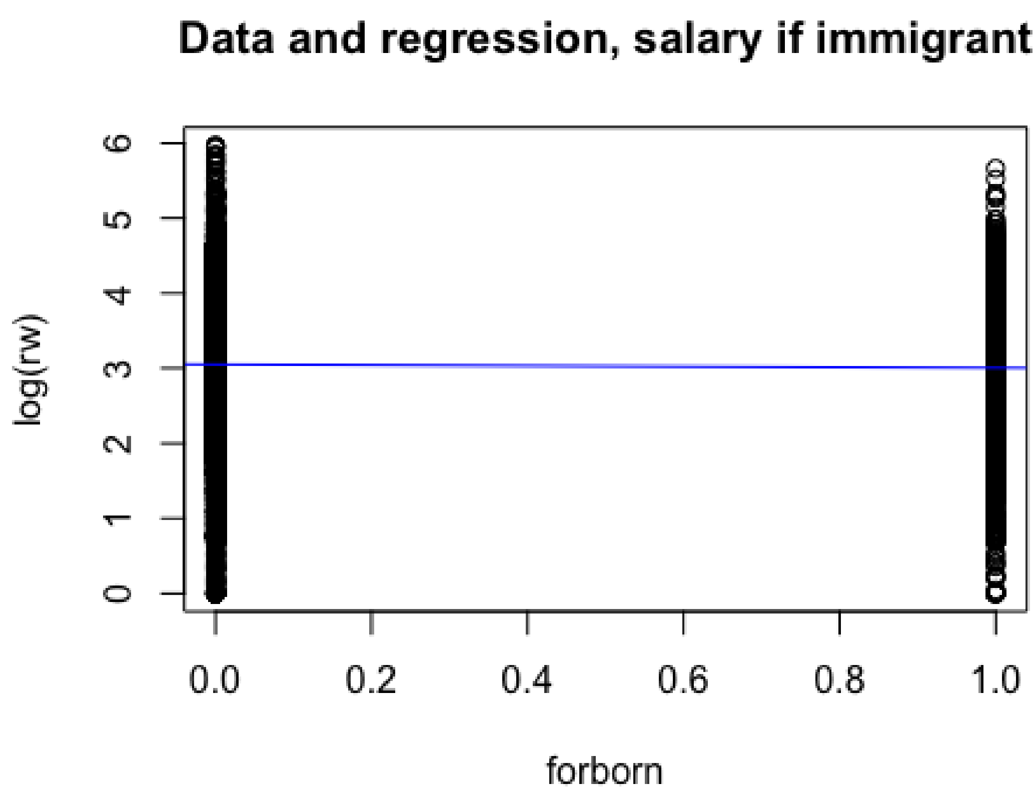
$$\log(wr_i) = \beta_0 + \beta_1 prinusyr_i + \beta_2 age_i + \beta_3 age_i^2 + \beta_4 female_i + \beta_5 rural_i + \beta_6 HS_i + \beta_7 Somecollege_i + \beta_8 College_i + \beta 9 Advanced_i + \beta 10 manag03_i + \beta 11 Black_i + \beta_{12} Hispanic_i + \beta_{13} Asian_i + \beta_{14} Other_i + u_i$$

As before, a 95% confidence interval of the estimate of "prinusyr" and perform a Wald test will be calculated to check the null hypothesis of having all estimates equal to 0.

Results

First question: Quantify the immigrant wage gap and explore possible explanations.

```
##                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  3.04996206 0.001721807 1771.37280 0.000000e+00
## forborn     -0.04839966 0.004668480  -10.36733 3.558784e-25

## [1] 0.0007266576
```



As the homoskedasticity of the error terms is just a particular case of the heteroskedasticity, the function coeftest.hc1 builds heteroskedastict robust estimates.

This first simple regression finds a negative causal effect of being an immigrant on the real wage. The economic interpretation is : In average the real wage of an Im is 4.84% lower than the real wage of a non-immigrants. The t values of

$$\widehat{\beta_1}$$

and of the intercept are higher than 1.96 in absolute value and the P values are less than 0.05, which means that the intercept and the estimate are significant are the 5% level.

The graph is quite illegible, since we have a very high number of data plotted on a dummy variable. The "slope" is only connecting two dots: 3.049962 and 3.001562 dollars.

But we know that the estimate

$$\widehat{\beta_1}$$

suffers from omitted variables bias and thus does not illustrates the right effect of being an immigrant on the real wage. The R^2 tells us whether the regressor is good at predicting values of the dependent variable in the sample. Here the

$$R^2$$

is extremely low (0.000727), meaning our model explains almost anything about the wage gap. Let's have a look to the second multiple regression that adds several regressors.

```
##                        Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)        1.6317464728 1.075800e-02 151.677461  0.000000e+00
## forborn           -0.0556727701 4.576984e-03 -12.163636  5.032465e-34
## age                0.0502241853 5.509178e-04  91.164572  0.000000e+00
## I(age^2)          -0.0004891205 6.461249e-06 -75.700611  0.000000e+00
## female            -0.2197684604 2.595265e-03 -84.680536  0.000000e+00
## rural             -0.0962835787 3.354008e-03 -28.707021 9.361393e-181
## educHS             0.1705335976 4.542256e-03  37.543809 4.382806e-307
## educSome college   0.2719891176 4.672734e-03  58.207699  0.000000e+00
## educCollege        0.6168506479 5.219585e-03 118.180020  0.000000e+00
## educAdvanced       0.8338295836 5.934140e-03 140.513969  0.000000e+00
## wbhaoBlack        -0.1446000682 4.235276e-03 -34.141825 1.587383e-254
## wbhaoHispanic     -0.0637648409 4.090462e-03 -15.588665  9.572650e-55
## wbhaoAsian         0.0447621663 6.234883e-03   7.179311  7.037756e-13
## wbhaoOther        -0.0741622422 1.072659e-02  -6.913868  4.734357e-12
## manag03            0.2745540556 4.474671e-03  61.357370  0.000000e+00

## [1] 0.3552957
```

In this second regression, all the estimates are highly significant to the 5% level because the absolute value of our t statistic is greater than 1.96 for every estimates. Indeed, all the P-values are less than 0.05. We can notice that the regression dropped the dummy variable of employment.Perfect multicollinearity arises when one of the regressors is a perfect linear combination of the other regressors. Employment was perfect multicollinear with another variable. Because of this, we will not have any causal effect from the employment rate on the wage.

```
##                        Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)       1.6317464728 1.075800e-02 151.677461  0.000000e+00
## forborn          -0.0556727701 4.576984e-03 -12.163636  5.032465e-34
## age               0.0502241853 5.509178e-04  91.164572  0.000000e+00
## I(age^2)         -0.0004891205 6.461249e-06 -75.700611  0.000000e+00
## female           -0.2197684604 2.595265e-03 -84.680536  0.000000e+00
## rural            -0.0962835787 3.354008e-03 -28.707021 9.361393e-181
## educHS            0.1705335976 4.542256e-03  37.543809 4.382806e-307
## educSome college  0.2719891176 4.672734e-03  58.207699  0.000000e+00
## educCollege       0.6168506479 5.219585e-03 118.180020  0.000000e+00
## educAdvanced      0.8338295836 5.934140e-03 140.513969  0.000000e+00
## wbhaoBlack       -0.1446000682 4.235276e-03 -34.141825 1.587383e-254
## wbhaoHispanic    -0.0637648409 4.090462e-03 -15.588665  9.572650e-55
## wbhaoAsian        0.0447621663 6.234883e-03   7.179311  7.037756e-13
## wbhaoOther       -0.0741622422 1.072659e-02  -6.913868  4.734357e-12
## manag03           0.2745540556 4.474671e-03  61.357370  0.000000e+00

## [1] 0.3552957
```

We found that females earn 22% less than males on average; living in a rural environment decreases one's wage by 9.6%; education has a positive effect on wages and the level of the diploma is positively correlated with the salary;The Asian ethnicity has a positive estimate and earn in average 4.5% more than White people, other ethnicities earn less than whites.

What we can observe is that the estimate of the forborn variable is a bit more negative after adding control variables. It ranges from -0.0483 to -0.0556. Holding all other variables constant, an immigrant earns on average 5.57% less than a non-immigrant. We can build a 95% confidence interval on the estimate

$$\widehat{\beta_1}$$

of forborn to get an interval containing the real causal effect with 95% of chance.

```
## [1] -0.06464366 -0.04670188
```

We are sure at 95% that the true causal effect of being an immigrant on the wage is between -6.46% and -4.67%.

Even though we should not rely too much on the adjusted

$$R^2$$

, in our case it is still relevant to comment on it. We first had a

$$R^2$$

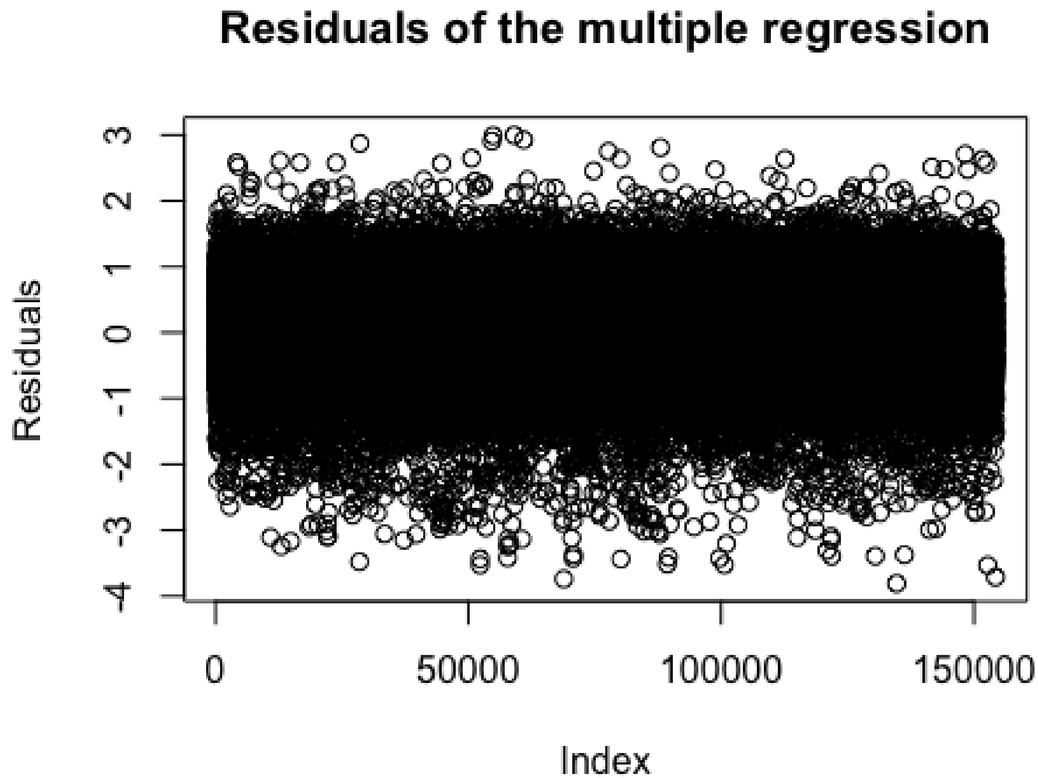
of 0.000727, which became 0.355 in the second regression. The second is almost 500 times larger than the first, which means that the second model is more reliable in explaining the wage gap.

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## Loading required package: survival

## Wald test
##
## Model 1: log(rw) ~ forborn
## Model 2: log(rw) ~ forborn + age + I(age^2) + female + rural + educ +
##     wbhao + manag03
##   Res.Df Df      F    Pr(>F)
## 1 154277
## 2 154264 13 6527.9 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Wald test tells us that we are rejecting the joint null hypothesis of all omitted variables' estimates to be equal to 0. Indeed, the p-value of the test is very close to 0.

In conclusion, the estimate of the variable forborn is highly significant, different of 0 and its value is between -0.0646 and -0.0467.

Before moving through the question b) we would like to make sure that the assumptions of the regression model are still valid, to avoid a bias in our interpretation. Assumption 1:




```
## [1] 1.202593e-16
```

The residuals of this multiple regression are random, does not depend on X and have a mean of approximately 0:

$$E(u_i|X_{1i},\ldots,X_{ki}) = 0$$

Assumption 1 holds.

Assumption 2: As the data are collected randomly on the population sample, we assume therefore that all of the regressors are independent and identically distributed (i.i.d).

Assumption 3:

```
## [1] 10675550

## [1] 0.5202272

## [1] 0.1356807

## [1] 0.1884656

## [1] 0.1207766

## [1] 7034201
```

We have

$$0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$$

so the third assumption is valid as well.

Assumption 4: There is no longer perfect multicollinearity in the multiple regression model, since the only one was employment and was removed in the process. This assumption holds.

All the assumptions for multiple OLS regression hold.

Let's consider the following threats to internal validity: – Omitted variables : we still have some variables missing that could explain the wage gap (cf the variable student). – Functional form misspecification : It is difficult to say what is the correct functional form, but quadratic seems better than linear model. – Measurement error : This is potentially important. Since the data are from a survey there might be measurement error both in the dependent variable as in the independent variables. – Sample selection : As we have seen, the sample is very large and made up of random people, so sample selection within this population is unlikely to be a problem. – Simultaneous causality : age and education are certainly causal, we are older, higher education than we have on average. – Heteroskedasticity and/or correlated error terms : Heteroskedasticity-robust standard errors were used. The data represent a random sample so that correlation across the error terms is unlikely to be a problem.

Threats to external validity : - Difference in populations : Would the model be relevant to apply in France? Certainly not. Immigrants who migrated to USA or France are definitely not from the same country/nationality. Maybe they migrate for different reasons, with different characteristics, different background. Moreover, an American or a French man may have a different approach and way of thinking with immigrants, so we have a very low probability that our explanation of the immigrant wage gap in the United States is valid in France. - Difference in settings : The political system, the labor law, the educational system, the culture are different and therefore this study is only valid for the United States

Second question: Quantify the immigrant wage gap and explore possible explanations, whether and how the wage gap varies by time since immigrants entered the US

```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept)  3.051620397 0.0017064033 1788.33484 0.000000e+00
## prinusyr    -0.003688708 0.0002673336  -13.79814 2.774766e-43

## [1] 0.001292021
```

As we can see in this first simple regression, the bigger the regressor prinusyr, the lower the wage. The value of the estimate is -0.0036887, which means that every migrant that arrived one period later than another one will earn 0.37% less. The t-statistics is greater in absolute value (13,8) than 1.96 and the p-value is approximately 0, which means that the estimate is highly significant to the 5% level. For example, in average an immigrant that arrived before 1950 (code 1) will earn 0.37% less than a non-immigrant (coded 0). Or an immigrant that arrived between 1986 and 1987 (coded 10) will earn in average 0.74% (2*0.37) more than an immigrant that arrived in the period 1990-1991 (coded 12).

The

$$R^2$$

of the regression is very low 0.0013, the reason is probably that this simple regression suffers from omitted variable bias. Let's use our second model with more variables to see the differences.

```
##                     Estimate   Std. Error    t value      Pr(>|t|)
## (Intercept)      1.6346114447 1.074472e-02 152.131559  0.000000e+00
## prinusyr        -0.0036308276 2.508441e-04 -14.474438  1.888290e-47
## age              0.0501919995 5.491911e-04  91.392588  0.000000e+00
## I(age^2)        -0.0004898082 6.454024e-06 -75.891902  0.000000e+00
## female          -0.2199379568 2.594623e-03 -84.766833  0.000000e+00
```

```
## rural             -0.0962455248 3.352456e-03 -28.708963 8.855645e-181
## educHS             0.1706900129 4.515934e-03  37.797278 3.387434e-311
## educSome college  0.2717306955 4.642341e-03  58.533124  0.000000e+00
## educCollege        0.6169957196 5.194630e-03 118.775673  0.000000e+00
## educAdvanced       0.8347897861 5.909416e-03 141.264338  0.000000e+00
## wbhaoBlack        -0.1437409462 4.234765e-03 -33.943074 1.317386e-251
## wbhaoHispanic     -0.0635283363 3.976847e-03 -15.974547  2.137950e-57
## wbhaoAsian         0.0479034811 6.161898e-03   7.774144  7.642517e-15
## wbhaoOther        -0.0742602499 1.072599e-02  -6.923392  4.426686e-12
## manag03            0.2743304059 4.475268e-03  61.299208  0.000000e+00

## [1] 0.3555506
```

In this second regression, the estimate

$$\hat{\beta}_1$$

on the real wage is almost the same as before, about -0.363%, against -0.37% in the first simple regression. The p-values of all the estimates are highly significant to the 5% level. The interpretation of the estimate of “prinusyr” is the same as before. An immigrant that arrived in the United-States during the period 2002-2003 (coded 18) will earn in average 6.534% (= 18*0.363) less than a non-immigrant (coded 0).This difference is enormous and we will come back to it later.

The

$$R^2$$

of this regression is greater and values 0.3556, which is way higher than the

$$R^2$$

of the first simple regression model which valued 0.0013. It suggests that the variance of the wages are better explained in the second model.

Calculation of the 95% confidence interval of the estimate :

```
## [1] -0.004122482 -0.003139173
```

The real value of the estimate is in the interval (-0.412%, -0.314%) with a 95% probability. The causal effect is pretty low for immigrants that arrived in the United-States a long time ago, but it accumulates over time and becomes important for later immigrants.

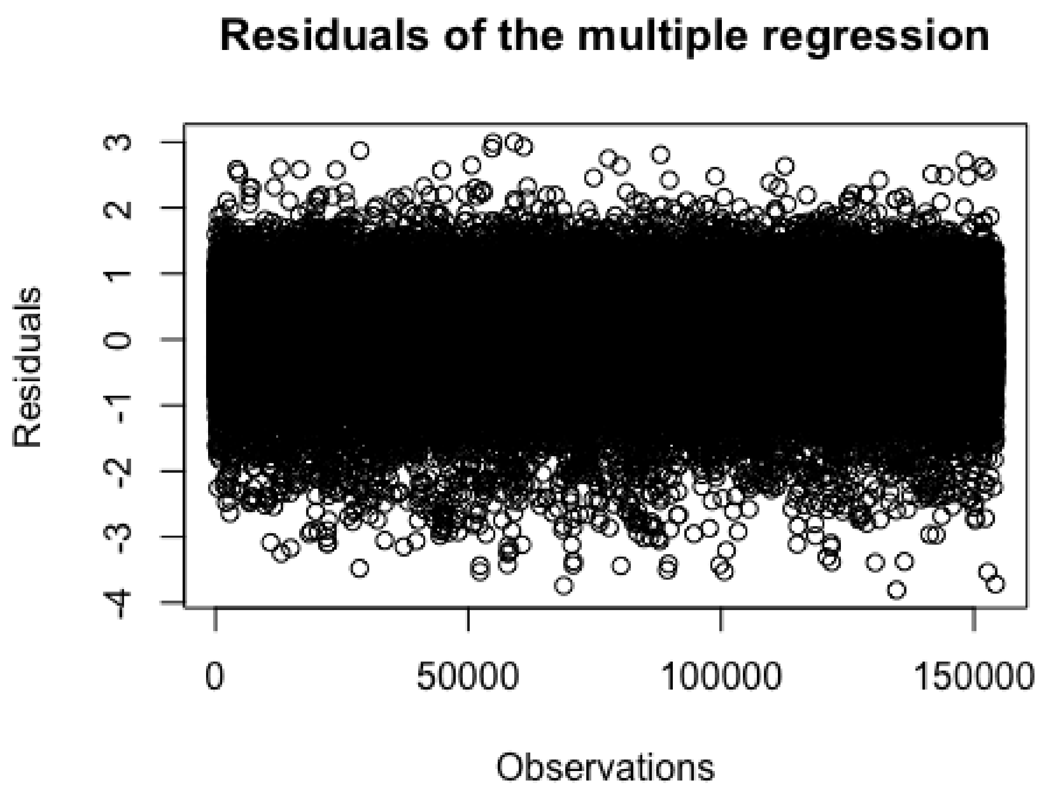
Wald test to test the null hypothesis : all the estimates are 0.

```
## Wald test
##
## Model 1: log(rw) ~ prinusyr
## Model 2: log(rw) ~ prinusyr + age + I(age^2) + female + rural + educ +
##     wbhao + manag03
##   Res.Df Df      F    Pr(>F)
## 1 154277
```

```
## 2 154264 13 6524.8 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

According to the Wald test, we strongly reject the null hypothesis of all estimates to be 0.

Again it is important to check that the assumptions of the regression model are valid, so that there is no bias in our interpretation. Assumption 1:




```
## [1] 1.432724e-16
```

The residuals of this multiple regression are random, does not depend on X and have a mean of 0:

$$E(u_i|X_{1i},\ldots,X_{ki}) = 0$$

Assumption 1 holds.

Assumption 2: As the data are collected randomly on the population sample, we assume therefore that all of the regressors are independent and identically distributed (i.i.d).

Assumption 3:

```
## [1] 10675550
```

```
## [1] 0.5202272

## [1] 0.1356807

## [1] 16115.58

## [1] 0.1884656

## [1] 0.1207766

## [1] 7034201
```

We have

$$0 < E(X_{1i}^4) < \infty, \ldots, 0 < E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$$

so the third assumption holds.

Assumption 4: There is no longer perfect multicollinearity in the multiple regression model, since the only one was employment and was removed in the process. This assumption holds.

All the assumptions for multiple OLS regression hold. I am allowed to comment the results of it.

Let's consider the following threats to internal validity: – Omitted variables : we still have some variables missing that could explain the wage gap (cf the variable student,or a variable on the wage gap between NIms to see if the decreasing wage affected only Im overtime or not). – Functional form misspecification : same arguments in question a) – Measurement error :same arguments in question a) – Sample selection : same arguments in question a) – Simultaneous causality : same arguments in question a) – Heteroskedasticity and/or correlated error terms : same arguments in question a)

Threats to external validity : - Difference in populations : same arguments in question a) - Difference in settings : same arguments in question a). Moreover, the economic evolution is not the same in France as in the USA

Summary and conclusion

As a conclusion of the first question, We found a causal negative effect of -5.57% on the real wage of being an immigrant. On average an immigrant earns 5,57% less than a non-immigrant in the US.

The multiple regression model of the second question suggests that there is a negative causal effect of the time of arrival of the immigrants on their wage. This effect is estimated at -0.363% It means that an immigrant earns in average 0.363% less than another immigrant that arrived in the United-States one period before.

By combining the results of questions a) and b), we see that there is a difference in the real wages received by immigrants and non-immigrants, but also between immigrants themselves when they have arrived at a period different in the United States

Social interpretation and bias

This wage gap that we studied could be the result of many things. The immigrant's productivity and salary upon entering the host country is lower than the productivity and salary of a comparable native worker because the education and experience acquired abroad are not perfectly transferable across borders. Moreover, immigrants can accept a job even if they know that they are paid less than a non-immigrant because, in some cases, they still earn proportionally more than in their country of origin.

One of the limitations of this model lies in the data on education. We assume in our model that having an American degree has the same value as having a foreign degree. Of course we know that is not true. Solving this problem can be difficult, but will provide a more concrete analysis. Another limitation is how attractive are immigrants compared to non-immigrants. Some studies have shown that attractive people are more likely to find professional success and are often offered more jobs, higher salaries, and promotions.Having a variable measuring the level of attractiveness through the golden ratio also known as the beauty ratio will help avoid bias. Will the cost of this variable measure be worth the results?

A criticism that applies to every analysis on humans is that humans are biased, biased data is everywhere.

References

- Lecture 1 to 9 by Edwin Leuven, at University of Oslo
- What's the Wage Gap in the States?; U.S. Census Bureau, American Community Survey 1-Year Estimates, 2021 https://www.nationalpartnership.org/our-work/economic-justice/wage-gap/
- Introduction to Econometrics, Fourth edition, James H. Stock & Mark W. Watson
- https://www.dermatologytimes.com/view/art-assessment-what-beauty Harrar H, Myers S, Ghanem AM. Art or Science? An Evidence-Based Approach to Human Facial Beauty a Quantitative Analysis Towards an Informed Clinical Aesthetic Practice. Aesthetic Plast Surg. 2018;42(1):137-146.

Appendix - All code for this assignment

```
library(foreign)
data = read.dta("cepr_org_2019.dta")

library(lmtest)
library(sandwich)
library(data.table)
library(readxl)
library(fixest)
library(stats)

data = as.data.table(data)
data
```

```
# Quick visual description of the data

library("dplyr")
data1=select_if(data, is.numeric)
dstat = function(x, ...){
c(mean = mean(x, ...), sd = sd(x, ...),
min = min(x, ...), max = max(x, ...), N = sum(!is.na(x)))
}
t(sapply(data1, dstat, na.rm=T))
# Summarizing our variables to get the number of NA
summary(data$reason79)
summary(data$cmsacode05)

summary(data1$nmemp2)
(291114/291390)*100
# Number of respondents with the variable "schft" :
df_st <- na.omit(data[,c(24,72,157)])
rows_student <- nrow(df_st)
rows_student
nb_imm_st <- sum(df_st[df_st$forborn==1]$forborn)
nb_imm_st

# Number of respondents without the variable "schft" :
df_no_st <- na.omit(data[,c(24,157)])
nb_imm_not_st <- sum(df_no_st[df_no_st$forborn==1]$forborn)
nb_imm_not_st
nb_notimm_not_st <- nrow(df_no_st)-nb_imm_not_st
nb_notimm_not_st

# Ratio of immigrants and non immigrants that answered to the querstion "Are
you a full-time student ?" :
nb_imm_st/nb_imm_not_st*100
(rows_student-nb_imm_st)/nb_notimm_not_st*100
df = data[,c(15,16,18,24,28,49,84,94,118,157)]
df
nb_immigrants <- sum(df[df$forborn == 1,]$forborn)
nb_non_immigrants <- nrow(df) - nb_immigrants


tab <- matrix(c(nb_non_immigrants, nb_immigrants), ncol=2)
colnames(tab) <- c("Number of non immigrants", "Number of immigrants")
rownames(tab) <- c(" ")
tab <- as.table(tab)
tab
layout(mat = c(2, 1))
plot(density(df[df$forborn == 0,]$rw, na.rm = TRUE), xlim = c(0,150), main =
"Density of the real wages for non immigrants", xlab = "Real wage")
plot(density(df[df$forborn == 1,]$rw, na.rm = TRUE), xlim = c(0,150), main =
"Density of the real wages for immigrants", xlab = "Real wage")
```

```r
print("Summary of the variables for immigrants")
t(sapply(df[df$forborn == 1,c(-3, -8)], dstat, na.rm=T))

print("Summary of the variables for non immigrants")
t(sapply(df[df$forborn == 0,c(-3, -8)], dstat, na.rm=T))

print("% of non-immigrants reporting a value in prinusyr")
4134/251854*100

# Reprocess prinusyr for non-immigrants
df[df$forborn == 0,]$prinusyr = 0

print("Ethnic groups of immigrants")
summary(df[df$forborn == 1,]$wbhao)

print("Ethnic groups of non-immigrants")
summary(df[df$forborn == 0,]$wbhao)

# % of each ethnic between the Im and secondly between the NIm
White_Im=(8037/39536)*100;
Black_Im=(3293/39536)*100
Hispanic_Im=(17350/39536)*100
Asian_Im=(10800/39536)*100
Other_Im=(56/39536)*100

White_NIm=(192221/251854)*100;
Black_NIm=(26646/251854)*100
Hispanic_NIm=(21421/251854)*100
Asian_NIm=(7272/251854)*100
Other_NIm=(4294/251854)*100


tab <- matrix(c(White_Im, Black_Im, Hispanic_Im, Asian_Im, Other_Im), ncol=5)
colnames(tab) <- c("White", "Black","Hispanic","Asian","Other")
rownames(tab) <- c(" ")
tab <- as.table(tab)
print("% of each ethnic group between the immigrants people")
tab


tab <- matrix(c(White_NIm, Black_NIm, Hispanic_NIm, Asian_NIm, Other_NIm),
ncol=5)
colnames(tab) <- c("White", "Black","Hispanic","Asian","Other")
rownames(tab) <- c(" ")
tab <- as.table(tab)
print("% of each ethnic group between the non immigrants people")
tab
```

```
# Proportion of missing data for the wage variable in the sample
print("Summary of the variable rw");summary(df$rw)
print("% of missing data for the wage"); (137111/291390)*100

df1 <- na.omit(df[, c(3,4,10)])

# Number of immigrants by ethnic group in the full sample
nb_white_imm <- sum(df[(df$forborn == 1) & (df$wbhao == "White"),]$forborn)
nb_black_imm <- sum(df[(df$forborn == 1) & (df$wbhao == "Black"),]$forborn)
nb_hisp_imm <- sum(df[(df$forborn == 1) & (df$wbhao == "Hispanic"),]$forborn)
nb_asian_imm <- sum(df[(df$forborn == 1) & (df$wbhao == "Asian"),]$forborn)
nb_other_imm <- sum(df[(df$forborn == 1) & (df$wbhao == "Other"),]$forborn)

# Number of immigrants by ethnic group who answered to the survey about their
real wage
nb_nimm_ans <- nrow((df1[(df1$forborn == 0),]))
nb_white_imm_ans <- sum(df1[(df1$forborn == 1) & (df1$wbhao ==
"White"),]$forborn)
nb_black_imm_ans <- sum(df1[(df1$forborn == 1) & (df1$wbhao ==
"Black"),]$forborn)
nb_hisp_imm_ans <- sum(df1[(df1$forborn == 1) & (df1$wbhao ==
"Hispanic"),]$forborn)
nb_asian_imm_ans <- sum(df1[(df1$forborn == 1) & (df1$wbhao ==
"Asian"),]$forborn)
nb_other_imm_ans <- sum(df1[(df1$forborn == 1) & (df1$wbhao ==
"Other"),]$forborn)

# Construction of the table
row1 <- c(nb_non_immigrants, nb_white_imm, nb_black_imm, nb_hisp_imm,
nb_asian_imm, nb_other_imm)
row2 <- c(nb_nimm_ans, nb_white_imm_ans, nb_black_imm_ans, nb_hisp_imm_ans,
nb_asian_imm_ans, nb_other_imm_ans)
row3 <-
c(nb_nimm_ans/nb_non_immigrants*100,nb_white_imm_ans/nb_white_imm*100,
nb_black_imm_ans/nb_black_imm*100, nb_hisp_imm_ans/nb_hisp_imm*100,
nb_asian_imm_ans/nb_asian_imm*100, nb_other_imm_ans/nb_other_imm*100)

tab_full_imm <- rbind(as.integer(row1), as.integer(row2), as.integer(row3))

colnames(tab_full_imm) <- c("Non immigrants", "White immigrants", "Black
immigrants", "Hispanic immigrants", "Asian immigrants", "Other immigrants")
rownames(tab_full_imm) <- c("Total people in survey", "Number of people
answering to their salary", "Ratio per group (in %)")
tab_full_imm <- as.table((tab_full_imm))
tab_full_imm
```

```r
coeftest.hc1 = function(x, ...) {
coeftest(x, vcovHC(x, type = "HC1"), ...)[1:x$rank,]
}

reg1 = lm(log(rw) ~ forborn, df)
coeftest.hc1(reg1)
summary(reg1)$r.square



plot(log(rw) ~ forborn, df, main = "Data and regression, salary if
immigrant")
abline(reg1, col = "blue")
reg2 = lm(log(rw) ~ forborn + age + I(age^2) + female + rural + educ + wbhao
+ manag03 + empl, data = df, na.action = na.omit)
reg2_robust <- coeftest.hc1(reg2)
reg2_robust
summary(reg2)$adj.r.square
# Removal of the variable empl of employment to remove multicolinearity
reg2 = lm(log(rw) ~ forborn + age + I(age^2) + female + rural + educ + wbhao
+ manag03, data = df, na.action = na.omit)
reg2_robust <- coeftest.hc1(reg2)
reg2_robust
summary(reg2)$adj.r.square
CI_forborn <- c(reg2_robust[2,1]-1.96*reg2_robust[2,2],
reg2_robust[2,1]+1.96*reg2_robust[2,2])
CI_forborn
library(AER)
# Wald test to test joint null hypothesis. Test = "F" means that we perform
the null hypothesis test.
waldtest(reg1, reg2, test = "F")
plot(reg2$residuals, main = "Residuals of the multiple regression", ylab =
"Residuals")
mean(reg2$residuals)
mean(df$age^4);mean(df$female);mean(df$forborn);mean(df$rural);mean(na.omit(d
f$manag03)^4);mean(na.omit(df$rw)^4)
reg1_bis = lm(log(rw) ~ prinusyr, df)
coeftest.hc1(reg1_bis)
summary(reg1_bis)$r.squared

plot(log(rw) ~ prinusyr, df, main = "Data and regression, salary by date of
arrival")
abline(reg1_bis, col = "blue")
reg2_bis = lm(log(rw) ~ prinusyr + age + I(age^2) + female + rural + educ +
wbhao + manag03, data = df, na.action = na.omit)
reg2_bis_robust <- coeftest.hc1(reg2_bis)
reg2_bis_robust
summary(reg2_bis)$adj.r.square
CI_prinusyr <- c(reg2_bis_robust[2,1]-1.96*reg2_bis_robust[2,2],
```

```r
reg2_bis_robust[2,1]+1.96*reg2_bis_robust[2,2])
CI_prinusyr
# Wald test to test joint null hypothesis. Test = "F" means that we perform
the null hypothesis test.
waldtest(reg1_bis, reg2_bis, test = "F")
plot(reg2_bis$residuals, main = "Residuals of the multiple regression", ylab
= "Residuals", xlab = "Observations")
mean(reg2_bis$residuals)
mean(df$age^4);mean(df$female);mean(df$forborn);mean(na.omit(df$prinusyr)^4);
mean(df$rural);mean(na.omit(df$manag03)^4);mean(na.omit(df$rw)^4)
```