Case Study: The Gender Wage Gap has been Decreasing Over Time?

Based on the slides of Prof. Pedro Raposo.

Miguel Salema

Católica-Lisbon SBE

April 12, 2024

Section 1

Introduction

Setup

Load the necessary packages and data:

```r
# Run this only ONCE
#install.packages(haven)
#install.packages(tidyverse)
#install.packages(stargazer)

library(tidyverse) # core tidyverse packages
library(stargazer) # for the output tables

df <- read_csv("../data/qp_sample4.csv") #load the data
```

Key Variables - Definition

Name	Description
wage_lhr	Reports the real hourly wages in log terms. The hourly wage is measured in euros and it is the ratio between total regular and non-regular payroll (base wage, regular payments, non-regular benefits, and overtime payments) in the reference month and total hours of work (normal and overtime). It was deflated using the Consumer Price Index (with base-year 1986).
firm	Firm identifier.
year	Year information ranging between 1991 and 2017.
male	Dichotomous variable indicating whether the individual is a male.
age	Reports the person's age in years.
tenure	Reports the number of months an employee has worked for his firm.
educ	Reports 8 categories of education ranging from no education to having PhD. Information is compatible through the all period.
job_title	Reports more than 30000 categories of occupations.
sector	Reports 31 categories of industries. Information is compatible through the all period.

View the Data

Lets see how the data looks like:

```
head(df) # see the first rows
## # A tibble: 6 x 9
##    year   firm wage_lhr  male   age tenure  educ job_title sector
##   <dbl>  <dbl>    <dbl> <dbl> <dbl>  <dbl> <dbl>     <dbl>  <dbl>
## 1  1999 470756   0.631      1    45    180     2     22334     13
## 2  2015 681779   0.481      0    20      6     5     64894     22
## 3  2004 213629  -0.0606     0    30    101     2     70452     14
## 4  2013 744477  -0.0816     1    28     24     4    113072     21
## 5  2012 439219   0.579      1    33    153     5    139534     15
## 6  2009 559628   0.135      0    34     93     4    102792     28
```

```
View(df) # this will open df in a separate window
```

Descriptive Statistics

```r
stargazer(as.data.frame(df), # requires data.frame
          digits=1)# rounding
```

Statistic	N	Mean	St. Dev.	Min	Max
year	4,742,129	2,005.1	7.6	1,991	2,017
firm	4,742,129	459,438.0	232,080.0	1	1,064,886
wage_lhr	4,742,129	0.4	0.6	−0.9	7.5
male	4,742,129	0.6	0.5	0	1
age	4,742,129	38.0	10.9	18	64
tenure	4,742,129	99.8	103.8	0	600
educ	4,742,129	3.8	1.6	0	7
job_title	4,742,129	81,124.1	40,628.2	1	163,876
sector	4,742,129	19.1	6.9	3	31

Calculate mean wage difference

Calculate the mean wage difference between male and female:

```
# group_by interacts with summarise
# so, mean() will be calculated for both men and women
df %>%
  group_by(male) %>% # group by the gender variable
  summarise(average_wage = mean(wage_lhr)) # apply the mean
## # A tibble: 2 x 2
##    male average_wage
##   <dbl>        <dbl>
## 1     0        0.293
## 2     1        0.526
```

How do we interpret these log differences?

```
exp(0.5261582 - 0.2934248) - 1
## [1] 0.262045
```

Variance-Covariance Matrix

	wage_lhr	male	age	educ	tenure
tenure	16	2.18	584.43	−24.53	10771.91
educ	0.44	−0.08	−4.03	2.52	−24.53
age	1.32	0.37	117.74	−4.03	584.43
male	0.06	0.24	0.37	−0.08	2.18
wage_lhr	0.34	0.06	1.32	0.44	16

Variance-Covariance Matrix

Code for the variance-covariance plot:

```
library(reshape2)
df %>%
  select(wage_lhr, # select what variables to keep
         male,
         age,
         educ,
         tenure) %>%
  cov() %>% # This is the line that computes the matrix
  round(digits = 2) %>% # round the numbers
  melt() %>% # the rest of the code is for the plot
  as_tibble() %>%
  ggplot(aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value), color = "white", size = 4) +
  scale_fill_gradientn(colours = c("#A0C7BE", "#708090"), trans = "pseudo_log") +
  labs(x = "",
       y = "") +
  my_theme +
  theme(legend.position = "none")
```

Correlation Matrix

Correlation Matrix

Code for the variance-covariance plot:

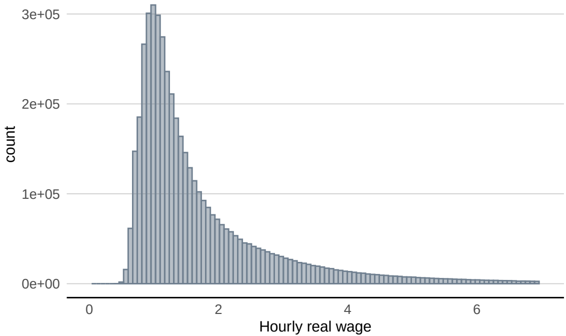
```
library(reshape2)
df %>%
  select(wage_lhr, # select what variables to keep
         male,
         age,
         educ,
         tenure) %>%
  cor() %>% # This is the line that computes the matrix
  round(digits = 2) %>% # round the numbers
  melt() %>% # the rest of the code is for the plot
  ggplot(aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  geom_text(aes(Var2, Var1, label = value), color = "white", size = 4) +
  scale_fill_gradientn(colours = c("#A0C7BE", "#708090")) +
  labs(x = "",
       y = "",
       fill = "Correlation") +
  my_theme +
  theme(legend.position = "right")
```

Graph the Data: Log Wage Distribution

```r
df %>%
  ggplot(aes(x = wage_lhr)) +
  geom_histogram(bins = 100, color= "slategray", fill = "slategrey", alpha=0.6) +
  xlim(-1, 3) +
  labs(x = "Log of hourly real wage") +
  my_theme
```

Graph the Data: Wage Distribution

```
df %>%
  ggplot(aes(x = exp(wage_lhr) )) + #We can the variable directly
  geom_histogram(bins = 100, color= "slategray", fill = "slategrey", alpha=0.5) +
  xlim(0, 7) +
  labs(x = "Hourly real wage") +
  my_theme
```



Men and Women Comparison

Men and Women Comparison

Code for the previous plot:

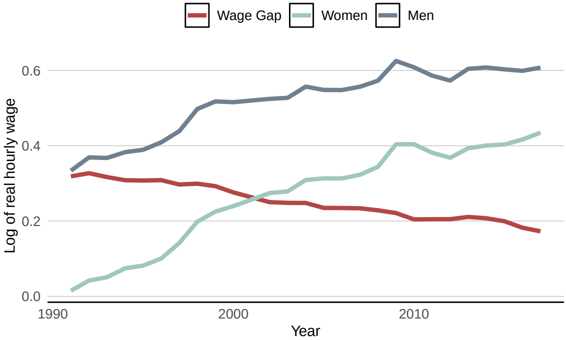
```
df %>%
  mutate(male = as.character(male)) %>%
  ggplot(aes(x=wage_lhr,
             color= male,
             fill= male)) +
  geom_histogram(aes(y=..density..),
                 alpha=0.5,
                 position="identity",
                 bins = 50) +
  geom_density(size = 1, alpha=0) +
  xlim(-1, 3) +
  scale_color_manual(values = c("#b34745", "#708090"),
                     labels = c("Women","Men"),
                     name = "") +
  scale_fill_manual(values = c("#b34745", "#708090"),
                     labels = c("Women","Men"),
                     name = "") +
  labs(x = "Log of hourly real wage") +
  my_theme
```

Raw Gender Wage Gap Over Time

```r
df_wage_gap <- df %>%
  group_by(male,
           year) %>%
  summarise(average_wage = mean(wage_lhr)) %>%
  pivot_wider(names_from = male,
              values_from = average_wage,
              names_prefix = "wage_") %>%
  mutate(diff = wage_1 - wage_0)
```

```r
View(df_wage_gap) #open this dataset
```

Plot the Gender Wage Gap

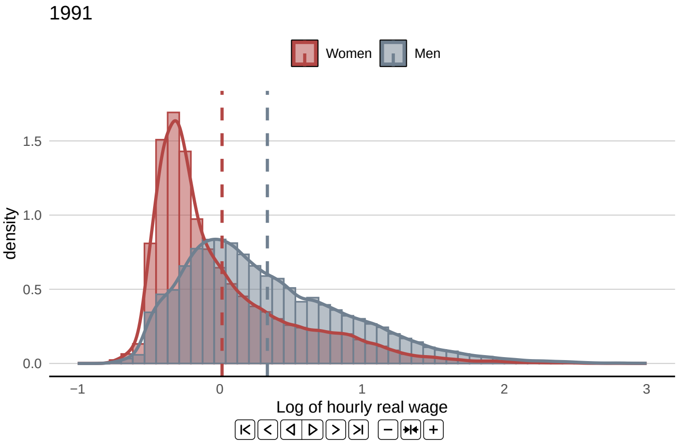


Evolution of the Gender Wage Gap

Code for the previous plot:

```
df_wage_gap %>%
  pivot_longer(!year, names_to = "series", values_to = "wage") %>%
  ggplot(aes(x=year, y=wage, color=series)) +
  geom_line(size = 1.5) +
  scale_color_manual(values = c("#b34745", "#A0C7BE", "#708090"),
                     labels=c("Wage Gap", "Women", "Men"),
                     name="") +
  labs(y = "Log of real hourly wage",
       x = "Year") +
  my_theme
```

Wage Distribution Over Time



Section 2

Using Linear Models

Deriving the OLS Estimates

Estimate the relationship between wages and education with a linear model:

$$wage_i = \beta_0 + \beta_1 male_i + \varepsilon_i$$

In a univariate regression, we can estimate β_1 with a simple fraction:

$$\hat{\beta}_1 = \frac{cov(wage, male)}{var(male)}$$

Deriving the OLS Estimates

Calculate $\hat{\beta}_1$ in R:

```
beta1 <- cov(df$wage_lhr, df$male)/var(df$male)
beta1
## [1] 0.2327335
```

Notice that $\hat{\beta}_1$ is not the same as a correlation, but it's similar:

```
cor(df$wage_lhr, df$male)
## [1] 0.1971039
```

Deriving the OLS Estimates

To estimate β_0 in a univariate regression:

$$\hat{\beta}_0 = \bar{wage} - \hat{\beta}_1 \bar{male}$$

Estimate $\hat{\beta}_0$ in R:

```
mean(df$wage_lhr)- beta1 * mean(df$male)
## [1] 0.2934248
```

Univariate Linear Model

There is no need to compute the coefficients by hand. We can use the `lm` function:

```
m1 <- lm(wage_lhr ~ male, data = df)
m1
##
## Call:
## lm(formula = wage_lhr ~ male, data = df)
##
## Coefficients:
## (Intercept)         male
##      0.2934       0.2327
```

Notice the values are the same as the ones we calculated before.

Without Independent Variables

If we estimate a regression without independent variables we calculate the average of the dependent variable.

$$wage_i = \beta_0 + \varepsilon_i$$

Run a model without independent variables:

```
m2 <- lm(wage_lhr ~ 1 , data=df)
m2
##
## Call:
## lm(formula = wage_lhr ~ 1, data = df)
##
## Coefficients:
## (Intercept)
##      0.4264
```

It's the same as computing the mean:

```
mean(df$wage_lhr)
## [1] 0.4264361
```

Without Independent Variables

Compare with a model that has gender for independent variable:

```
stargazer(m1, m2)
```

	Dependent variable:	
	wage_lhr	
	(1)	(2)
male	0.233^{***}	
	(0.001)	
Constant	0.293^{***}	0.426^{***}
	(0.0004)	(0.0003)
Observations	4,742,129	4,742,129
R^2	0.039	0.000
Adjusted R^2	0.039	0.000
Residual Std. Error	0.573 (df = 4742127)	0.584 (df = 4742128)
F Statistic	$191{,}678.100^{***}$ (df = 1; 4742127)	
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Section 3

Case Study: The Gender Wage Gap has been Decreasing Over Time?

Going Beyond BRM

Using data from Quadros de Pessoal.

Table 1: Gender wage gap: 5 regressions/models

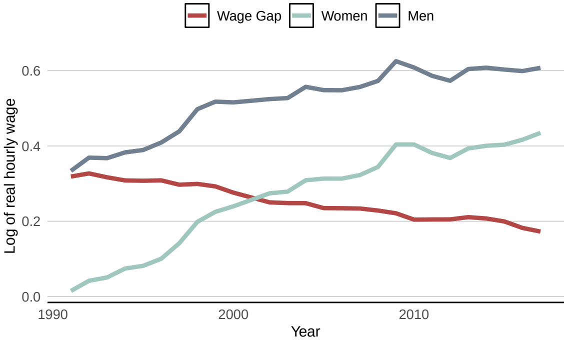
	(1)	(2)	(3)	(4)	(5)
Male	0.233	0.278	0.216	0.191	0.117
Control variables:					
Age		Yes	Yes	Yes	Yes
Age2		Yes	Yes	Yes	Yes
Tenure		Yes	Yes	Yes	Yes
Tenure2		Yes	Yes	Yes	Yes
Education dummies		Yes	Yes	Yes	Yes
Year dummies		Yes	Yes	Yes	Yes
Industry dummies			Yes		
Firm dummies				Yes	Yes
Job title dummies					Yes
No. observations	4,742,129	4,742,129	4,742,129	4,742,129	4,550,039
Adjusted R2	0.0388	0.4551	0.5163	0.6916	0.7667

Going Beyond BRM

	(1)	(2)	(3)	(4)	(5)
1991*Male	0.319	0.265	0.190	0.161	0.093
1992*Male	0.327	0.273	0.198	0.171	0.100
1993*Male	0.317	0.269	0.196	0.166	0.092
1994*Male	0.308	0.269	0.201	0.173	0.100
1995*Male	0.308	0.274	0.207	0.183	0.111
1996*Male	0.309	0.280	0.211	0.183	0.111
1997*Male	0.297	0.283	0.214	0.188	0.118
1998*Male	0.299	0.287	0.222	0.195	0.124
1999*Male	0.293	0.285	0.222	0.201	0.132
2000*Male	0.276	0.281	0.219	0.199	0.131
2002*Male	0.250	0.280	0.222	0.203	0.134
2003*Male	0.248	0.282	0.225	0.206	0.135
2004*Male	0.248	0.286	0.227	0.205	0.123
2005*Male	0.235	0.280	0.222	0.202	0.124
2006*Male	0.235	0.285	0.225	0.204	0.129
2007*Male	0.234	0.290	0.233	0.209	0.132
2008*Male	0.229	0.291	0.233	0.208	0.129
2009*Male	0.221	0.283	0.225	0.200	0.122
2010*Male	0.204	0.272	0.212	0.187	0.110
2011*Male	0.205	0.271	0.212	0.185	0.109
2012*Male	0.205	0.267	0.207	0.180	0.105
2013*Male	0.211	0.280	0.218	0.190	0.116
2014*Male	0.207	0.282	0.220	0.189	0.116
2015*Male	0.200	0.279	0.216	0.186	0.115
2016*Male	0.182	0.266	0.206	0.178	0.110
2017*Male	0.173	0.263	0.203	0.177	0.110
Control variables:					
Age		Yes	Yes	Yes	Yes
Age2		Yes	Yes	Yes	Yes
Tenure		Yes	Yes	Yes	Yes
Tenure2		Yes	Yes	Yes	Yes
Education dummies		Yes	Yes	Yes	Yes
Industry dummies			Yes		
Firm dummies				Yes	Yes
Job title dummies					Yes
No. observations	4,742,129	4,742,129	4,742,129	4,742,129	4,550,039
Adjusted R2	0.0388	0.4551	0.5163	0.6916	0.7667

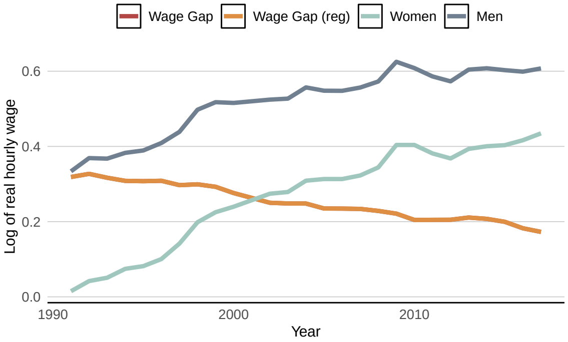
Plot the Raw Gender Wage Gap

In the 1st class we plotted the average differences of hourly wage between men and women:



Visualize the Uncontrolled Model

What happens if we add the coefficients from the model 1 in page 31 to the graph? We can see that the regression coefficients match exactly the mean difference.



Visualize the Uncontrolled Model

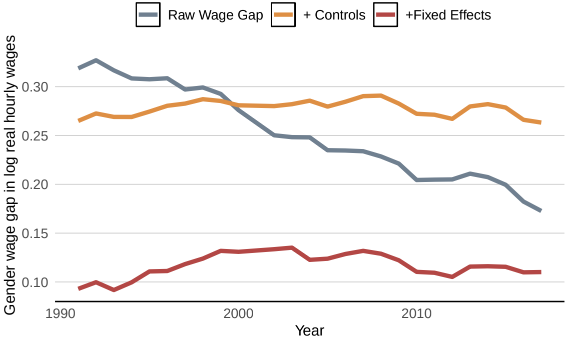
Code for the previous plot:

```
# model 1 of page 31:
m_raw <- lm(wage_lhr ~ as.factor(year) + as.factor(year):male - 1,
            data = df)
# -1 will remove the intercept

df_wage_gap %>%
  pivot_longer(!year, names_to = "series", values_to = "wage") %>%
  # add the regression coeffs to the data
  bind_rows(tibble(wage = m_raw$coefficients[27:52],
                   year = unique(df$year) %>% sort(),
                   series = "reg diff")) %>%
  ggplot(aes(x=year, y=wage, color=series)) +
  geom_line(size = 1.5) +
  scale_color_manual(values = c("#b34745", "#de8f44", "#A0C7BE", "#708090"),
                     labels=c("Wage Gap", "Wage Gap (reg)", "Women", "Men"),
                     name="") +
  labs(y = "Log of real hourly wage",
       x = "Year") +
  my_theme
```

Compare the models

What happens if we add the coefficients from the model 1, 2 and 5 in page 31 to the graph? We can see that the regression coefficients almost did not change over time for model 2 and 5.



Compare the models

code for the previous plot:

```
# model 1:
m_raw <- lm(wage_lhr ~ as.factor(year) + as.factor(year):male - 1,
            data = df)
# -1 will remove the intercept

# model 2:
m_inter <- lm(wage_lhr ~ as.factor(year) + as.factor(year):male - 1 + age + I(age^2)
              data = df)

# library for the high dimentional fixed effects
library(fixest)
m_full <- feols(wage_lhr ~ as.factor(year):male + age + I(age^2) + tenure + I(tenure
              data = df)

# identify the coefficients of interest in each object
select_raw <- names(m_raw$coefficients) %>% str_detect("male")
select_inter <- names(m_inter$coefficients) %>% str_detect("male")
select_full <- names(m_full$coefficients) %>% str_detect("male")
# these select vectors are logical (TRUE/FALSE), indicating whether the coefficient
```

Compare the models

```r
# plot the selected coefficients. Build a tidy dataset
df_plot <- tibble(
  # keep only the selected coefficients
  coeffs = c(m_raw$coefficients[select_raw],
             m_inter$coefficients[select_inter],
             m_full$coefficients[select_full]),
  # repeat the year sequence 3 times
  year = unique(df$year) %>% sort() %>% rep(times = 3),
  # each serie has the size of the number of years
  series = c("raw", "inter", "full") %>%
    rep(each = n_distinct(df$year)) %>%
    fct_inorder()
)

df_plot %>%
  ggplot(aes(x=year, y=coeffs, color=series)) +
  geom_line(linewidth = 1.5) +
  scale_color_manual(values = c("slategrey", "#de8f44", "#b34745"),
                     labels=c("Raw Wage Gap", "+ Controls", "+Fixed Effects"),
                     name="") +
  labs(y = "Gender wage gap in log real hourly wages",
       x = "Year") +
  my_theme
```

What explains the unexplained part?

- Employers' discrimination?
- Discrimination or prejudice? (Flabbi 2010)
- Greedy jobs? (Goldin 2021)
- Omitted variable bias?

References

Flabbi, Luca. 2010. "Gender Discrimination Estimation in a Search Model with Matching and Bargaining." *International Economic Review* 51 (3): 745–83.

Goldin, Claudia. 2021. *Career and Family: Women's Century-Long Journey Toward Equity*. Princeton University Press.