Data Analysis 2 - ECBS5142 - Assignment 1

In this assignement, we will try to discover the gender wage gap with the level of education for occupations in Production (occ2012 code from 7700 to 8965).

Table of Contents

In [1]:
```
%%capture
!pip install -r requirements.txt;
```

In [2]:
```
# import libs
import os
import sys
import warnings

import numpy as np
import pandas as pd
from mizani.formatters import percent_format
from plotnine import *
from datetime import datetime
import statsmodels.api as sm
import statsmodels.formula.api as smf
from scipy.stats import norm, chisquare
from IPython.core.display import HTML
from stargazer.stargazer import Stargazer
import statsmodels.nonparametric.kernel_regression as loess

from mizani.transforms import log_trans
from mizani.formatters import percent_format
from mizani.formatters import log_format

warnings.filterwarnings("ignore")
```

In [3]:
```
# Import the prewritten helper functions
from py_helper_functions import *
```

In [4]:
```
# read the data from the csv file
all_df = pd.read_csv('morg-2014-emp.csv')
```

In [5]:
```
# Filter the data for occ2012 between 7700 and 8965
comp_sample = all_df[(all_df['occ2012'] >= 7700) & (all_df['occ2012'] <= 8965)][['hhid',
#drop the all_df
del(all_df)
```

In [6]:
```
# Add a column 'hourly_wage' to the DataFrame
comp_sample['hourly_wage'] = comp_sample['earnwke'] / comp_sample['uhours']
```

In [7]:
```
# Add the natural log of wage (ln_wage) column
comp_sample['ln_wage'] = np.log(comp_sample['hourly_wage'])
```

In [8]:
```
# add column female to have boolean for male or female
```

```
comp_sample['female'] = comp_sample['sex'].apply(lambda x: 1 if x == 2 else 0)
```

In [9]:
```
# Add the sex_text column for descriptive values
comp_sample['sex_text'] = comp_sample['female'].apply(lambda x: '[1] female' if x == 1 e
```

In [10]:
```
# Describe the comp_sample
comp_sample.info()
```

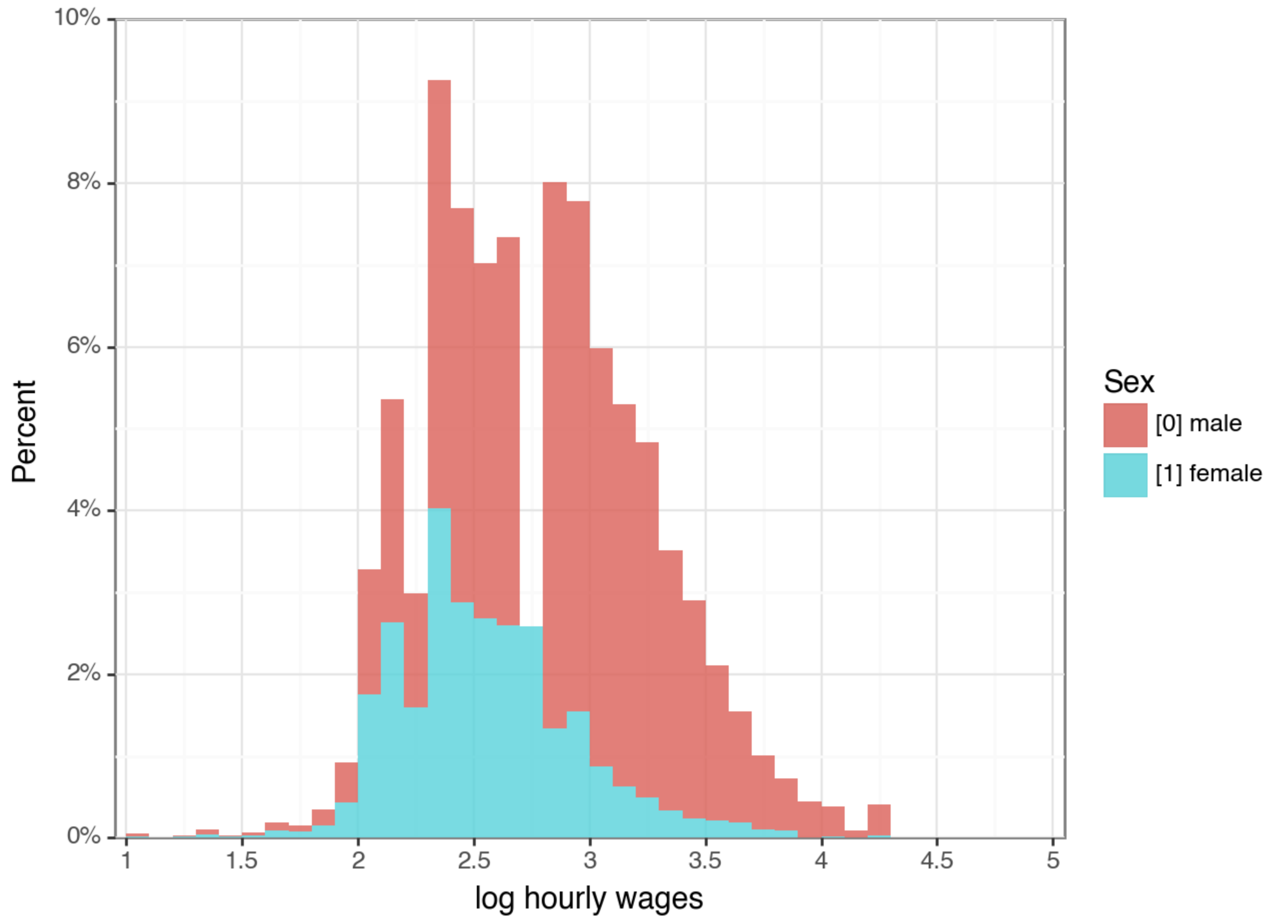
```
<class 'pandas.core.frame.DataFrame'>
Index: 9205 entries, 2 to 149293
Data columns (total 10 columns):
 #   Column       Non-Null Count  Dtype  
---  ------       --------------  -----  
 0   hhid         9205 non-null   int64  
 1   earnwke      9205 non-null   float64
 2   uhours       9205 non-null   int64  
 3   grade92      9205 non-null   int64  
 4   sex          9205 non-null   int64  
 5   occ2012      9205 non-null   int64  
 6   hourly_wage  9205 non-null   float64
 7   ln_wage      9205 non-null   float64
 8   female       9205 non-null   int64  
 9   sex_text     9205 non-null   object 
dtypes: float64(3), int64(6), object(1)
memory usage: 791.1+ KB
```

Comment

In the Production Occupations sample, we have a total of 9205 observations, none of which has missing values. Let's examine the distribution of the sample.

Distribution of wage by gender

In [11]:
```
(
    ggplot(comp_sample, aes(x="ln_wage", y="stat(count)/sum(stat(count))", fill='factor(
    + geom_histogram(
        binwidth=0.1,
        boundary=0,
        size=0.25,
        alpha=0.8,
        show_legend=True,
        na_rm=True,
    )
    + labs(x="log hourly wages", y="Percent", fill="Sex")
    + expand_limits(x=0.01, y=0.01)
    + scale_x_continuous(expand=(0.01, 0.01), limits=(1, 5), breaks=seq(1, 5, 0.5))
    + scale_y_continuous(
        expand=(0.0, 0.0),
        limits=(0, 0.1),
        breaks=seq(0, 0.1, 0.02),
        labels=percent_format(), #mizani
    )
    + theme_bw()
)
```



Out[11]: <Figure Size: (640 x 480)>

Comment

From the above histogram, it seems like in our sample, there are a lot more observations for male's wage than female's wage (almost 3 to 1!). We see that the Production Occupations mighty male-dominant.

The unconditional gender gap

Here we will examine the hourly wage gap between male and female in our sample.

Reg1 - Regression of ln(wage) on gender

In [12]:
```
reg1 = smf.ols(formula="ln_wage~female", data=comp_sample).fit(cov_type="HC1")
reg1.summary()
```

Out[12]:

OLS Regression Results

Dep. Variable:	ln_wage	**R-squared:**	0.073
Model:	OLS	**Adj. R-squared:**	0.073
Method:	Least Squares	**F-statistic:**	832.3
Date:	Mon, 20 Nov 2023	**Prob (F-statistic):**	2.78e-175
Time:	02:45:51	**Log-Likelihood:**	-6553.2
No. Observations:	9205	**AIC:**	1.311e+04
Df Residuals:	9203	**BIC:**	1.312e+04

Df Model:	1
Covariance Type:	HC1

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	2.8461	0.006	453.203	0.000	2.834	2.858
female	-0.3097	0.011	-28.849	0.000	-0.331	-0.289

Omnibus:	3731.807	**Durbin-Watson:**	1.938
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	89269.591
Skew:	-1.396	**Prob(JB):**	0.00
Kurtosis:	17.998	**Cond. No.**	2.44

Notes:
[1] Standard Errors are heteroscedasticity robust (HC1)

Analysis Explanation

From the regression, we see that the P value is 0.000, which indicates that we have strong evidences to reject the hypothesis that there is no difference in wage between male and female.

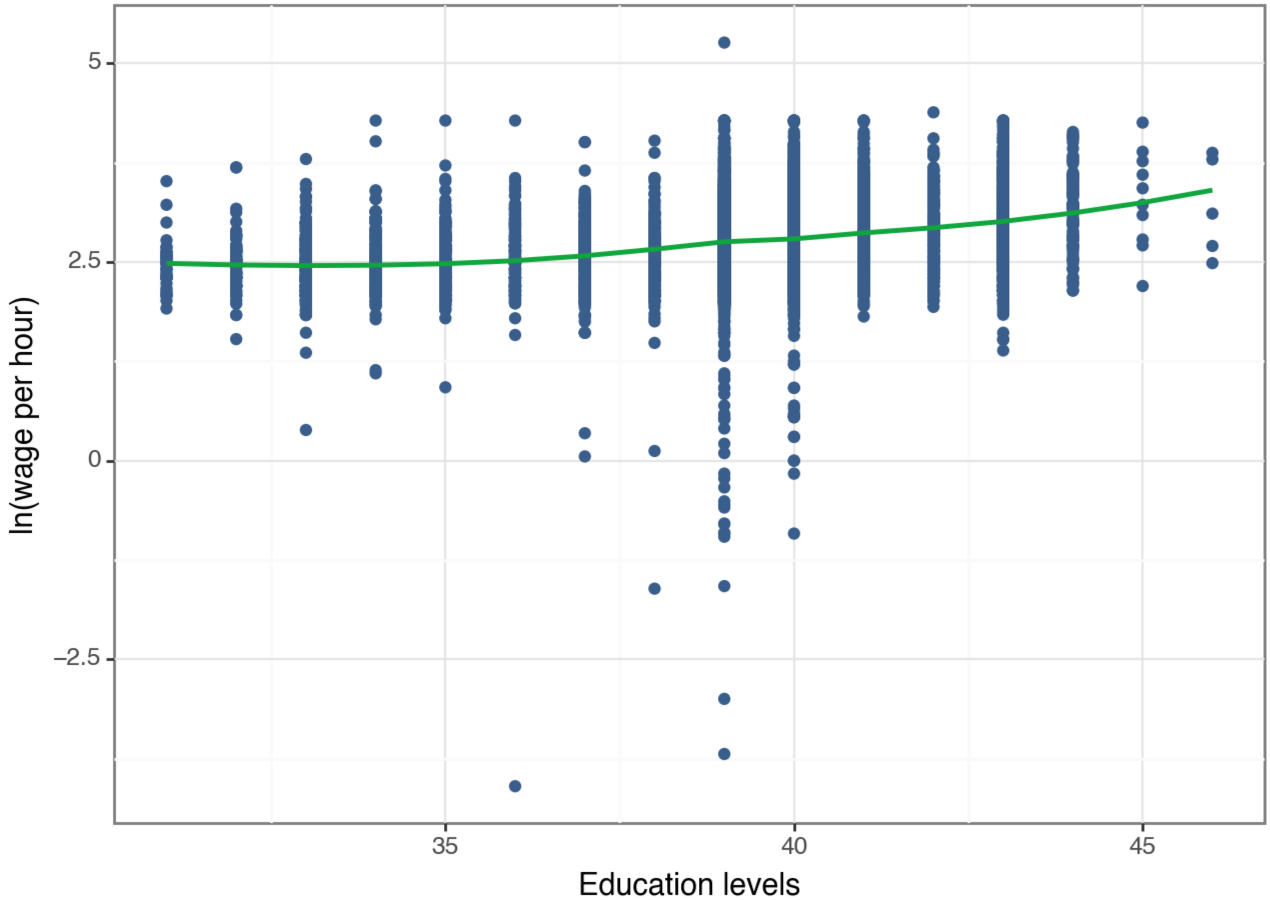
The coefficient of -0.3097 and the standard error of 0.011 suggest that female employees earn between 29.8%-32.0% less than male employees. If our sample is a representative sample, this behavior is very likely to happen in the population since our p-value is < 0.01.

However, the R-squared for this model is only 0.073, which means the model only explains about 7.3% of the variance in the wage. There might be other factors that contribute to the wage aside from gender.

The gender wage gap and education level

Let's plot a scatter plot and see how the loess regression looks like:

In [13]:
```
(
    ggplot(comp_sample, aes(x='grade92', y='ln_wage')) +
    geom_point(color=color[0]) +
    geom_smooth(method='loess', color=color[1], se=False) +
    labs(x="Education levels", y="ln(wage per hour)") +
    theme_bw()
)
```



Out[13]: <Figure Size: (640 x 480)>

Comment

From the loess regression, it seems like the education level below 36 (9th grade or below) does not have a lot of impact on wage. Starting from level 36 (High school and above), there is a positive correlation between education level and wage as the loess line is going up.

Let's construct some regression models to examine the gender wage gap with the level of education:

Reg2 - Regression of ln(wage) on gender and level of education

In [14]:
```
reg2 = smf.ols(formula="ln_wage~female+grade92", data=comp_sample).fit(cov_type="HC1")
reg2.summary();
```

Reg3 - Regression of level of education on gender

In [15]:
```
reg3 = smf.ols(formula="grade92~female", data=comp_sample).fit(cov_type="HC1")
reg3.summary();
```

Table: Gender wage gap and level of education – different specifications

In [16]:
```
stargazer = Stargazer([reg1, reg2, reg3])
stargazer.custom_columns(['ln(hourly wage)', 'ln(hourly wage)', 'education level'], [1,
stargazer.show_model_numbers(True)
stargazer.cov_spacing = 1.5
stargazer.covariate_order(
    [
        "female",
        "grade92",
```

```
        "Intercept"
    ]
)
stargazer.rename_covariates(
    {
        "Intercept": "Constant",
        "grade92": "education level"
    }
)
stargazer
```

Out[16]:

	ln(hourly wage)	ln(hourly wage)	education level
	(1)	(2)	(3)
female	-0.310^{***}	-0.290^{***}	-0.378^{***}
	(0.011)	(0.010)	(0.056)
education level		0.053^{***}	
		(0.002)	
Constant	2.846^{***}	0.754^{***}	39.291^{***}
	(0.006)	(0.079)	(0.026)
Observations	9205	9205	9205
R^2	0.073	0.128	0.006
Adjusted R^2	0.073	0.128	0.005
Residual Std. Error	0.493 (df=9203)	0.478 (df=9202)	2.252 (df=9203)
F Statistic	832.273^{***} (df=1; 9203)	866.432^{***} (df=2; 9202)	45.563^{***} (df=1; 9203)
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Analysis Explanation

From model (1), female in this sample earn 30%-32% less than male. This is significant at 1%.

From model (2), female with the same level of education earn 28%-30% less than male. This is significant at 1%.

Comparing the coefficient of female between the models (1) and (2), there is a slight difference. The omitted variable bias is

$$-0.310 - (-0.290) = -0.02$$

From model (3), there is a negative correlation between level of education and female (female's level of education is lower than male's level of education), which result in the difference that we see between the models (1) and (2) coefficients. The diffence of -0.02 should be equal to the product of model (3) female's coefficient and model (2) level of education coefficient. It is indeed equal:

$$-0.378 * 0.053 = -0.020034 \approx -0.02$$

But how significant is the inclusion of level of education changes the wage gap between male and female?

We can see from the models (1) and (2), the point estimate of each coefficients (-0.310 & -0.290) are each outside of the CI of the other ([-0.321, -0.299] & [-0.30, -0.28]). However, the 2 CIs are overlapped. We should do a formal test to decide if the level of education significantly change the wage gap.

In [17]:
```
# Coefficients of 'female' from both models
coef_female_reg1 = reg1.params['female']
coef_female_reg2 = reg2.params['female']

# Standard errors of 'female' from both models
se_female_reg1 = reg1.bse['female']
se_female_reg2 = reg2.bse['female']

# Calculate the difference in coefficients and the standard error of this difference
diff_coef = coef_female_reg1 - coef_female_reg2
diff_se = np.sqrt(se_female_reg1**2 + se_female_reg2**2)

# Calculate the t-statistic for the difference
t_stat = diff_coef / diff_se

# Calculate the p-value for the t-statistic
p_value = 2 * (1 - norm.cdf(np.abs(t_stat)))

print(f"The p-value when testing if the coef of female in reg1 and reg2 are the same: {p
```

```
The p-value when testing if the coef of female in reg1 and reg2 are the same: 0.17984167
406673057
```

The p-value 0.18 > 0.05, meaning we do not have sufficient evidence to reject that the two coefficients are the same in the two models (1) and (2). In other words, in the population, the wage gap between male and female might not significantly change after including the difference in level of education.

In [ ]:
```
```