HW1 - Jonas Nepozitek

Setup

In [ ]:

```
"""
Homework 1 Gender Wage Gap
Jackson School of Global Affairs

 Created by Ardina Hasanbasri for GLBL 5021

Additional reference code and data used:
Békés & Kézdi (2021) see more code below
https://gabors-data-analysis.com/

"""

# Note: Feel free to move this code to jupyter notebook if you prefer.

#----------------------------------#
# SETTING UP YOUR WORKSPACE #
# ----------------------------------#

# These imports are similar to lecture 1.2 code (feel free to add or delete)
import warnings
import numpy as np
import pandas as pd
import pyfixest as pf
import seaborn as sns
import matplotlib.pyplot as plt
import pyreadstat
from statsmodels.nonparametric.smoothers_lowess import lowess
import statsmodels.formula.api as smf
from typing import List

cps= pd.read_csv("https://osf.io/download/4ay9x/")
```

```
C:\Users\nepoz\AppData\Local\Temp\ipykernel_22748\375386066.py:31: DtypeWarning: Columns (16) have mixed types. Speci
fy dtype option on import or set low_memory=False.
  cps= pd.read_csv("https://osf.io/download/4ay9x/")
```

1.1

In [2]:

```
#-------------------------------------------------------------#
# Sample Selection and Creating New Data                      #
# ------------------------------------------------------------#

cps = cps.query("uhours>=20 & earnwke>0 & age>=24 & age<=64")

# Create variables
cps["female"] = (cps.sex == 2).astype(int)
cps["w"] = cps["earnwke"] / cps["uhours"]
cps["lnw"] = np.log(cps["w"])

# Add demographic variables
cps["white"] = (cps["race"] == 1).astype(int)
cps["afram"] = (cps["race"] == 2).astype(int)
cps["asian"] = (cps["race"] == 4).astype(int)
cps["hisp"] = (cps["ethnic"].notna()).astype(int)
cps["othernonw"] = (
    (cps["white"] == 0) & (cps["afram"] == 0) & (cps["asian"] == 0) & (cps["hisp"] == 0)
).astype(int)
cps["nonUSborn"] = (
    (cps["prcitshp"] == "Foreign Born, US Cit By Naturalization")
    | (cps["prcitshp"] == "Foreign Born, Not a US Citizen")
).astype(int)

cps["married"] = ((cps["marital"] == 1) | (cps["marital"] == 2)).astype(int)
cps["divorced"] = ((cps["marital"] == 3) & (cps["marital"] == 5)).astype(int)
cps["wirowed"] = (cps["marital"] == 4).astype(int)
cps["nevermar"] = (cps["marital"] == 7).astype(int)

cps["child0"] = (cps["chldpres"] == 0).astype(int)
cps["child1"] = (cps["chldpres"] == 1).astype(int)
cps["child2"] = (cps["chldpres"] == 2).astype(int)
cps["child3"] = (cps["chldpres"] == 3).astype(int)
cps["child4pl"] = (cps["chldpres"] >= 4).astype(int)
```

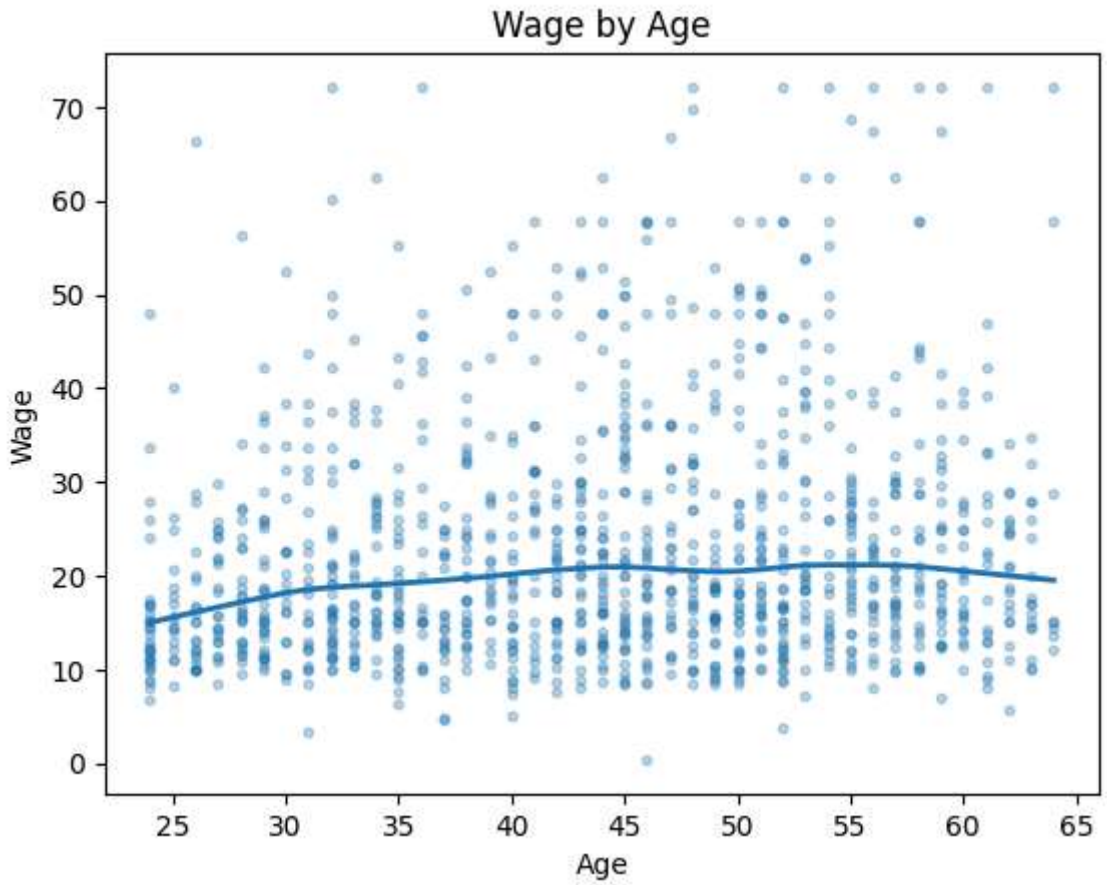
```
# Now let's select an industry to work with.
# The code below shows the industry codes and their counts for first 25 rows.
cps["ind02"].value_counts(dropna=False).to_frame()[1:25]
cps = cps.query('ind02=="Motor vehicles and motor vehicle equipment manufacturing (3361, 3362, 3363)"')
```

1.2

In [3]:
```
lowess_fit = lowess(
    endog=cps['w'],
    exog=cps['age'],
    frac=0.3   # smoothing parameter
)

plt.scatter(cps["age"], cps["w"], alpha=0.3, s=10)
plt.plot(lowess_fit[:, 0], lowess_fit[:, 1], linewidth=2)
plt.xlabel("Age")
plt.ylabel("Wage")
plt.title("Wage by Age")
```

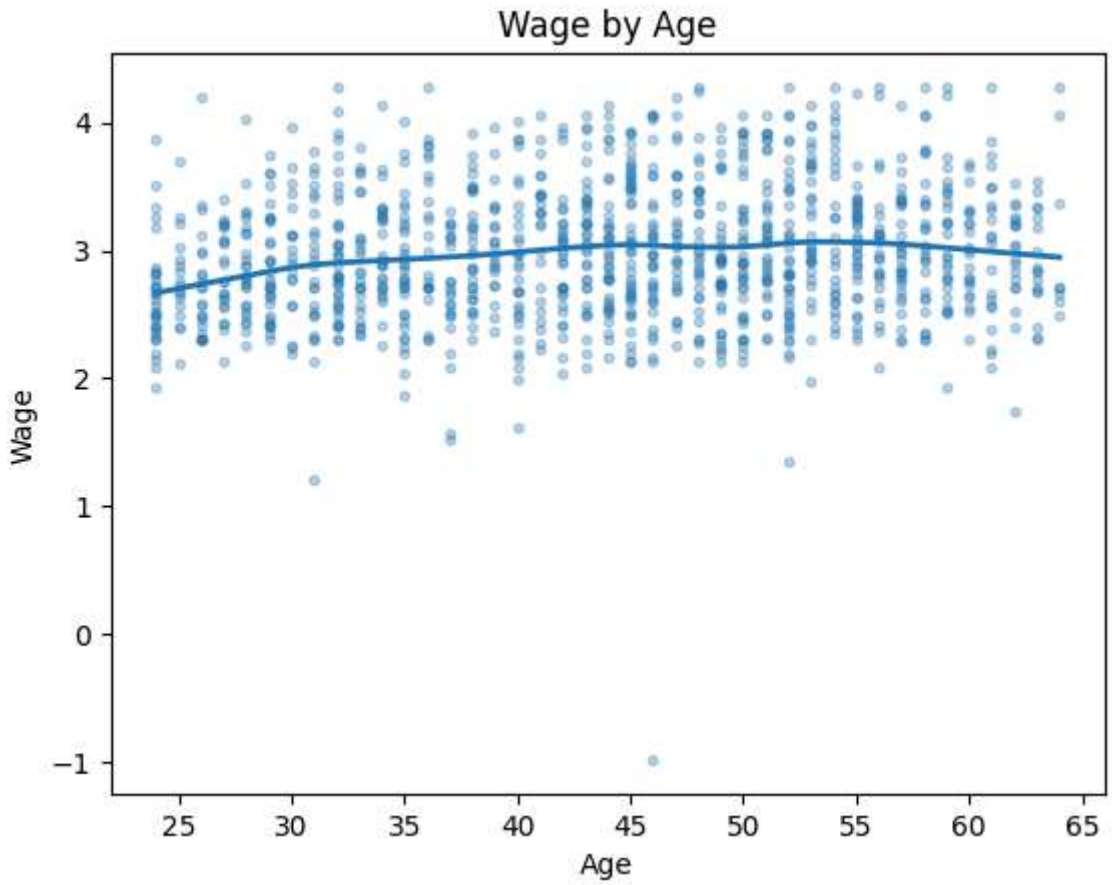
Out[3]:
```
Text(0.5, 1.0, 'Wage by Age')
```



In [4]:
```
lowess_fit = lowess(
    endog=cps['lnw'],
    exog=cps['age'],
    frac=0.3   # smoothing parameter
)

plt.scatter(cps["age"], cps["lnw"], alpha=0.3, s=10)
plt.plot(lowess_fit[:, 0], lowess_fit[:, 1], linewidth=2)
plt.xlabel("Age")
plt.ylabel("Wage")
plt.title("Wage by Age")
```

Out[4]:
```
Text(0.5, 1.0, 'Wage by Age')
```



We can see that wage slowly increases for older workers. However, after a certain age is reached, wage might start falling again.

1.3

In [5]:
```
cps["age2"] = cps["age"] ** 2
cps["age3"] = cps["age"] ** 3
```

In [6]:
```
group_A_codes = [31, 32, 33, 34, 35, 36, 37, 38]
group_B_codes = [39]
group_C_codes = [40, 41, 42]
group_D_codes = [43]
group_E_codes = [44, 45, 46]
```

```
mapping = {}
mapping.update({c: 'less_than_hs' for c in group_A_codes})
mapping.update({c: 'hs_grad' for c in group_B_codes})
mapping.update({c: 'some_college_associate' for c in group_C_codes})
mapping.update({c: 'college_grad' for c in group_D_codes})
mapping.update({c: 'post_grad' for c in group_E_codes})

cps['educ'] = cps['grade92'].map(mapping).fillna('other')
```

1.4

In [7]:
```
cps['educ'].value_counts(normalize=True)
```

Out[7]:
```
educ
hs_grad                  0.413882
some_college_associate   0.295630
college_grad             0.154242
post_grad                0.076264
less_than_hs             0.059983
Name: proportion, dtype: float64
```

Most people (41.4%) have high school as the highest level of education, 29.6% of people have some college/associate degree, 15.4% are college grads, 7.6% have post graduate degrees, and 6% of people don't even have high school education in the dataset.

2.1

In [8]:
```
reg_w_f = pf.feols("lnw~female", data=cps,vcov="HC1")
reg_w_fa = pf.feols("lnw~female+age", data=cps,vcov="HC1")
reg_w_faa = pf.feols("lnw~female+age+age2", data=cps,vcov="HC1")
reg4_w_faaa = pf.feols("lnw~female+age+age2+age3", data=cps,vcov="HC1")
```

In [9]:
```
pf.etable([reg_w_f, reg_w_fa, reg_w_faa, reg4_w_faaa],
          head_order="h",
          model_heads=["ln wage","ln wage", "ln wage", "ln wage"],
          labels={"Intercept": "Constant"}
)
```

Out[9]:

	ln wage			
	(1)	(2)	(3)	(4)
female	-0.213*** (0.032)	-0.225*** (0.032)	-0.231*** (0.032)	-0.231*** (0.032)
age		0.008*** (0.001)	0.050*** (0.011)	0.073 (0.065)
age2			-0.000*** (0.000)	-0.001 (0.002)
age3				0.000 (0.000)
Constant	3.048*** (0.018)	2.707*** (0.056)	1.859*** (0.215)	1.548 (0.862)
Observations	1167	1167	1167	1167
S.E. type	hetero	hetero	hetero	hetero
R^2	0.033	0.059	0.071	0.071
Adj. R^2	0.032	0.058	0.068	0.068

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

a)

female: On average female automotive workers have 22.5% lower wages than male automotive workers, holding age constant.

age: The average predicted increase in wages as age increases by one is 0.8% among people, controlling for their gender (in this case, this would be male automotive workers).

b)

I would prefer the second regression if I was to analyse a possible gender gap, as it would allow me to capture a possible confounding variable (age). Adding age_squared and age_cubed does not improve the predictive power. Plus, the results in Model 4 are not significant.

2.2

In [10]:
```
reg_w_fe = pf.feols("lnw~female+educ", data=cps,vcov="HC1")
```

In [11]:
```
pf.etable([reg_w_fe],
          head_order="h",
          model_heads=["ln wage"],
          labels={"Intercept": "Constant"}
)
```

Out[11]:

	ln wage
	(1)
female	-0.191***
	(0.028)
educ[T.hs_grad]	-0.441***
	(0.043)
educ[T.less_than_hs]	-0.784***
	(0.055)
educ[T.post_grad]	0.363***
	(0.064)
educ[T.some_college_associate]	-0.291***
	(0.044)
Constant	3.330***
	(0.038)
Observations	1167
S.E. type	hetero
R^2	0.285
Adj. R^2	0.282

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

On average, automotive workers have 78.4% lower wages when not being high school graduates compared to college grads, controlling for their gender (in this case, it is male automotive workers). Similarly, it is 44.1% less for high school grads, 29.1% less for people with some college/associate degree, and 36.3% more for people with post-graduate degrees compared to college grads, controlling for their gender.

2.3

In [12]:
```
reg_w_f = pf.feols("lnw~female", data=cps,vcov="HC1")
reg_w_fa = pf.feols("lnw~female+age", data=cps,vcov="HC1")
reg_w_fae = pf.feols("lnw~female+age+educ", data=cps,vcov="HC1")
```

In [13]:
```
pf.etable([reg_w_f, reg_w_fa, reg_w_fae],
          head_order="h",
          model_heads=["ln wage","ln wage", "ln wage"],
          labels={"Intercept": "Constant"}
)
```

Out[13]:

	ln wage		
	(1)	(2)	(3)
female	-0.213***	-0.225***	-0.202***
	(0.032)	(0.032)	(0.028)
age		0.008***	0.007***
		(0.001)	(0.001)
educ[T.hs_grad]			-0.454***
			(0.042)
educ[T.less_than_hs]			-0.791***
			(0.055)
educ[T.post_grad]			0.333***
			(0.065)
educ[T.some_college_associate]			-0.300***
			(0.043)
Constant	3.048***	2.707***	3.016***
	(0.018)	(0.056)	(0.059)
Observations	1167	1167	1167
S.E. type	hetero	hetero	hetero
R^2	0.033	0.059	0.309
Adj. R^2	0.032	0.058	0.306

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

The gender wage gap (the absolute value of the female coefficient) increases when adding age as a variable in the multilinear regression, indicating that age acted as a suppressor of the magnitude of the wage gap.

The gender wage gap decreases when adding the education category, indicating that some of the gap can be explained by differing education levels between men and women, with men potentially having higher levels of education (which can drive their wage growth).

Holding education constant and age constant, however, the wage gap is still 20.2%.

2.4

In [14]:
```
cps_fem = cps.loc[cps['female'] == 1]
cps_male = cps.loc[cps['female'] == 0]
```

In [15]:
```
reg_w_a_fem = pf.feols("lnw~age", data=cps_fem,vcov="HC1")
reg_w_a_male = pf.feols("lnw~age", data=cps_male,vcov="HC1")
reg_w_faaf = pf.feols("lnw~female+age+age*female", data=cps,vcov="HC1")
```

In [16]:
```
pf.etable([reg_w_a_fem, reg_w_a_male, reg_w_faaf],
          head_order="h",
          model_heads=["ln wage","ln wage", "ln wage"],
          labels={"Intercept": "Constant"}
)
```

Out[16]:

	ln wage		
	(1)	(2)	(3)
age	0.002 (0.002)	0.010*** (0.001)	0.010*** (0.001)
female			0.120 (0.131)
age:female			-0.008** (0.003)
Constant	2.744*** (0.114)	2.624*** (0.065)	2.624*** (0.065)
Observations	319	848	1167
S.E. type	hetero	hetero	hetero
R^2	0.002	0.042	0.064
Adj. R^2	-0.001	0.041	0.062

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

1)

As age increases, the average wage of women increases by 0.2%. However, the coefficient is not significant.

2)

As age increases, the average wage of men increases by 1%

3)

As age increases, the average wage of men increases by 1% (because we're holding gender constant).

The female coefficient tells us that on average, females have 12% higher wages than male, holding age constant. However, the results are not significant.

As age increases, the average change in wage will be 0.8 percentage points less for women than for men.

The models are connected thus: The age coefficient in model 3 is the same as the age coefficient in model 2, because in model 3, we're holding gender constant (meaning, we're calculating the coefficient for male workers). The age:female interaction term (-0.008) is the same as taking the age coefficient for the female sample (0.002) and subtracting the age coefficient for the male sample (0.01).

3.1

In [52]:
```
reg_w_f = pf.feols("lnw~female", data=cps,vcov="HC1")
reg_w_fae = pf.feols("lnw~female+age+educ", data=cps,vcov="HC1")
reg_w_faemo = pf.feols("lnw~female+age+educ+married+ownchild", data=cps,vcov="HC1")
```

In [53]:
```
pf.etable([reg_w_f, reg_w_fae, reg_w_faemo],
          head_order="h",
          model_heads=["ln wage","ln wage", "ln wage"],
          labels={"Intercept": "Constant"}
)
```

Out[53]:

	ln wage		
	(1)	(2)	(3)
female	-0.213***	-0.202***	-0.189***
	(0.032)	(0.028)	(0.028)
age		0.007***	0.007***
		(0.001)	(0.001)
educ[T.hs_grad]		-0.454***	-0.446***
		(0.042)	(0.043)
educ[T.less_than_hs]		-0.791***	-0.794***
		(0.055)	(0.055)
educ[T.post_grad]		0.333***	0.316***
		(0.065)	(0.066)
educ[T.some_college_associate]		-0.300***	-0.294***
		(0.043)	(0.043)
married			0.075**
			(0.028)
ownchild			0.019
			(0.012)
Constant	3.048***	3.016***	2.956***
	(0.018)	(0.059)	(0.063)
Observations	1167	1167	1167
S.E. type	hetero	hetero	hetero
R^2	0.033	0.309	0.317
Adj. R^2	0.032	0.306	0.312

Significance levels: * p < 0.05, ** p < 0.01, *** p < 0.001. Format of coefficient cell: Coefficient (Std. Error)

3.2

As can be seen from the output above, gender, age, education levels and possibly marriage status are important in determining the wage gap between men and women in the automotive sector. Marriage is especially interesting as a variable, as it still plays a role even when controlling for age. On average, married people earn more than unmarried people, holding other variables (including age) constant. Having kids, on the other hand, does not play a significant role in determining a wage gap.

Before controlling for all these variables, the gender wage gap was 21.3%. After controlling, it is 18.9%.