KorlackiAssignment1

January 29, 2026

[194]:

```
# Imports
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
import statsmodels.api as sm
from typing import List
from stargazer.stargazer import Stargazer
from IPython.display import display, Latex
pd.options.mode.chained_assignment = None
```

[195]:

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

```python
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)
```

```
C:\Users\2025\AppData\Local\Temp\ipykernel_7652\2317969728.py:1: DtypeWarning:
Columns (16) have mixed types. Specify dtype option on import or set
low_memory=False.
  cps= pd.read_csv("https://osf.io/download/4ay9x/")
```

0.1 Question 1

1.1

[196]:

```python
data = cps.query('ind02=="Postal Service (491)"')
num = len(data)
```

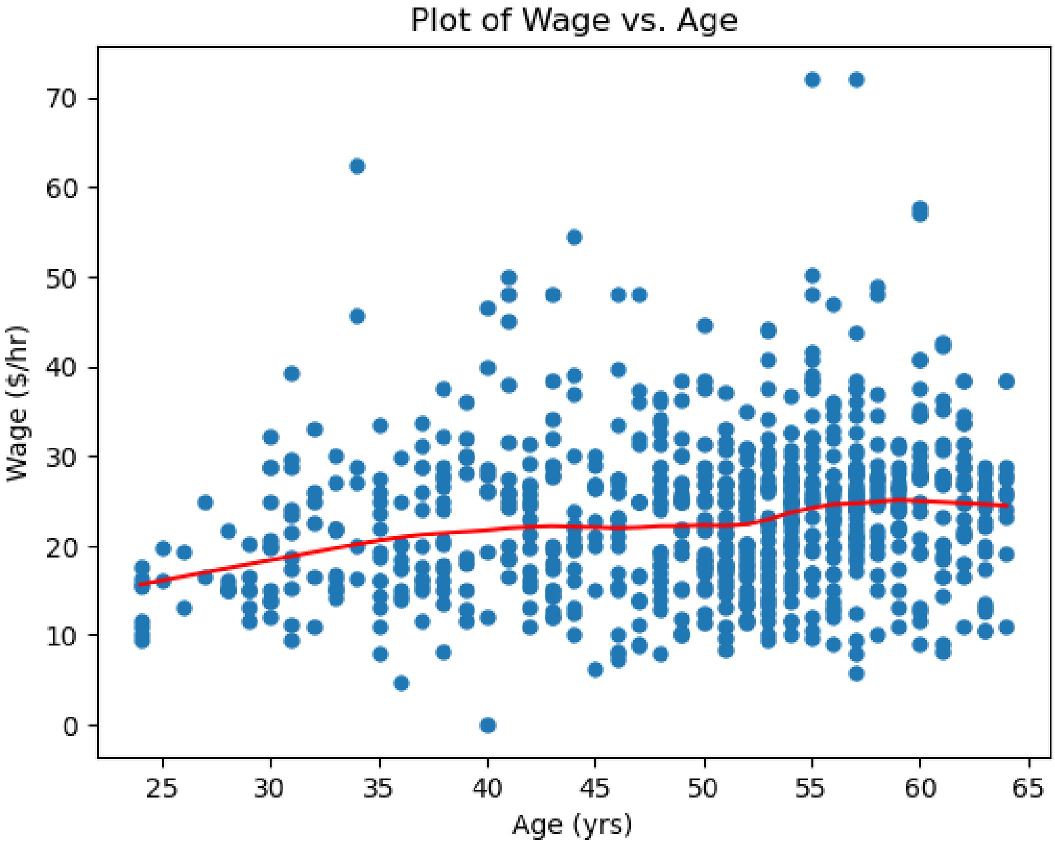
1.2

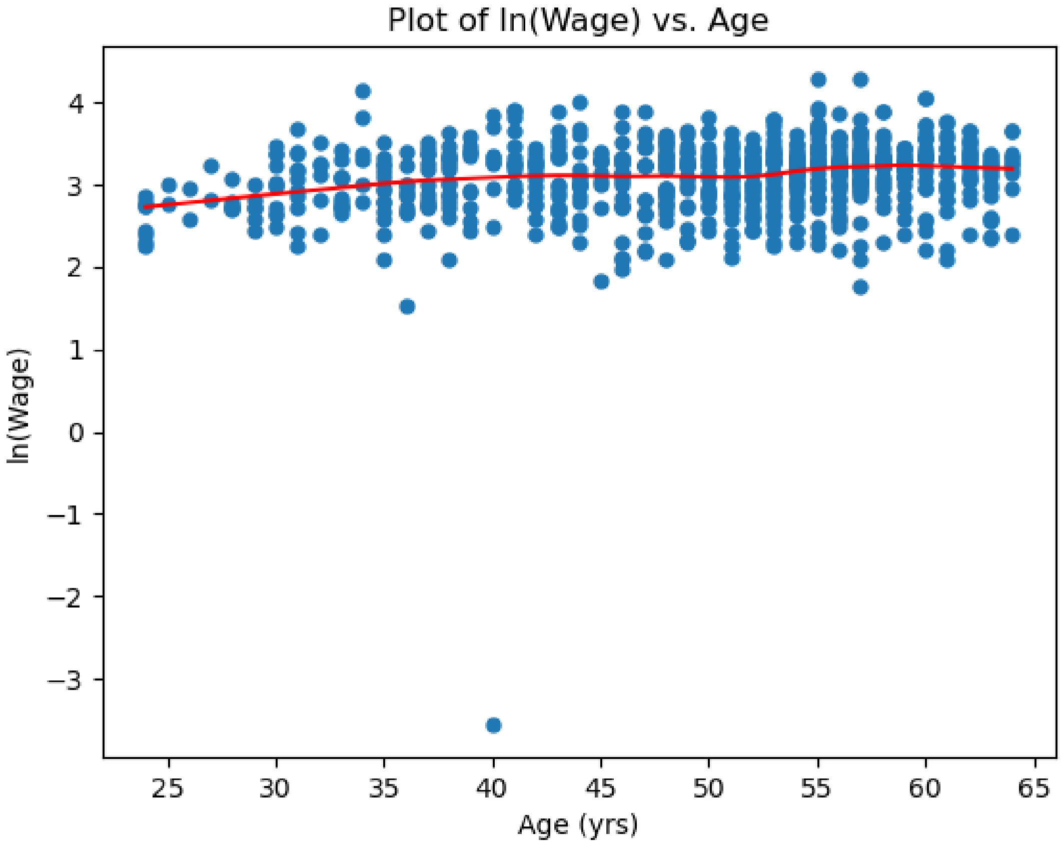
[197]:

```python
# Fig of Wage vs Age
plt.figure()
plt.scatter(data['age'],data['w'],s=25)
lowessfit1 = lowess(data['w'],data['age'],frac=1/3.0)
plt.plot(lowessfit1[:,0],lowessfit1[:,1],'r-')
plt.xlabel("Age (yrs)")
plt.ylabel("Wage ($/hr)")
plt.title("Plot of Wage vs. Age")
# Fig of ln(Wage) vs Age
plt.figure()
plt.scatter(data['age'],data['lnw'],s=25)
lowessfit2 = lowess(data['lnw'],data['age'],frac=1/3.0)
plt.plot(lowessfit2[:,0],lowessfit2[:,1],'r-')
plt.xlabel("Age (yrs)")
plt.ylabel("ln(Wage)")
plt.title("Plot of ln(Wage) vs. Age")
```

[197]:

```
Text(0.5, 1.0, 'Plot of ln(Wage) vs. Age')
```


Plot of Wage vs. Age
Wage ($/hr)
0
10
20
30
40
50
60
70
Age (yrs)
25
30
35
40
45
50
55
60
65



Both plots show that in the postal service, wages gradually increase in age. However, the lowess curve in the first plot suggests that wages tend to plateau between 40 and 50 years of age, before increasing again after that. Interestingly, while the data is generally clustered around the lowess curve, the ln(wage) vs age plot shows a clear outlier in the data.

1.3

```
[198]: # Add age**2 and age**3
data.loc[:,'age2'] = np.square(data['age'])
data.loc[:,'age3'] = np.power(data['age'],3)
```

1.4

```
[199]: data.loc[:,'preHS'] = (data['grade92'] <= 38).astype(int)
data.loc[:,'HS'] = (data['grade92'] == 39).astype(int)
data.loc[:,'Coll'] = ((data['grade92'] >= 40) & (data['grade92'] <= 42)).
 ↪astype(int)
data.loc[:,'BS'] = (data['grade92'] == 43).astype(int)
data.loc[:,'Adv'] = (data['grade92'] >= 44).astype(int)

cond = [
    (data['grade92'] <= 38),
```

```python
    (data['grade92'] == 39),
    (data['grade92'] >= 40) & (data['grade92'] <= 42),
    (data['grade92'] == 43),
    (data['grade92'] >= 44)
]
vals = ['preHS', 'HS', 'Coll', 'BS','Adv']
data.loc[:,'edu'] = np.select(cond,vals,default='')

counts = data['edu'].value_counts(normalize=True)
print(counts.sort_index())
```

```
edu
Adv      0.013495
BS       0.164642
Coll     0.431849
HS       0.372470
preHS    0.017544
Name: proportion, dtype: float64
```

The majority (80.42%) of the workers in the postal service had either high completed high school or some college, with another 16.46% having a Bachelor's degree. However, very few postal workers had advanced degrees or hadn't completed high school. The general lack of advanced degrees in postal workers generally agrees with an assumption of not needing advanced skills or education in this field. However, seeing that the overwhelming majority had at least completed high school can likely be explained by that generally being a requirement for government jobs of this sort.

0.2 Question 2

2.1

[200]:

```python
x1 = data['female']
x1 = sm.add_constant(x1)
y1 = data['lnw']
model1 = sm.OLS(y1,x1)
res1 = model1.fit()

x2 = data.loc[:,['female','age']]
x2 = sm.add_constant(x2)
y2 = data['lnw']
model2 = sm.OLS(y2,x2)
res2 = model2.fit()

x3 = data.loc[:,['female','age','age2']]
x3 = sm.add_constant(x3)
y3 = data['lnw']
model3 = sm.OLS(y3,x3)
res3 = model3.fit()

x4 = data.loc[:,['female','age','age2','age3']]
```

```
x4 = sm.add_constant(x4)
y4 = data['lnw']
model4 = sm.OLS(y4,x4)
res4 = model4.fit()

stargazer = Stargazer([res1, res2, res3, res4])
stargazer.covariate_order(['female','age','age2','age3','const'])
stargazer.custom_columns(['Model 1','Model 2','Model 3','Model 4'])
stargazer.dependent_variable_name('ln(wage)')
display(Latex(stargazer.render_latex()))
```

	Dependent variable: ln(wage)			
	Model 1	Model 2	Model 3	Model 4
	(1)	(2)	(3)	(4)
female	-0.148***	-0.130***	-0.131***	-0.130***
	(0.034)	(0.034)	(0.034)	(0.034)
age		0.008***	0.023	0.084
		(0.002)	(0.015)	(0.092)
age2			-0.000	-0.002
			(0.000)	(0.002)
age3				0.000
				(0.000)
const	3.146***	2.731***	2.407***	1.560
	(0.023)	(0.090)	(0.345)	(1.319)
Observations	741	741	741	741
R^2	0.025	0.054	0.055	0.056
Adjusted R^2	0.024	0.052	0.052	0.051
Residual Std. Error	0.457 (df=739)	0.451 (df=738)	0.451 (df=737)	0.451 (df=73
F Statistic	19.095*** (df=1; 739)	21.153*** (df=2; 738)	14.416*** (df=3; 737)	10.915*** (df=4;
Note:				*p<0.1; **p<0.05; ***p<

a. For model 2, a coefficient of -0.130 for the female variable suggests that, on average, female postal workers make on average about 13% less than male postal workers. A cofficient of 0.008 for the age variable means that, on average, a postal worker makes 0.8% more for every year they are older.

b. Out of these 4 models, model 2 is likely the most appropriate. Adding the age^2 and age^3 variables does not appear to substantially improve the predictive capacity of the model, with the R^2 values being nearly the same. Additionally, in model 2, the age coefficient is significant at 99% confidence while for models 3 and 4, all of the age coefficients are not significant.

2.2

[201]:
```
x5 = data.loc[:,['female','preHS','Coll','BS','Adv']]
x5 = sm.add_constant(x5)
y5 = data['lnw']
```

```
model5 = sm.OLS(y5,x5)
res5 = model5.fit()

stargazer2 = Stargazer([res5])
stargazer2.covariate_order(['female','preHS','Coll','BS','Adv','const'])
stargazer2.custom_columns(['Model 5'])
stargazer2.rename_covariates({'preHS' : 'Not Finished High School', 'Coll':
 ↪'Some College','BS':'Bachelors','Adv':'Advanced Degrees', 'female':
 ↪'Female','const':'Const'})
stargazer2.dependent_variable_name('ln(wage)')
display(Latex(stargazer2.render_latex()))
```

	Dependent variable: ln(wage)
	Model 5
	(1)
Female	-0.137***
	(0.034)
Not Finished High School	-0.436***
	(0.128)
Some College	0.109***
	(0.037)
Bachelors	0.105**
	(0.049)
Advanced Degrees	0.310**
	(0.145)
Const	3.081***
	(0.032)
Observations	741
R^2	0.062
Adjusted R^2	0.056
Residual Std. Error	0.450 (df=735)
F Statistic	9.706*** (df=5; 735)
Note:	*p<0.1; **p<0.05; ***p<0.01

In the model that examines the effect of gender and education on wages in the postal service, female postal workers, on average, make about 13.7% less then men (similar to the female and age regressions performed in part 2.1). In contrast, education is correlated with having a positive effect on the wages postal workers make. The regression uses the 'completed high school' category as the baseline for education, and it shows that not finishing high school corresponds to, on average, having 43.6% lower wages. Those postal workers that completed some college or have a Bachelor's make about 10.9% and 10.5%, respectively, compared to their high school peers, and those with advanced degrees make an average of about 31.0% more.

2.3

```
[202]: x6 = data.loc[:,['female','age','preHS','Coll','BS','Adv']]
       x6 = sm.add_constant(x6)
       y6 = data['lnw']
       model6 = sm.OLS(y6,x6)
       res6 = model6.fit()

       stargazer3 = Stargazer([res1, res2, res6])
       stargazer3.covariate_order(['female','age','preHS','Coll','BS','Adv','const'])
       stargazer3.custom_columns(['Model 1','Model 2','Model 6'])
       stargazer3.rename_covariates({'age': 'Age','preHS' : 'Not Finished High␣
        ↪School', 'Coll':'Some College','BS':'Bachelors','Adv':'Advanced Degrees',␣
        ↪'female':'Female','const':'Const'})
       stargazer3.dependent_variable_name('ln(wage)')
       display(Latex(stargazer3.render_latex()))
```

	Dependent variable: ln(wage)		
	Model 1	Model 2	Model 6
	(1)	(2)	(3)
Female	-0.148***	-0.130***	-0.119***
	(0.034)	(0.034)	(0.033)
Age		0.008***	0.008***
		(0.002)	(0.002)
Not Finished High School			-0.441***
			(0.126)
Some College			0.114***
			(0.036)
Bachelors			0.107**
			(0.049)
Advanced Degrees			0.269*
			(0.143)
Const	3.146***	2.731***	2.660***
	(0.023)	(0.090)	(0.092)
Observations	741	741	741
R^2	0.025	0.054	0.091
Adjusted R^2	0.024	0.052	0.084
Residual Std. Error	0.457 (df=739)	0.451 (df=738)	0.443 (df=734)
F Statistic	19.095*** (df=1; 739)	21.153*** (df=2; 738)	12.283*** (df=6; 734)
Note:			*p<0.1; **p<0.05; ***p<0.01

As additional variables like education and age are included in the regression, the magnitude of the coefficient for the female variable tends to decrease. This suggests that for postal workers, there may be additional factors other than gender that influence the wage gap, with the gap going from 14.8% in model 1 to 11.9% in model 6, and that controlling for these other factors can help better estimate the actual gender wage gap. However, even in model 6, the coefficient for the female variable is still significant at 99% confidence, suggesting that gender does still play a factor in explaining the wage gap.

2.4

[203]:
```
dataf = data.query('female==1')
datam = data.query('female==0')
data['ageXfemale'] = data['female']*data['age']

x7 = dataf.loc[:,['age']]
x7 = sm.add_constant(x7)
y7 = dataf['lnw']
model7 = sm.OLS(y7,x7)
res7 = model7.fit()

x8 = datam.loc[:,['age']]
x8 = sm.add_constant(x8)
y8 = datam['lnw']
model8 = sm.OLS(y8,x8)
res8 = model8.fit()

x9 = data.loc[:,['female','age','ageXfemale']]
x9 = sm.add_constant(x9)
y9 = data['lnw']
model9 = sm.OLS(y9,x9)
res9 = model9.fit()

stargazer4 = Stargazer([res7, res8, res9])
stargazer4.covariate_order(['female','age','ageXfemale','const'])
stargazer4.custom_columns(['Model 7','Model 8','Model 9'])
stargazer4.rename_covariates({'age': 'Age','ageXfemale':'Age*Female', 'female':
  ↪'Female','const':'Const'})
stargazer4.dependent_variable_name('ln(wage)')
display(Latex(stargazer4.render_latex()))
```

Model 7 shows that for female postal workers, every year they are older corresponds to a wage that is 0.8% higher, on average, while model 8 shows that the wage increase for men with respect to age is 0.9% on average. This is confirmed by the interaction term in Model 9 having a coefficient of -0.001, which is a correction on the age coefficient for women. Model 9 also shows that women make an average of 8.4% less than men.

0.3 Question 3

3.1

[204]:
```
data['carrier'] = (data['occ2012']==5550).astype(int)
data['clerk'] = (data['occ2012']==5540).astype(int)
data['sorter'] = (data['occ2012']==5560).astype(int)
data['otherocc'] =  ( (data['occ2012']!=5540)& (data['occ2012']!=5550) &␣
  ↪(data['occ2012']!=5560) ).astype(int)
data['union'] = (data['unionmme'] == "Yes").astype(int)
```

	Dependent variable: ln(wage)		
	Model 7	Model 8	Model 9
	(1)	(2)	(3)
Female			-0.084
			(0.175)
Age	0.008***	0.009***	0.009***
	(0.003)	(0.002)	(0.002)
Age*Female			-0.001
			(0.003)
Const	2.626***	2.709***	2.709***
	(0.144)	(0.104)	(0.121)
Observations	330	411	741
R^2	0.021	0.042	0.054
Adjusted R^2	0.018	0.040	0.050
Residual Std. Error	0.516 (df=328)	0.390 (df=409)	0.451 (df=737)
F Statistic	6.947*** (df=1; 328)	18.111*** (df=1; 409)	14.108*** (df=3; 737)
Note:			*p<0.1; **p<0.05; ***p<0.01

```
x10 = data.loc[:,['female','age','preHS','Coll','BS','Adv','afram','asian',␣
 ↪'hisp','othernonw','nonUSborn','divorced','wirowed','nevermar','child1',␣
 ↪'child2','child3', 'child4pl','carrier','clerk','sorter','union']]
x10 = sm.add_constant(x10)
y10 = data['lnw']
model10 = sm.OLS(y10,x10)
res10 = model10.fit()

stargazer5 = Stargazer([res1, res6, res10])
stargazer5.
 ↪covariate_order(['female','age','preHS','Coll','BS','Adv','afram','asian',␣
 ↪'hisp','othernonw','nonUSborn','divorced','wirowed','nevermar','child1','child2',␣
 ↪'child3', 'child4pl','carrier','clerk','sorter','union','const'])
stargazer5.custom_columns(['Model 1','Model 6','Model 10'])
stargazer5.rename_covariates({'female' : 'Female','age' : 'Age','preHS' : 'Not␣
 ↪Finished High School','Coll' : "Some College",'BS': 'Bachelors','Adv':
 ↪'Advanced Degrees', 'afram':'African American','asian':'Asian','hisp':
 ↪'Hispanic','othernonw':'Other Non-White','nonUSborn':'Non-US␣
 ↪Born','divorced':'Divorced','wirowed':'Widowed', 'nevermar':'Never␣
 ↪Married','child1':'1 Child','child2' : '2 Children','child3': '3␣
 ↪Children','child4pl': '4+ Children','carrier' : 'Postal Carrier','clerk' :
 ↪'Postal Clerk','sorter' : 'Postal Sorter','union' : 'Union Member','const':␣
 ↪'Const'})
```

```python
stargazer5.dependent_variable_name('ln(wage)')
display(Latex(stargazer5.render_latex()))
```

3.2 The baseline or reference for Model 10 is a white, male, non-union, married, postal woker born in the US with a high school education who is not a carrier, clerk, or sorter and has no children. As these additional variables are added, the part of the wage gap that is attributed solely to gender continues to shrink, with model 10 showing only about a 10.9% lower wage, on average, for women compared to men (vice 14.8% in model 1). This shows how some of these other variables can help account for the difference in wages when they are appropraitely controlled for. As noted previously, education plays a substantial factor, with a more than 40% decrease in earnings for those that haven't completed high school to a 27.8% average increase for those with advanced degrees, with respect to those that only completed high school. While the magnitudes of each education coefficients are slightly different compared to Model 6, the general trends and statistical significance remains constant. Age also continues to play a major factor, with each additional age corresponding to, on average, a 0.7% higher wage, which is statistically significant at 99%.

That leads to some of the other variables that were considered. As far as race is concerned, the only statistically significant coefficient was for Hispanic postal workers, but even then the regression coefficient is only significant at 90% confidence. These results suggest that, at least based on this model and data, race does not play a major factor in postal worker's wages. The same holds for citizenship status. Looking at marital status and number of children, only one of the coefficients is of notable interest. Never married postal workers have a, on average 10% lower wage than those that are married. This could perhaps be that many US government pay scales include adjustments for those workers with dependents such as spouses and children. It may warrant further investigation how closely marital status is linked to the number of children, but logically it follows that once a person is married, it is more likely they may have further dependents. As for the type of work, looking at the 3 most common types of occupations - carriers, sorters, and clerks - all of these are negatively correlated with income compared to other occupations in this sector, although these results are not or barely statistically significant. This suggests that other roles - such as management and administration - might make more than your rank and file of the postal service. Finally, it is unsurprising that those postal workers that are union members, on average, make 12.6% more than their non-union peers.

	Dependent variable: ln(wage)		
	Model 1	Model 6	Model 10
	(1)	(2)	(3)
Female	-0.148***	-0.119***	-0.103***
	(0.034)	(0.033)	(0.034)
Age		0.008***	0.007***
		(0.002)	(0.002)
Not Finished High School		-0.441***	-0.434***
		(0.126)	(0.126)
Some College		0.114***	0.098***
		(0.036)	(0.036)
Bachelors		0.107**	0.106**
		(0.049)	(0.048)
Advanced Degrees		0.269*	0.278*
		(0.143)	(0.143)
African American			-0.044
			(0.044)
Asian			-0.065
			(0.090)
Hispanic			-0.134*
			(0.068)
Other Non-White			0.047
			(0.102)
Non-US Born			0.001
			(0.074)
Divorced			0.000***
			(0.000)
Widowed			0.015
			(0.136)
Never Married			-0.100**
			(0.048)
1 Child			-0.027
			(0.129)
2 Children			0.234
			(0.171)
3 Children			0.088
			(0.065)
4+ Children			0.033
			(0.047)
Postal Carrier			-0.059
			(0.039)
Postal Clerk			-0.082*
			(0.050)
Postal Sorter			-0.040
			(0.060)
Union Member			0.126***
			(0.035)
Const	3.146***	2.660***	2.722***
	(0.023)	(0.092)	(0.123)
Observations	741	741	741
R^2	0.025	0.091	0.129