Authors: Dinmukhamed Alchinbayev, Qinlin Wang, Sizhe Sun

Table 1. Summary statistics: 2008 Current Population Survey (CPS)

	count	mean	sd	skewness	min	p5	p25	p50	p75	p95	max
earnings per hour ($)	4733	10.19	6.21	2.11	1.05	3.68	5.89	8.53	12.75	22.69	78.71
years of education	4733	13.30	2.36	-0.09	1.00	10.00	12.00	13.00	16.00	18.00	18.00
years	4733	38.33	11.30	0.17	18.00	21.00	29.00	38.00	47.00	58.00	64.00
years of work experience	4733	19.04	11.40	0.27	0.00	2.00	10.00	19.00	27.00	39.00	52.00
=1 if female	4733	0.49	0.50	0.06	0.00	0.00	0.00	0.00	1.00	1.00	1.00
=1 if black	4733	0.10	0.30	2.69	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if white	4733	0.90	0.30	-2.69	0.00	0.00	1.00	1.00	1.00	1.00	1.00
=1 if married	4733	0.60	0.49	-0.43	0.00	0.00	0.00	1.00	1.00	1.00	1.00
=1 if union member	4733	0.16	0.37	1.82	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if northeast region of U.S.	4733	0.22	0.42	1.34	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if midwest region of U.S.	4733	0.24	0.43	1.19	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if south region of U.S.	4733	0.31	0.46	0.82	0.00	0.00	0.00	0.00	1.00	1.00	1.00
=1 if west region of U.S.	4733	0.22	0.42	1.33	0.00	0.00	0.00	0.00	0.00	1.00	1.00
=1 if full time worker (as opposed to part-time worker)	4733	0.88	0.32	-2.36	0.00	0.00	1.00	1.00	1.00	1.00	1.00
=1 if lives in metropolitan area	4733	0.79	0.40	-1.45	0.00	0.00	1.00	1.00	1.00	1.00	1.00
Observations	4733										

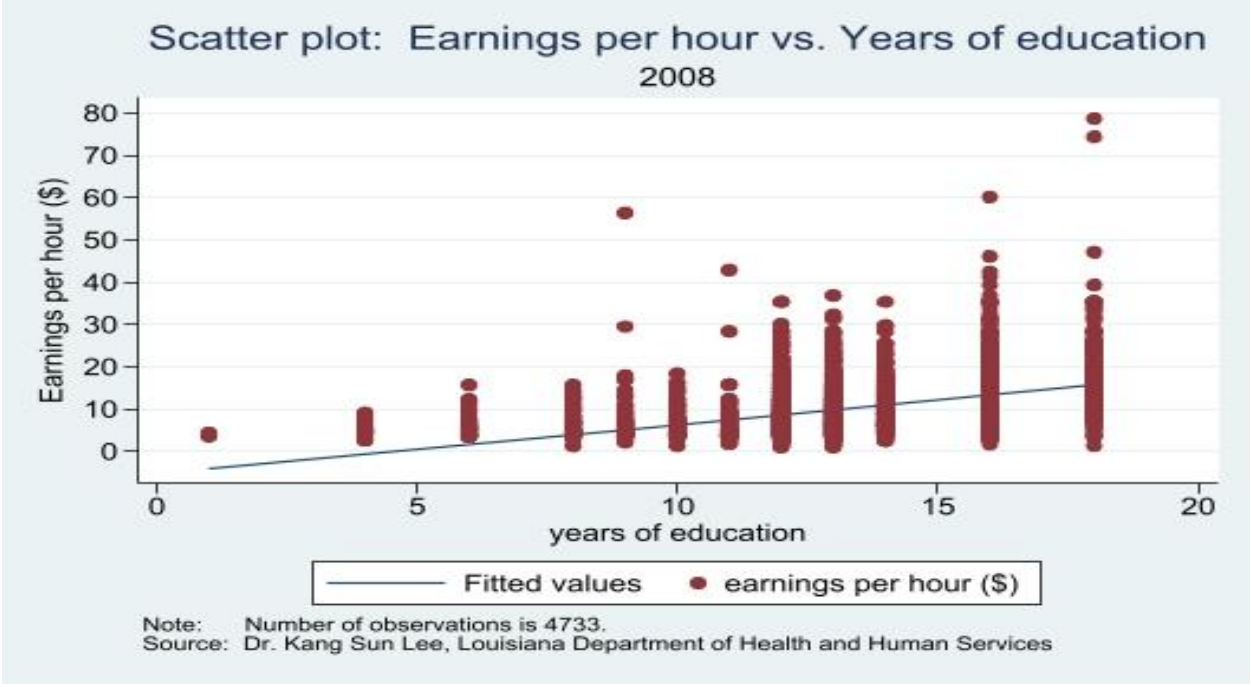
Note: .
Source: Dr. Kang Sun Lee, Louisiana Department of Health and Human Services.

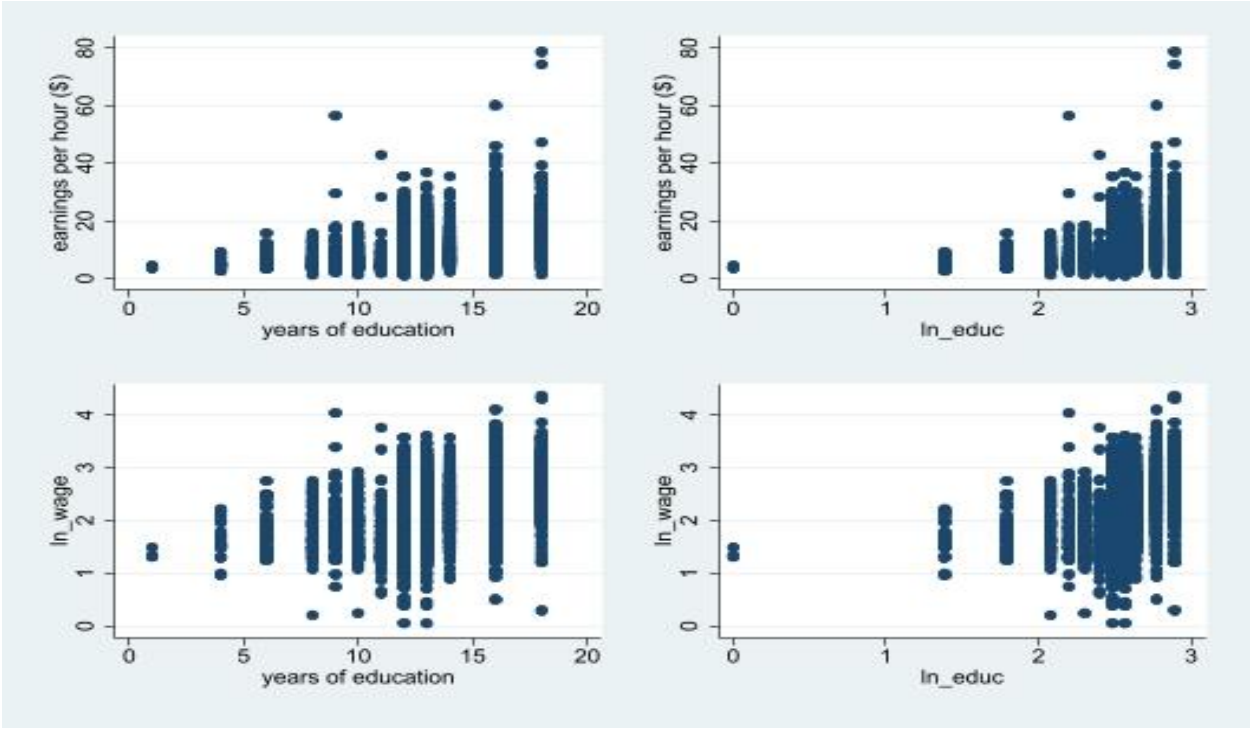
Based on the database which contains data on measures of the hourly wage rate, years of education, age of a person, and other 12 variables from the 2008 Current Population Survey (CPS), we summarized the number of observations, mean, standard deviation, skewness, minimum value, P5, P25, P50, P75, and maximum value for those 15 variables by using STATA.

```
. reg wage educ, r

Linear regression                               Number of obs     =      4,733
                                                F(1, 4731)        =     722.86
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1924
                                                Root MSE          =     5.5846

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   1.156924   .0430306    26.89   0.000     1.072564    1.241284
       _cons |  -5.202605   .5498755    -9.46   0.000    -6.280617   -4.124593
------------------------------------------------------------------------------
```

The coefficient of the educ variable indicates that every additional year of education increases the hourly earning by 1.15 $/hour. This effect is statistically significant at the 5% level.

If we use the standard deviation as a measure of significant effect, an increase of 1.15 $/hour is not big enough. The standard deviation of wage is 6.21 $/hour. It means that to increase the wage by 1 standard deviation we would need 5.4 years of education (6.21/1.15). 5.4 years of education is more than twice the standard deviation of the education variable (5.4/2.36=2.28). Even if we account for a bachelor's degree which usually requires 4 years of education, it will only increase hourly wage by 4.6 $/hour, which is less than one standard deviation. It starts being meaningful if a person gets a bachelor's degree and then studies for one more year.


Scatter plot: Earnings per hour vs. Years of education
2008
Earnings per hour ($)
0
10
20
30
40
50
60
70
80
0
5
10
15
20
years of education
Fitted values
earnings per hour ($)
Note: Number of observations is 4733.
Source: Dr. Kang Sun Lee, Louisiana Department of Health and Human Services


earnings per hour ($)
years of education
earnings per hour ($)
ln_educ
ln_wage
years of education
ln_wage
ln_educ

e)

Table 1. Regression results for person's earnings in 2008

	(1) ln_wage	(2) ln_wage	(3) ln_wage	(4) ln_wage	(5) ln_wage
ln_educ	1.104***				
	(0.0626)				
years of education		0.105***	0.114***	0.115***	0.105***
		(0.00315)	(0.00301)	(0.00294)	(0.00379)
years of work experience			0.0121***	0.0121***	0.0122***
			(0.000611)	(0.000598)	(0.000597)
=1 if female				-0.249***	-0.551***
				(0.0132)	(0.0798)
female_educ					0.0227***
					(0.00597)
Constant	-0.671***	0.770***	0.426***	0.524***	0.654***
	(0.162)	(0.0424)	(0.0424)	(0.0419)	(0.0523)
Observations	4733	4733	4733	4733	4733
R^2	0.164	0.202	0.263	0.315	0.317
Adjusted R^2	0.164	0.202	0.263	0.314	0.316
F	311.4	1112.7	862.0	735.9	545.9
rmse	0.503	0.491	0.472	0.455	0.455

Note 1: Robust standard errors are displayed in parenthesis.
Significance levels: * p<0.10; ** p<0.05; *** p<0.01
Source: Dr. Kang Sun Lee, Louisiana Department of Health and Human Services.

f) The interpretation of the coefficient in Model 1 is that with every increase of 1 percent in education we get an increase of 1.1 percent in the person's earnings. The coefficient in Model 2 indicates that every additional year of education increases hourly wages by 10.5%.

In this case, we can look at the effect of an associate's degree. If a person works a standard full-time they get around 1,920 hours of work per year (40*4*12=), assuming no holidays and leaves. If they earn 15 $/hour their annual earning would be 15*1,920=28,800$, ignoring taxes. If they get an associate's degree and increase their years of education by 2 years their hourly earnings will increase by 20%. That is 18*1,920=34,560$ annually. Thus, the effect is practically significant.

g) Comparing models (1) and (2) with each other, we find that model 2 has a larger adjusted R-squared value and a smaller root mean standard error value. Model #2's adjusted R-squared value equals 0.202 meaning that 20.2% of the variation in the hourly wage rate can be explained by the regression of the ln(wage) on the years of education. The smaller RMSE value for the regression in model #2 means that using the years of education yields more accurate predictions than the predictions made by using the percentage change in years of education in model#1. Therefore, model#2 provides a better fit for the data.

h) We can see that according to model 3, every additional year of education increases a person's hourly wage by 11.4%, while every additional year of experience increases it by 1.2%. If we compare models 2 and 3 we can observe that the coefficient of educ changed from 10.5% to 11.4%, while the

standard errors decreased. Furthermore, we generally expect the number of years of experience and years of education to be correlated, since every year a person studies they don't work, decreasing their experience by that one year. Additionally, it makes economic sense that a more experienced worker would be appreciated and, consequently, paid more than an inexperienced one. This leads us to believe that model 2 has OVB. However, given that every additional year of experience increases the person's wage only by 1.2%, meaning that having 10 years of experience leads to an increase of only 12%, whereas one year of education alone leads to an increase of 11.4%, we conclude that this change is not practically meaningful.

i) In model#4, we run the Log-linear regression of ln wage on years of education, years of work experience, and the gender of "female". The estimated coefficient of "female" means, holding the years of education and work experience constant, if the gender of the person is female, the hourly wage rate decreases by 24.9%. The estimated coefficient of "years of education" means, holding everything else constant, if the years of education increase by 1, the hourly wage rate increases by 11.5%. The estimated coefficient of "years of work experience" means, holding everything else constant, if the years of work experience increase by 1, the hourly wage rate increases by 1.21%

j) If we are interested in the causal relationship between education and earnings then we can use model 3. Since years of education generally are not correlated with gender, the gender dummy is not relevant in the study of the causal effect of education on earnings. Additionally, since the coefficient of years of education (educ) did not change in a significant way, we can assume that gender does not cause omitted variable bias.

k) We can infer the following: every additional year of education increases the wage by 10.5% for males, while females get paid less than males by 55% on average. We can also see that on average, for every additional year of education females get a 12% increase in hourly earnings, 2% more than males. It should also be noted that this difference between the return on education between males and females is statistically significant, as are other effects. Given that, we empirically observe the gender pay gap, which is reflected in the differences in pay between males and females when controlled for education and experience. We can also see that education generally increases wages, especially for women, for whom returns on education are higher than for males by 2%, highlighting the economic benefits of education.

l)

```
Linear regression                               Number of obs     =      4,733
                                                F(7, 4725)        =     318.20
                                                Prob > F          =     0.0000
                                                R-squared         =     0.3213
                                                Root MSE          =     .45325

------------------------------------------------------------------------------
             |               Robust
     ln_wage | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .1041582   .0038048    27.38   0.000      .096699    .1116173
       exper |   .0120771    .000597    20.23   0.000     .0109066    .0132476
      female |  -.5591927   .0796371    -7.02   0.000    -.7153186   -.4030668
 female_educ |   .0232517   .0059566     3.90   0.000     .0115741    .0349293
   northeast |          0  (omitted)
     midwest |  -.0503708   .0194223    -2.59   0.010    -.0884475   -.0122941
       south |  -.1016694   .0181967    -5.59   0.000    -.1373434   -.0659955
        west |  -.0524149   .0203546    -2.58   0.010    -.0923194   -.0125105
       _cons |   .7269702   .0548663    13.25   0.000     .6194067    .8345337
------------------------------------------------------------------------------

. test northeast midwest south west

 ( 1)  o.northeast = 0
 ( 2)  midwest = 0
 ( 3)  south = 0
 ( 4)  west = 0
       Constraint 1 dropped

       F(  3,  4725) =   10.62
            Prob > F =    0.0000
```

In Part(l), we extend Model 5 from Part (e) by incorporating geographic region dummies (northeast, midwest, south, and west) When we run the Log-Linear regression of *ln_wage on educ, exper, female, female_edu, northeast, midwest, south, and west*, the control variable of "northeast" was omitted. In this case, the category of "Northeast" is chosen as the baseline category. From the estimated coefficient of the midwest, south, and west, we can know that the earnings per hour (hourly wage rate) in the midwest region of the U.S., holding the years of education, years of work experience, gender of female, and "*female_educ*" (which is an "interaction" variable that you need to create by multiplying *female* by *educ*.) constant, is lower than that in the northeast region by 5.04 percent. In the southern region of the U.S., holding everything else constant, the earnings per hour are lower by 10.17 percent compared to the

northeast region. In the western region of the U.S., holding other variables constant, the earnings per hour are lower by 5.24 percent compared to the northeast region.

In Part (m), in order to test the null hypothesis that geographic regions jointly do not matter for earnings, we did the F-test based on the results of regression in Part (l). As the results of the F-test shown, there are three constraints (midwest, south, and west) and the constraint of the northeast was been dropped. Based on the table given by the question, the critical value we conduct for the F-test at the 5% significance level is 2.60 and the critical value we conduct for the F-test at the 1% significance level is 3.78. Comparing the result of F-test 10.62 with both of the critical values for two significance levels, we can reject the null hypothesis that geographic regions jointly do not matter for earnings. Because F-test equals 10.62 which is much larger than the critical value of 5% (2.60), so we have more evidence to reject the null hypothesis.