Econometrics Assignment 5

Chase Bookin & Cole Price

July 22, 2020

1.

A)
The coefficient on Midwest loses its statistical significance when moving from model 1 to model 2 because the baseline categorical dummy variable changes from South to West. In each model, Midwest is being compared to the baseline region. While Midwest is significantly different statistically from the South region, it is not significantly different statistically from West. This is also reflected by the proximity of the coefficients for Midwest and West in model 1, 0.0502 and 0.0485, respectively.

B)
Looking at model 1, we see the coefficient for Midwest is 0.0502, meaning that holding all else constant, switching the region from South to Midwest is expected to increase the GPA by 0.0502. The coefficient of 0.100 on Northeast means that holding all else constant in the regression, switching from South to Northeast is expected to result in an increase in GPA of 0.100. To find the average GPA gap between a student in the Midwest and a student in the Northeast, we find the difference between the two coefficients, and see that on average, students in the Northeast have a GPA that is higher than that of their Midwest peers by 0.0498.

C)
The new interpretation of the Northeast coefficient would be the estimated effect on GPA when the region is switched from Midwest to Northeast.

2.

A)
If accidents is a concave function of miles, we would see that in the beta_3 coefficient of miles squared. In this case, the beta-3 coefficient would be negative.

B)
To strictly measure the impact of an additional mile driven on accident risk, we would set the change in expected accidents equal to the following: beta_2 + (beta_3 x miles)ˆ2 - (beta_3 x (miles - 1))ˆ2. In other words, we take the miles coefficient and add the difference of the miles squared coefficient multiplied by number of miles and the miles squared coefficient multiplied by one less than the number of miles. This is necessary to capture the varying impact of number of miles on accident risk given that the function is not linear with respect to miles and therefore does not have a constant slope.

C)
In order to see at what level of miles driven the function reaches its peak, we need to take the derivative of the regression model with respect to miles. The derivative with respect to miles is equal to the following: beta_2 + 2(beta_3)(miles). Then we set this equal to zero and solve for miles, yielding the peak accident level with respect to miles of negative beta_2 divided by the quantity 2 times beta_3, or -beta_2 / (2 x beta_3).

D)
If we re-ran this regression using log(accidents) as the Y variable, we would be using a log-linear regression model. Therefore, the interpretation of the coeffcient of 0.0078 on alcohol would be that as the driver's total alcohol consumption over the past five years increases by one unit, it is expected that the number of accidents will increase by 0.78%.

3)

B)

Table 1: Data summary

Name	Piped data
Number of rows	4733
Number of columns	15
Column type frequency:	
numeric	15
Group variables	None

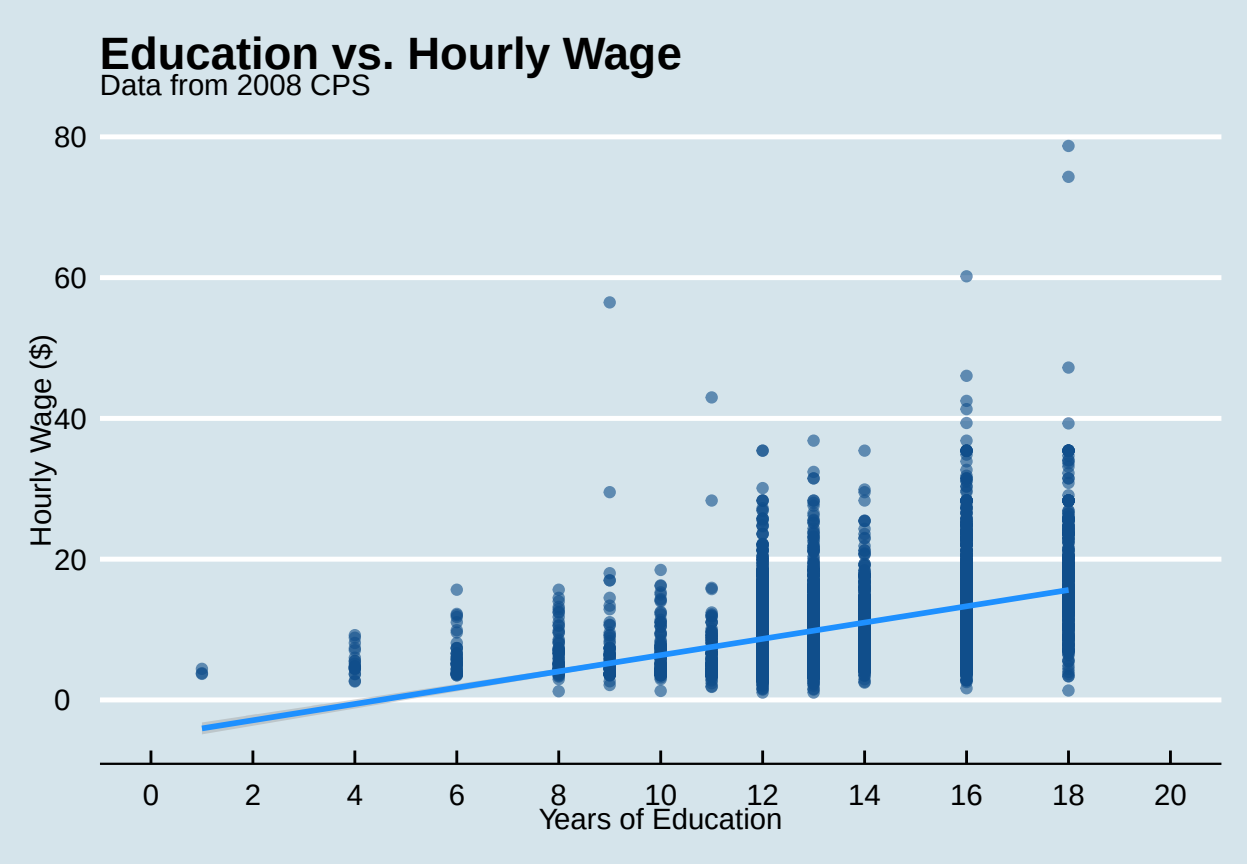
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
Wage	0	1	10.19	6.21	1.05	5.89	8.53	12.75	78.71
Education	0	1	13.30	2.36	1.00	12.00	13.00	16.00	18.00
Age	0	1	38.33	11.30	18.00	29.00	38.00	47.00	64.00
Experience	0	1	19.04	11.40	0.00	10.00	19.00	27.00	52.00
Female	0	1	0.49	0.50	0.00	0.00	0.00	1.00	1.00
Black	0	1	0.10	0.30	0.00	0.00	0.00	0.00	1.00
White	0	1	0.90	0.30	0.00	1.00	1.00	1.00	1.00
Married	0	1	0.60	0.49	0.00	0.00	1.00	1.00	1.00
Union	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00
Northeast	0	1	0.22	0.42	0.00	0.00	0.00	0.00	1.00
Midwest	0	1	0.24	0.43	0.00	0.00	0.00	0.00	1.00
South	0	1	0.31	0.46	0.00	0.00	0.00	1.00	1.00
West	0	1	0.22	0.42	0.00	0.00	0.00	0.00	1.00
Full Time	0	1	0.88	0.32	0.00	1.00	1.00	1.00	1.00
Metropolitan	0	1	0.79	0.40	0.00	1.00	1.00	1.00	1.00

Source: 2008 Current Population Survey

From this summary table of the CPS data, we see that the mean wage is 10.2 dollars per hour with a standard deviation of 6.21 dollars. The median is 8.53 dollars, and the mean is likely pulled to the right of the median due to large salaries including the maximum hourly wage of 78.7 dollars. The average years of education in the data is 13.3 with a standard deviation of 2.36. The average experience is 19 with a fairly wide spread of 11.4 years. 48.5 percent of the observations are from females, 9.87% are black, and 90.1% are white. The most common region is the South, with 31% of the data, followed by Midwest, then West and Northeast.

C)



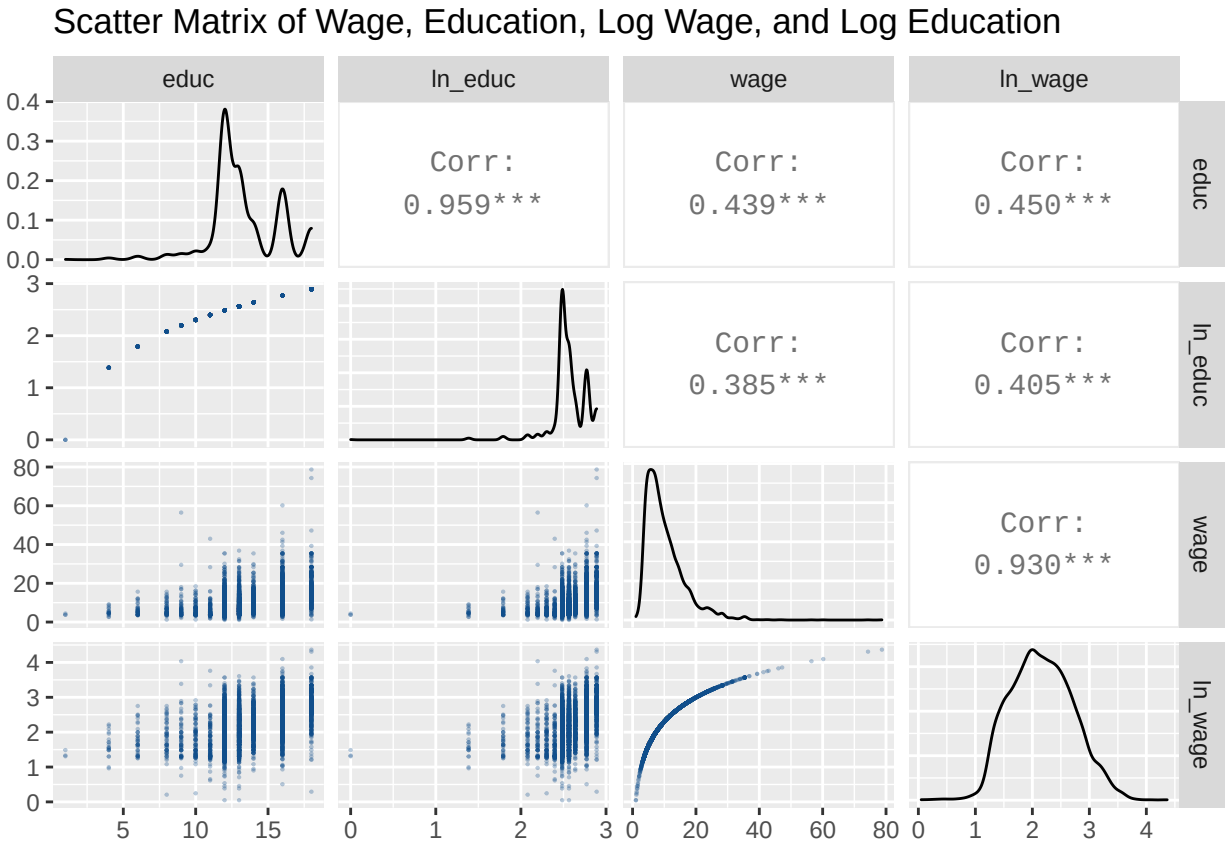

```
null <- 0
se <- 0.04303063
estimate <- 1.156924

t <- (estimate - null) / se
# Critical t-value at 5% significance level is 1.96.
```

When we regress wage on education using the CPS dataset, we find the coefficient on education is approximately 1.16. This means that for an additional year of education, we expect the hourly wage to increase by 1.16 dollars. The robust standard error of the education term is 0.043 and the intercept is -5.20.

Education is statistically significant at the 5% level. We find that the t-value of the education coefficient is approximately 26.89, much larger than the critical value of 1.96. Therefore, we reject the null hypothesis that the education coefficient is equal to zero.

The estimated value of the education coefficient is both statistically significant and practically significant. Each additional year of education is expected to increase hourly wages by 1.16 dollars, which adds up quickly over time, especially given the mean hourly wage of 10.2 dollars.

D)



Data from 2008 CPS

E)

F)
In model 1, the coefficient estimate on ln_educ means that for an increase of 1 percent in education, we expect an increase of 1.1% in hourly wage, as this is a log-log regression model. In the log-linear model 2, the coefficient estimate on educ means that for an increase of 1 year in education, we expect about a 10.5% increase in hourly wage. Both ln_educ and educ are significant at the 1% level in their respective models. The educ coefficient in model 2 is practically significant as well, as a 10.5% increase in hourly salary is a strong increase from one additional year of education. Similarly, the ln_educ coefficient in model 1 is practically significant. Considering an example increasing from 12 to 13 years of education, this 8.33% increase in education would be expected to result in a 9.2% increase in salary using the ln_educ coefficient in model 1. The expected percentage payoff of an additonal year of education is even higher for individuals with lower initial years of education.

G) Comparing model 1 to model 2, it appears that model 2 - which regresses ln_wage on education - is a better fit for the data. The adjusted R-squared value of model 2 is larger than that of model 1, with values of 0.202 and 0.164, respectively. Additionally, the residual standard error of model 2 is 0.491, lower than the 0.503 residual standard error of model 1. Finally, the F statistic of model 2 is around 1,200 and much larger than the F statistic of 927 in model 1, confirming that model 2 is a better fit for the data than model 1.

H)
Model 3 regresses ln_wage on education and experience. The coefficient on education means that for an increase of one year in education, we expect an 11.4% increase in hourly wage. This is greater than the coefficient estimate on on education in model 2, which does not include experience in the model. The coefficient on experience means that for an additional year of experience, we expect an increase of 1.2% in hourly wage. We indeed find that there is omitted variable bias. From model 3, we see that experience has

a statistically significant effect on ln_wage, and education and experience have a correlation of -0.148 in the data. Because experience is not controlled for in model 2, we see that the explanatory variable of education is correlated with the error term in the model, an indicator of OVB. Specifically, we find that the coefficient on education is biased downward in model 2. Intuitively, this stems from the negative correlation between education and experience and the positive effect of experience on wage. Without controlling for experience, as education increases and experience tends to decrease, therefore lowering the wage, this negative effect of less experience gets lumped with the effect of education. This means that in a Monte Carlo setting, the mean of the coefficient estimate on education is less than the true coefficient on education.

I)

In model 4, ln_wage is regressed on education, experience, and female. The coefficient estimate on education is 0.115, meaning that for an increase of one year in education, we expect an 11.5% increase in hourly wage, slightly greater than the coefficient estimate of 0.114 in model 3. The coefficient estimate on experience is identical to that of model 3 at 0.012, meaning that an additional one year of experience is expected to yield an increase of 1.2% in hourly wage. The new explanatory variable, female, has a coefficient estimate of -0.249, meaning that switching from a male to female subject, holding all else constant, is expected to decrease hourly wage by 24.9%.

J)

The coefficient estimate on education are nearly identical in model 3 and model 4, likely meaning that education and female are not highly correlated. We see that this is in fact the case, as the correlation between education and female in the data is 0.03. This extremely low correlation means the effect of gender when strictly considering the impact of education on hourly wage is nearly random. Because of this, if we were only interested in determining the causal effect of education on hourly wage, model 3 would suffice, as gender is largely uncorrelated with education.

K)