```
-----------------------------------------------------------------------------------------
      name:  PK
       log:  
/Users/priyakoirala/Desktop/school/econometrics/projects/project2/koirala_project2.lo
g
  log type:  text
 opened on:  13 Mar 2023, 23:49:13

. 
. /*=================================================================================
> The purpose of this assignment is to show how education affects wages. 
> The data set HTV2.dta is an observational data set on a random 
> sample of adults in the U.S. in 1991. Open the HTV2.dta data set. 
> =================================================================================*/
. 
. use "/Users/priyakoirala/Desktop/school/econometrics/projects/project2/HTV2.dta"
```

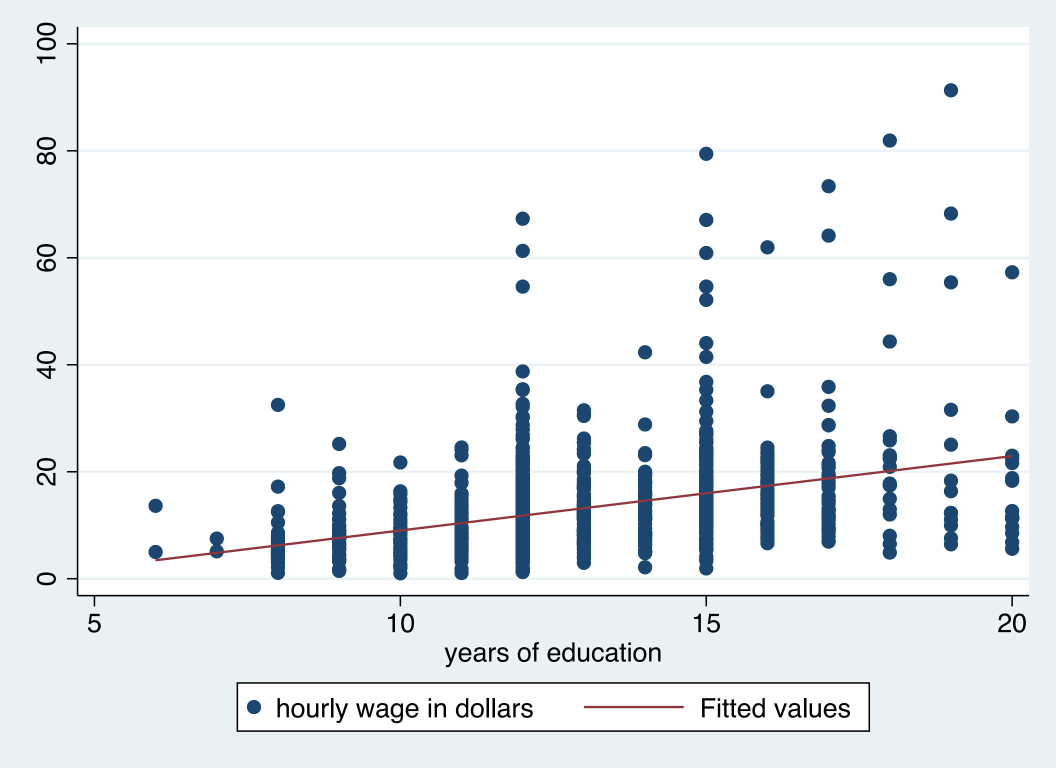
```
.
. /*==========================================================================
> (Q1): Summarize and describe the data set.
> (a) How many observations are in the data set?
> (b) How many variables are in the data set?
> (c) How are the wage and educ variables measured?
> ==========================================================================*/
.
. summarize

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        wage |      1,193    13.23942    9.116401   1.023529   91.30922
        educ |      1,193    13.03437    2.346208          6         20
          ne |      1,193     .210394    .4077594          0          1
          nc |      1,193    .3730092    .4838073          0          1
        west |      1,193    .1684828    .3744514          0          1
-------------+---------------------------------------------------------
       south |      1,193     .248114    .4320995          0          1
       exper |      1,193    10.72842    3.105527          1         19

.
. describe, f

Contains data from
/Users/priyakoirala/Desktop/school/econometrics/projects/project2/HTV2.dta
 Observations:         1,193
    Variables:             7                  20 Sep 2020 17:40
------------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
------------------------------------------------------------------------------------
wage            float   %9.0g                 hourly wage in dollars
educ            byte    %9.0g                 years of education
ne              byte    %9.0g                 =1 if person lives in the Northeast
nc              byte    %9.0g                 =1 if person lives in the Midwest
west            byte    %9.0g                 =1 if person lives in the West
south           byte    %9.0g                 =1 if person lives in the South
exper           byte    %9.0g                 years of work experience
------------------------------------------------------------------------------------
Sorted by:

.
. /* a. There are 1,193 observations in the data set.
>    b. There are 7 variables in the data set.
>    c. The wage variable is measured by hourly wage in dollars and educ variable
>       is measured by years of education. */
```

```
.
. /*===============================================================================
> (Q2): Estimate a bivariate regression relating education to wages.
> Assume homoskedasticity is true. You will consider whether it is actually true
> in (Q6).
> ===============================================================================*/
.
. reg wage educ

      Source |       SS           df       MS      Number of obs   =     1,193
-------------+----------------------------------   F(1, 1191)      =    174.98
       Model |  12689.9011         1  12689.9011   Prob > F        =    0.0000
    Residual |   86375.759     1,191  72.5237271   R-squared       =    0.1281
-------------+----------------------------------   Adj R-squared   =    0.1274
       Total |  99065.6602     1,192  83.1087753   Root MSE        =    8.5161

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   1.390671   .1051321    13.23   0.000     1.184407    1.596936
       _cons |  -4.887101   1.392335    -3.51   0.000    -7.618804   -2.155398
------------------------------------------------------------------------------
```

```
.
. /*===============================================================================
> (Q3): Consider the three assumptions that are necessary to achieve unbiased
> and consistent estimators. Does the model in (Q2) satisfy the first assumption?
> Why or why not?
> *==============================================================================*/
.
. /* The first assumption is that the error term has a conditional mean of zero
(conditional mean assumption). This means that no matter what value chosen for X, the
error term u must not show any systematic pattern and must have a mean of 0 (implies
unbiasedness).

The model in (Q2) does not achieve the first assumption because the error term is not
zero. This could be because there are other factors (variables) which contribute to
wage other than years of education. Such as, experience, skill level, family
background, etc. All of these factors and more could contribute to a higher or lower
wage. Since these factors are unaccounted for, there could be a bias in the results.
*/
```

```
.
. /*==============================================================================
> (Q4):  Consider the three assumptions that are necessary to achieve unbiased
> and consistent estimators. Does the model in (Q2) satisfy the second assumption?
> Why or why not?
> *=============================================================================*/
.
. /* The second assumption is that (Xi,Yi) are independently distributed (i.i.d).
This means the units of observation were selected at random from the population. And
that each random sample has the same distributional properties as the population.

This model could satisfy the second assumption because this data was obtained through
a random sampling of adults in the US in 1991. However, it is not to say the sample
has the same distributional properties as the population. */
```

```
.
. /*===========================================================================
> (Q5): Consider the three assumptions that are necessary to achieve unbiased
> and consistent estimators. Does the model in (Q2) satisfy the third assumption?
> Why or why not?
> *==========================================================================*/
.
. scatter wage educ || lfit wage educ

. graph export
"/Users/priyakoirala/Desktop/school/econometrics/projects/project2/Graph_Project2.pdf
", replace file
/Users/priyakoirala/Desktop/school/econometrics/projects/project2/Graph_Project2.pdf
saved as PDF format
```



```
. /* The third assumption is the no large outliers assumption. Though there are some
outliers present, the variable concerning wage can be naturally skewed and outliers
regarding money/wealth can be occasionally expected. */
```

```
.
. /*================================================================================
> (Q6): Consider the homoskedasticity assumption. Do you think the model in (Q2)
> exhibits homoskedasticity or heteroskedasticty? Why?
> *===============================================================================*/
.
. /* The model in (Q2) exhibits heteroskedasticity. According to the graph in (Q5),
you can see that the variances are unequal across the range of values, there are more
variances on the right side (higher years of education). */
```

```
.
. /*==============================================================================
> (Q7): Consider the normality of the errors assumption. Do you think the errors
> in the model in (Q2) follow a normal distribution? Why or why not? Suppose
> assumptions (1)-(4) are true, and the errors follow a normal distribution. Why
> should we care?
> *=============================================================================*/
.
. /* The model in (Q2) does not follow a normal distribution. With a normal
distribution, random variables are continuous and are symmetric around their means.
The graph (Q5) shows that the plots are skewed towards the right and are not linear,
they do not closely follow along the line. */
```

```
.
. /*==============================================================================
> (Q8): Return to your results from (Q2). Interpret beta1hat in a sentence.
> Round to two decimal places.
> *=============================================================================*/
.
. /* On average, an additional year of education will increase the hourly wage by
> $1.39.
>
> beta1hat = 1.390671 */
```

```
.
. /*==============================================================================
> (Q9): Is beta1 statistically significant when alpha=0.01? Use the t-statistic
> to justify your answer.
> *=============================================================================*/
.
. /* Yes, beta1 is statistically significant.
>
> H_0: beta1 = 0
> H_1: beta1 != 0
>
> t-statistic = (beta1hat - 0) / std error of beta1hat
>             = 1.39 / 0.11
>             = |12.64| > 2.58
>             = critical value for two-sided alternative where alpha = 0.01
>
> So using the t-statistic, we reject the null hypothesis of no statistical
> significance.
>
> We conclude that beta1 is statitically significant. */
```

```
.
. /*==============================================================================
> (Q10): Test the null hypothesis that beta1=0 vs. the alternative that beta1>0.
> Calculate the p-value for this hypothesis test. What do you conclude? Use
> alpha=0.01.
> *=============================================================================*/
.
. /* Yes, beta1 is greater than 0.
>
>    H_0: beta1 = 0
>    H_1: beta1 > 0
>
>    The value of test statistic for beta1 = 12.64
>    The p value for beta1 = 0.000
>
>    p-value = P(|Z| > |12.64| ) = 2 * P(Z < 12.64 ) = 2 * (12.64) =
>
>    12.64 > 0.01
>
>    So using p-value, we reject the null hypothesis that beta1 equals to 0.
>
>    We conclude that beta1 is greater than 0. */
```

```
.
. /*=============================================================================
> (Q11): Construct a 90% confidence interval for beta0. Interpret your confidence
> interval in a sentence. Round beta0hat and its standard error to two decimal
> places.
> *============================================================================*/
.
. /* CI(90%) for beta0:
> = [beta0-2.58*sehat(beta0), beta0+2.58*se(beta0))
> = [-4.89 - 2.58(1.39), -4.89 + 2.58(1.39)]
> = [-8.45, -1.33]
>
> beta0 = 1.39
> sehat(beta0) = estimate of standard error of beta0 = 0.11
>
> Lower Bound: -8.45
> Upper Bound: -1.33
>
> True value of beta0 lies between -8.45 and -1.33, meaning that without any
> years of education, wages fall between -$8.45 and -$1.33 */
.
. /*=============================================================================
> =============================================================================*/
. cap log close _all
```