Problem Set 4 - Solution

Labour Economics, Winter Semester 2025/26

Submit by Sunday, 18 January, 22:45h ***on Moodle!***

Learning objectives

- Read a scientific paper and extract key overarching information.
- Study functional forms of earnings profiles in education and over the life-cycle.
- Different ways of computing standard errors and interpretation.
- Fixed effects versus differences estimation.

Tasks

The purpose of this assignment is to estimate the rate of return to education using family fixed effect methods by replicating the results of a classic paper on the topic. In particular, you are asked to reproduce some of the results of the paper *Ashenfelter, O. and A. Krueger, "Estimates of the Economic Return to Schooling from a New Sample of Twins," American Economic Review, Vol. 84 (Dec., 1994): 1157-1173*. The dataset used by Ashenfelter and Krueger is on Moodle.

1. Read Ashenfelter and Krueger's paper. Answer the following:

   a) What is their research question and what are the difficulties in answering this question?

   Solution

   Their research question is to study the economic returns to schooling by contrasting the wage rates of identical twins with different schooling levels. One of the difficulties in answering this question is the potential for measurement error in estimates of the economic returns to schooling. However, it is possible that there may be other factors that influence the relationship between schooling and wage rates that are not controlled for in the study. Furthermore, difficulties were to ensure that the observed correlation between schooling and wages were not due to the correlation between schooling and worker's abilities.

   b) What is the autors' approach to overcome these difficulties and (to what extent) does it convince you?

   Solution

In their survey they took some unusual steps to measure a worker's schooling level accurately. They obtained independent estimates of each sibling's schooling level by asking the twins to report on both their own and their twin's schooling. These new data provide a simple and powerful method for assessing the role of measurement error in estimates of the economic returns to schooling.

At the same time, with this data they can control for the underlying similar genetic and upbringing background by including family fixed effects. Thus they remove any selection bias due to "family effects" and compare education and wage corrrelations within twin-pairs that are clean in this respect.

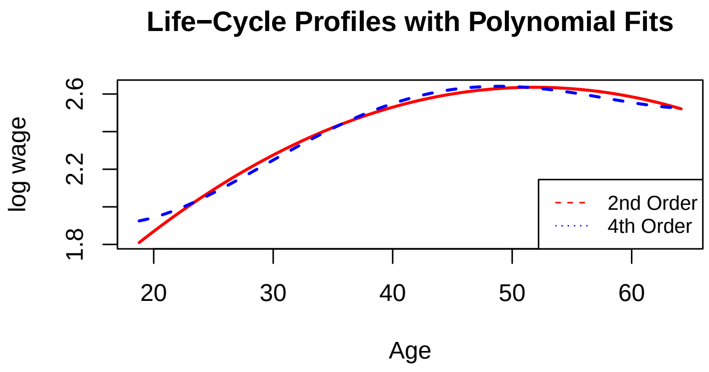
c) What are their key results? Are they surprising / convincing in your view?

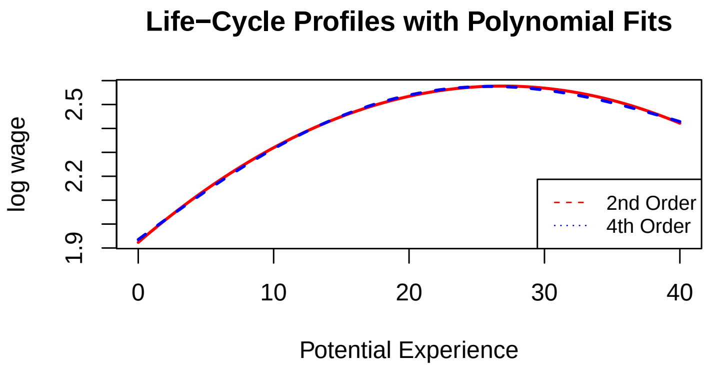
Solution

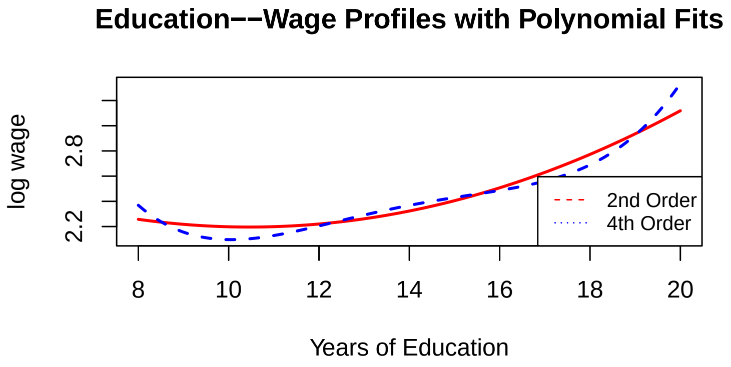
Increased schooling increases average wage rates by about 12–16 percent per year completed. Their results indicate that measurement error may lead to considerable underestimation of the returns to schooling in studies based on siblings. They find some weak evidence that unobserved ability may be negatively related to schooling level. Not sure this latter finding is terribly convincing to me.

2. Read in and explore "`twins_long.dta`". Refer to the article and information below for explanations of the variables. Plot workers' education, age, and experience against their (log) wages.

Solution







Profiles of age and potential experience are concave over the life cycle, as is known from the literature on the Mincer equation. Also, consistent with Mincer's original work, 2nd or 4th order profiles in experience do not really look different for these twins observed in the early 1990s.

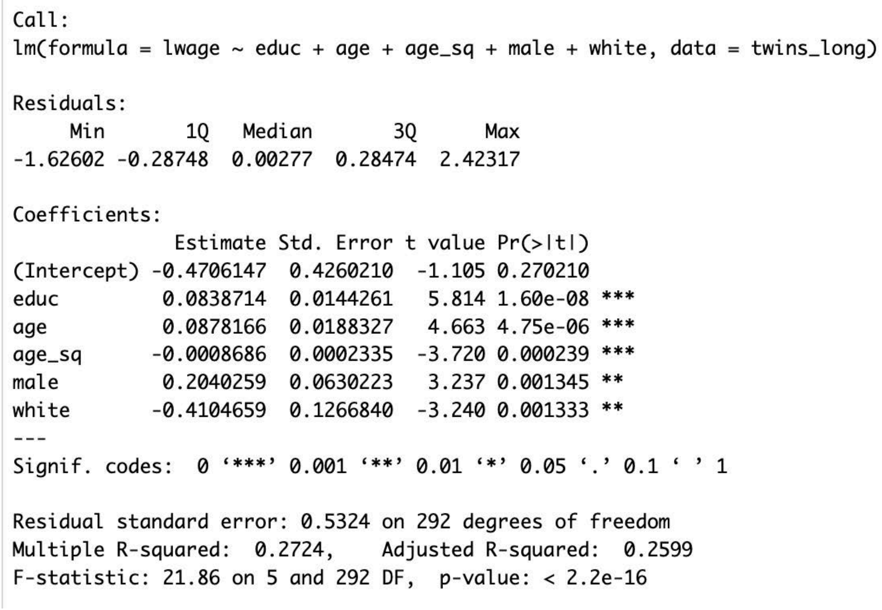
The education profile looks a bit convex, whereas Mincer postulated a linear profile. Perhaps removing the one observation with eight years of schooling might make this look more linear. Also, Mincer's linearity is in a regression where we control for experience, etc. We will see this below...

3. Run an OLS regression of log wages on a constant, schooling, age, age-squared, gender and a racial indicator ("`white`" dummy).

   a) What do the coefficients on age and age-squared imply about the life-cycle profile of earnings? Compare also to the plot from 2.

   Solution

   We find a positive relationship in age, so a higher age implies an higher wage: When we age by one year, wage changes by an average of 8.4 percent. With a negative relation in age squared, we can say, that our effect gets weaker, and after some point even negative, as the people get older. Our findings match with the plot from 2, as the wage increases until the participants got 50 years old and after this age, the trend goes negative. Life-cycle profiles are concave.



```
Call:
lm(formula = lwage ~ educ + age + age_sq + male + white, data = twins_long)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.62602 -0.28748  0.00277  0.28474  2.42317 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.4706147  0.4260210  -1.105 0.270210    
educ         0.0838714  0.0144261   5.814 1.60e-08 ***
age          0.0878166  0.0188327   4.663 4.75e-06 ***
age_sq      -0.0008686  0.0002335  -3.720 0.000239 ***
male         0.2040259  0.0630223   3.237 0.001345 ** 
white       -0.4104659  0.1266840  -3.240 0.001333 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5324 on 292 degrees of freedom
Multiple R-squared:  0.2724,	Adjusted R-squared:  0.2599 
F-statistic: 21.86 on 5 and 292 DF,  p-value: < 2.2e-16
```

b) Do your estimates match those in column (1) of Table 3 in Ashenfelter and Krueger? What kinds of biases is the OLS coefficient on `educ` likely to incorporate?

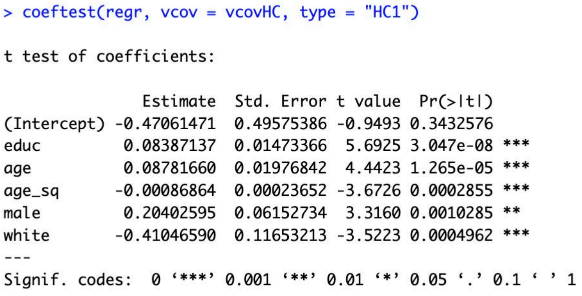
Solution

Yes, we obtain the same estimates. A bias for the coefficient on educ may be the measurement error ("attenuation bias") or the fact that some individuals may choose to enter higher-paying jobs and higher levels of education because they are more talented or motivated in the first place ("ability bias").

4. Post-estimation adjust the standard errors, once reporting coefficients with HC1 robust errors (e.g., `coeftest(model, vcov = vcovHC, type = "HC1")`) and once by additionally clustering at the family (`famid`)level. Do you know why one might want to do this and does the significance of the coefficients change?

Solution

HC1 robust

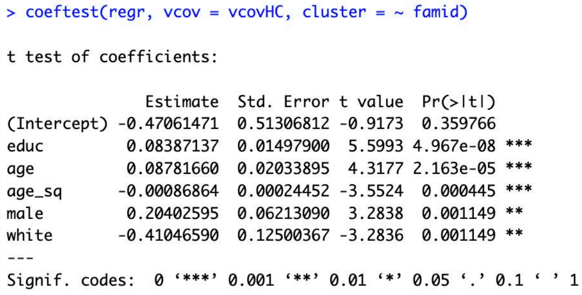


```
> coeftest(regr, vcov = vcovHC, type = "HC1")

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -0.47061471  0.49575386 -0.9493 0.3432576    
educ         0.08387137  0.01473366  5.6925 3.047e-08 ***
age          0.08781660  0.01976842  4.4423 1.265e-05 ***
age_sq      -0.00086864  0.00023652 -3.6726 0.0002855 ***
male         0.20402595  0.06152734  3.3160 0.0010285 ** 
white       -0.41046590  0.11653213 -3.5223 0.0004962 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Clustered at family level



```
> coeftest(regr, vcov = vcovHC, cluster = ~ famid)

t test of coefficients:

               Estimate  Std. Error t value  Pr(>|t|)    
(Intercept) -0.47061471  0.51306812 -0.9173  0.359766    
educ         0.08387137  0.01497900  5.5993 4.967e-08 ***
age          0.08781660  0.02033895  4.3177 2.163e-05 ***
age_sq      -0.00086864  0.00024452 -3.5524  0.000445 ***
male         0.20402595  0.06213090  3.2838  0.001149 ** 
white       -0.41046590  0.12500367 -3.2836  0.001149 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

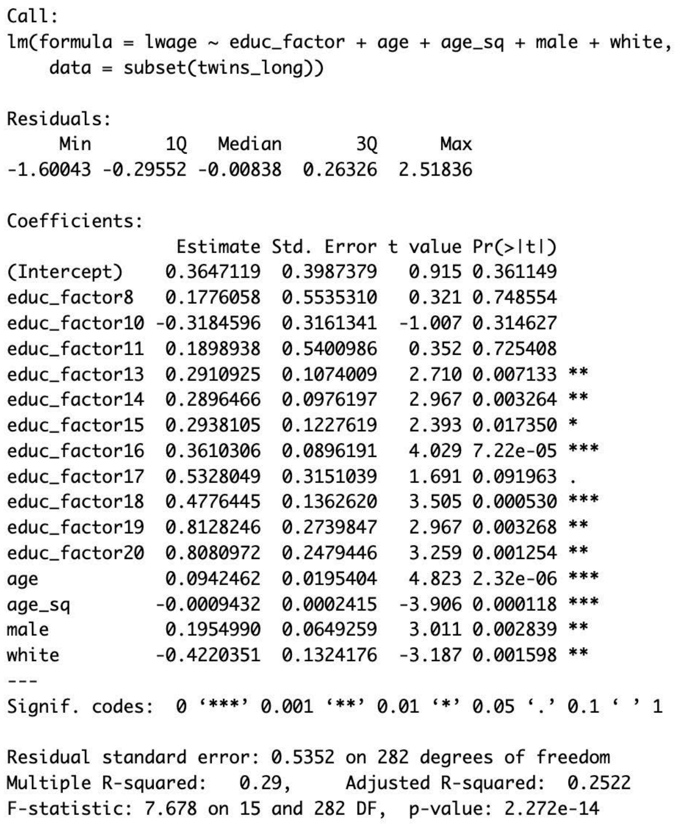
HC1 robustness means it adjusts our standard errors to allow for heteroscedasticity (i.e., that variances of error terms may vary over the distribution of the x-variables). Generally one wants to allow for this realistic possibility and thus be conservative with the reported standard errors (i.e., rather err by reporting too high than too low standard errors).

Clustering at the family level allows for correlation of error terms within family when constructing standard errors. This is plausible because random factors affecting one twin may likely also work on the other twin so they are not exactly two independently sampled observations. Therefore we do not have as much identifying variation as when $149 \times 2 = 298$ observations were independently sampled. Clustering adjusts for this.

Both of these can raise or lower standard errors (often they raise them). Both of these adjustments to standard errors do not change the coefficient estimates themselves!

5. Now create dummy variables for each level of schooling and run the above OLS regression with those education dummies instead of the educ variable. Ideally, omit the dummy for high school (`educ=12`) from the regression, so that all other education levels' coefficients are relative to this largest category. Do the estimated coefficients indicate that the effect of education of log wages is linear?

Solution

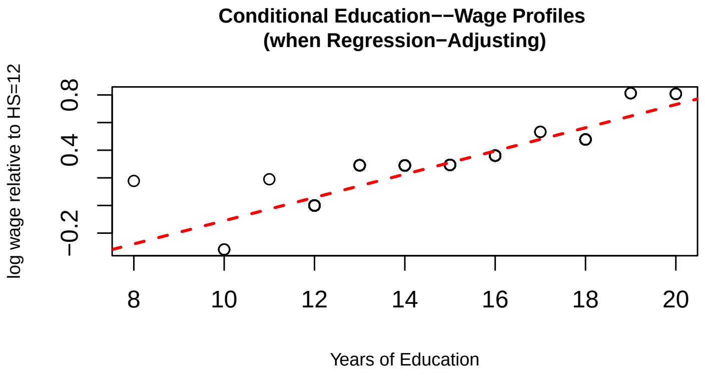


```
Call:
lm(formula = lwage ~ educ_factor + age + age_sq + male + white,
    data = subset(twins_long))

Residuals:
     Min       1Q   Median       3Q      Max 
-1.60043 -0.29552 -0.00838  0.26326  2.51836 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)    0.3647119  0.3987379   0.915 0.361149    
educ_factor8   0.1776058  0.5535310   0.321 0.748554    
educ_factor10 -0.3184596  0.3161341  -1.007 0.314627    
educ_factor11  0.1898938  0.5400986   0.352 0.725408    
educ_factor13  0.2910925  0.1074009   2.710 0.007133 ** 
educ_factor14  0.2896466  0.0976197   2.967 0.003264 ** 
educ_factor15  0.2938105  0.1227619   2.393 0.017350 *  
educ_factor16  0.3610306  0.0896191   4.029 7.22e-05 ***
educ_factor17  0.5328049  0.3151039   1.691 0.091963 .  
educ_factor18  0.4776445  0.1362620   3.505 0.000530 ***
educ_factor19  0.8128246  0.2739847   2.967 0.003268 ** 
educ_factor20  0.8080972  0.2479446   3.259 0.001254 ** 
age            0.0942462  0.0195404   4.823 2.32e-06 ***
age_sq        -0.0009432  0.0002415  -3.906 0.000118 ***
male           0.1954990  0.0649259   3.011 0.002839 ** 
white         -0.4220351  0.1324176  -3.187 0.001598 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5352 on 282 degrees of freedom
Multiple R-squared:   0.29,	Adjusted R-squared:  0.2522 
F-statistic: 7.678 on 15 and 282 DF,  p-value: 2.272e-14
```



The coefficients look broadly linear in years of education once we control for age, etc. At least when we ignore the high coefficient on educ==8, which is based on only one observation (might be outlier). This suggests we have to revert our opinion from 2. and that Mincer was right: conditional education profiles of log wages are approximately linear even if there are some upward and downward deviations.

6. Return to the specification that is linear in education. Compare your OLS estimate with one that incorporates a "family fixed effect" (`factor(famid)`). Focus on the coefficient for education (variation in the other variables is mostly removed by the family fixed effect). What are in your view the (economic) reasons for any differences?

Solution

Our regression equation is

$$w_{ij} = \alpha_i + \beta e_{ij} + \mu_j + \varepsilon_{ij} \tag{1}$$

with w_{ij} wage of twin $i \in \{1,2\}$ in family j, e_{ij} the education and μ_j a family effect as in Ashenfelter & Krueger.

Fixed effect estimation takes the mean of the data and subtracts it on both sides:

$$\tilde{w}_{ij} = \tilde{\alpha}_i + \beta \tilde{e}_{ij} + \tilde{\varepsilon}_{ij} \tag{2}$$

where $\tilde{x}_{ij} \equiv x_{ij} - \bar{x}_j$ and $\bar{x}_j \equiv \frac{1}{2}(x_{1j} + x_{2j})$. Fixed effects thus removes selection bias due to family effects. If $\tilde{\varepsilon}_{ij}$ were then uncorrelated with differences in twin couples' education $e_{ij} - \bar{e}_j$, we would get an unbiased coefficient estimate for β.

We also include a twin rank (either 1 or 2) dummy. This is for technical reasons here to get the exactly same estimate as in first differences below (economically, results don't change). In general one would want this control, e.g., if $i \in \{1,2\}$ were years, one would want to control for year fixed effects also (wages grow over time in the economy).

The results are actually higher with a family fixed effect $\hat{\beta}_{FE} = 0.092$. This is the same as in Ashenfelter Krueger Table 3 column (v). It may come from selection bias at the family level but in a different direction than we usually think ability bias goes. That is, usually we think that families with more motivated and talented offspring see higher education and earnings in both. Controlling for family FE, and removing OVB from it, would thus reduce the correlation between education and earnings. Here this goes up – albeit slightly and not statistically significantly compared to $\hat{\beta}_{OLS} = 0.084$ – and thus in the other direction.

7. An alternative to fixed effect estimation is to run a model on first differences. Write out the equation in first differences, estimate that differenced equation and compare the results to column (v) in Table 3. You can first-difference in `twins_long.dta` or compute differences between twins in the "wide" version of the same data `twins.dta`, also provided on Moodle. Should the coefficient of `educ` in the family fixed effect and the first-differenced model be the same? Are they the same?

Solution

Writing equation (1) in first differences gives:

$$\triangle w_j = \triangle\alpha + \beta\triangle e_j + \triangle\varepsilon_j \tag{3}$$

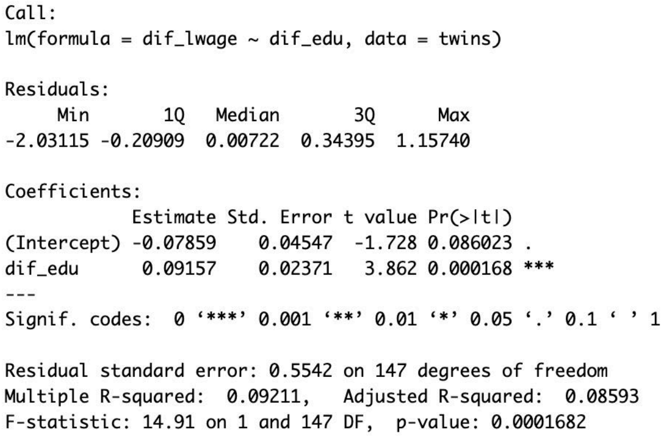
where $\triangle e_j \equiv e_{2j} - e_{1j}$. One can show that first-difference estimation in the probability limit (for large N) leads to the same estimates as fixed effects estimation. For two-period models (here only two twins), they are even numerically the same!

To see this, note that (2) is in fact

$$\frac{1}{2}\triangle w_j = \frac{1}{2}\triangle\alpha + \beta\frac{1}{2}\triangle e_j + \triangle\frac{1}{2}\varepsilon_j, \tag{4}$$

that is, exactly the same regression as (3).

Here is the estimation result:



```
Call:
lm(formula = dif_lwage ~ dif_edu, data = twins)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.03115 -0.20909  0.00722  0.34395  1.15740 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.07859    0.04547  -1.728 0.086023 .  
dif_edu      0.09157    0.02371   3.862 0.000168 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5542 on 147 degrees of freedom
Multiple R-squared:  0.09211,	Adjusted R-squared:  0.08593 
F-statistic: 14.91 on 1 and 147 DF,  p-value: 0.0001682
```

Notes: You can work in teams of 1–3 students. Please upload your pdf-file with responses. It should be clear which answers in the `.pdf` refer to which question. If you work in a team, each member has to upload the group's solution and note whom they worked with.

. . . for some further background information, turn to next page . . .

1) Background for the study

This classic twins study sought to answer what seems a simple question: By how much will another year of schooling most likely raise one's income? Attempts had been made to estimate the value of a year's education in previous studies, but previous estimates may have been imprecise for two reasons. The first, most obvious reason is the difficulty of extracting the education's effect on income from the effect that other variables related to education have on income. That is, a worker's natural ability, his family background, and his innate intelligence are all possible confounding factors that must be controlled for to estimate the effect of education on income accurately. Thus this study interviewed twins, collecting information about education, income, and background. Because monozygotic twins (twins from a single egg) are genetically identical and have similar family backgrounds, they provide an excellent control for confounding variables.

The second difficulty in measuring the effect of income on education has to do with the false reporting of education levels, and this study is the first to address it. Since people are more likely to report a higher education level than they have actually attained, especially in face to face interviews, the data will contain a number of people with lower education levels in the higher education categories. Thus, since education usually increases income, estimates for the precise amount of this increase will be too low. To correct for this bias the researchers interviewed the twins separately and recorded two entries for each individual's education level: his self-reported education level and the education level reported by his twin. This allowed them to estimate the "measurement error" of reported education levels and correct for it. The result was a much higher estimate of the effect a year of education is likely to have on one's income. In fact, this study's estimates were higher than those of all previous studies, which did not correct for measurement error in education level.

2) Brief description of the data

The data were collected by a team of five interviewers at the 16th Annual Twins Day Festival in Twinsburg, Ohio, in August 1991. A booth was set up at the festival's main entrance, and an ad inviting all adult twins to participate in the survey was placed in the festival program. In addition, the interviews roamed the festival grounds, approaching all adult twins for an interview, and almost every pair of twins accepted.

The key variables are:
famid = family id
age = age of the person
educ1 and educ2 = education attainment of twin 1 and 2 , respectively
lwage1 and lwage2 = the natural log (ln) of the hourly wage of twin 1 and 2 , respectively
male = an indicator variable equal to one if the person is male, zero otherwise
white = an indicator variable equal to one if the person is white, zero otherwise