**POLS 6481, Spring 2021**

Instructor: Prof. Scott Basinger

Homework Assignment 2

Distributed Wednesday, February 10, 2021

Due Tuesday, February 16, 2021

Is it generally true that height confers unfair advantages? Conventional wisdom holds that taller candidates win presidential elections, and taller executives tend to earn higher salaries.

Download the dataset *HeightWage\_MenWomenUS\_HW.csv* or *HeightWage\_MenWomenUS\_HW.dta*. These data are drawn from the National Longitudinal Study of Youth, 1979 cohort (age 14 to 22 years old in 1979). Our dependent variable is the wages earned in 1996 (*wage96*). Here is a codebook:

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| *male* | Male (1 = yes, 0 = no) |
| *white* | White (1 = yes, 0 = no) |
| *black* | Black (1 = yes, 0 = no) |
| *hispanic* | Hispanic (1 = yes, 0 = no) |
| *wage96* | Adult hourly wages (dollars) reported in 1996 (salary and wages in past calendar year divided by hours worked in past calendar year) |
| *height85* | Adult height (inches), self-reported in 1985 |
| *height81* | Adolescent height (inches), self-reported in 1981 |
| *athlets* | Participation in high school athletics (1 = yes, 0 = no) |
| *clubnum* | Number of club memberships in high school, excluding athletics, academic/honor society clubs, and vocational clubs |
| *momed79* | Mother’s years of education |
| *daded79* | Father’s years of education |
| *mompro2* | Mother in a professional/managerial occupation (1 = yes, 0 = no) |
| *poppro2* | Father in a professional/managerial occupation (1 = yes, 0 = no) |
| *siblings* | Number of siblings |
| *age* | Age (years) in 1996 |
| *esteem80* | Score on Rosenberg Self-Esteem Scale as an adolescent in 1980 (higher values indicate higher self-esteem) |
| *hgc96* | Highest grade of education completed in 1996 |

Also download the R script titled “*HW2q1-4prep.R*” and run the script to (a) eliminate cases with missing observations – but only for the variables you care about – and (b) eliminate variables you won’t care about and (c) limit the sample to male subjects, using the *male* variable.

1. Estimate the following bivariate regression model using ordinary least squares:

* What does this regression suggest about the relationships between height and wages?
* Can the slope coefficient, , be statistically distinguished from zero?

2. Estimate the following multivariate regression model using ordinary least squares, which controls for the individual’s height as an adolescent:

* How does the estimated slope coefficient from the multivariate model compare with the estimated slope coefficient from the simple regression model?
* How does the standard error of from the multivariate model compare to the bivariate regression model?
* Is from the multivariate model statistically significant, i.e., distinguishable from zero?
* Use the equations from lecture to solve for the standard errors for the simple and multiple regression coefficients for *height85*. Explain how the numerator changed, then explain how and specifically why the denominator changed, relative to the simple regression model.
* What is the correlation between *height85* and *height81*?
* How much of an effect did multicollinearity have on and/or its standard error?

The researchers were interested in whether the effect of adolescent height on wages is partly due to height encouraging participation in sports and belonging to clubs. Adding two variables to the analysis will allow you to evaluate whether taller adolescents simply had more opportunities to develop leadership and other skills (i.e., social capital).

3. Estimate the following regression models using ordinary least squares:

* What do these regressions suggest about how height and social capital (belonging to clubs and/or participating in athletics) affect wages?
* In the last regression above, which slope coefficients are “statistically significant”?
* In the last regression above, how does the estimated slope coefficient 2 compare with the estimated slope coefficients from the earlier regression models?
* How does the standard error of the estimated slope coefficient 2 from the multivariate model compare to standard errors from the earlier regression models?
* What is the correlation between *clubnum* and *athlets*?
* What is the correlation between each of these variables and *height81*?

Maybe there’s more to life than just making money. Perhaps taller people get greater social capital, which increases self-esteem. The dataset includes scores on the Rosenberg Self-Esteem Scale, taken as an adolescent in 1980; higher values indicate more self-esteem.

4. Estimate the following regression models using ordinary least squares:

* What do these regressions suggest about the relationship between height and self-esteem?
* What does the second regression suggest about the relationship between social capital (belonging to clubs and/or participating in athletics) and self-esteem?
* Which slope coefficients are “statistically significant”?
* Is multicollinearity a potential problem here? Use the casual methods described in lecture 8 to describe why it is or isn’t a problem?

Does raising education levels tend to improve economic growth? For their article, “Do Better Schools Lead to More Growth? Cognitive Skills, Economic Outcomes, and Causation,” Hanushek and Woessmann (2012) collected a dataset on economic growth of 50 countries from 1960 to 2000. Here is a codebook:

|  |  |
| --- | --- |
| **Variable name** | **Description** |
| *code* | Country code |
| *name* | Country name |
| *open* | Openness of the economy scale |
| *ed60* | Average years of schooling in 1960 |
| *ypc60* | GDP per capita in 1960 |
| *ypcgr* | Average annual growth rate (GDP per capita), 1960–2000 |
| *testavg* | Average combined math and science standardized test scores, 1964–2003 |
| *proprts* | Security of property rights scale |
| *edavg* | Average years of schooling, 1960–2000 |
| *region* | Region |

Download the dataset *globaled.csv* or *globaled.dta* and use it to answer questions 5–8:

1. Estimate a regression in which *ypcgr* is the dependent variable and *edavg* is the independent variable.

* Examine the coefficient for *edavg*; what effect does education have on growth?
* Examine the *t* statistic and *p* value; is the estimated coefficient statistically significant (*p* < .05)?
* How much of the variation in average economic growth is explained by average education?

5½. Optional question: estimate a regression in which *ypcgr* is the dependent variable and *ypc60* is the independent variable.

* Examine the coefficient for *ypc60*; what effect did GDP per capita in 1960 have on subsequent growth?
* Examine the *t* statistic and *p* value; is the estimated coefficient statistically significant (*p* < .05)?

1. Add the *ypc60* variable to the regression estimated in 5. The stated purpose of doing this is to control for GDP per capita, given the phenomenon that countries that are wealthier have slower growth rates, since poorer countries simply have more capacity to grow. You’ll also see that countries that are wealthier tend to have higher average years of education.

* Create a scatterplot with GDP per capita in 1960 on the horizontal axis and average years of education since 1960 on the vertical axis. How, and how strongly, are these variables related?
* Examine the coefficients; how did *edavg* and *ypc60* impact average growth since 1960?
* Examine the standard errors and *t* statistics for *edavg* and *ypc60*; is either of the estimated coefficients statistically significant (*p* < .05)?
* Examine the variance-covariance and/or correlation matrix. Why did the coefficient on *edavg* change so much between the simple regression and multiple regression?
* Once you control for GDP, how much does each additional year of schooling contribute to average economic growth?
* How much of the variation in average economic growth is explained by the combination of average education and GDP per capita in 1960?

1. Add another variable, *testavg*, to the regression estimated in 6. This variable average standardized test scores in math and science. Hanushek and Woessmann believed that quality of education would matter more than quantity of education.

* Create a scatterplot with average years of education on the horizontal axis and average standardized test scores on the vertical axis. How, and how strongly, are these variables related?
* Explain how omitting average standardized test scores (*testavg*) could have created omitted variable bias. Identify which cell of Wooldridge’s Table 3.2 this example would be classified in, treating the *edavg* variable as *x*1and the *testavg* variable as *x*2?
* Examine the coefficients for *edavg* and *ypc60* and *testavg*; what effect does each variable have on average economic growth?
* How has the coefficient on *edavg* changed from the regression you ran in 6.? How has the coefficient on *ypc60* changed from the regression you ran in 6.?
* Examine the standard errors of the coefficients and *t* statistics for *edavg* and *ypc60* and *testavg*; which estimated coefficients are statistically significant (*p* < .05)?
* Examine both the residual standard error (*sigma-hat*) and the standard errors of the regression coefficients for the *edavg* and *ypc60* variables; how did they change when you added *testavg* to the equation?
* Does adding the *testavg* variable introduce a problem of multicollinearity? Use two of the “casual” methods for detecting multicollinearity and calculate Variance Inflation Factors.

1. Drop the *edavg* variable, leaving only the *ypc60* and *testavg* variables in the regression.

* Have you created an omitted variable bias problem by dropping the *edavg* variable? Provide appropriate support for your conclusion (consult Lab 3 for ideas).
* What happens to the coefficient, standard error, and *t* statistic for *testavg* when the *edavg* variable is dropped?
* How much of the variation in average economic growth is explained by all three variables (*edavg* and *ypc60* and *testavg*), and how much of the variation is explained by just the last two variables (*ypc60* and *testavg*)?