Jason Yoon
ECN 140 B01
3/13/17
Empirical Project

Does Job-Training Improve Earnings?

1. Introduction

Individuals in the workforce will usually find ways to earn more value from their work; they will either try to improve their efficiency or gain more training and knowledge. The objective is to see if job-training affects earnings in a significant way and if policy-makers should invest more funds into training programs. The question is important because it will help show policy-makers a way to positively affect the wellbeing of the workforce by potentially increasing their earnings. Random individuals are given opportunity to undergo job-training and earnings are recorded 30 weeks after the assignments alongside the whole sample population. From the random assignments, effects of job-training on earnings can by studied and help enforce current economic opinion that more training increases earnings for individuals.

2. Data Description

The dataset is the accumulation of a large publicly funded training program in the late 80's and 90's. Random individuals were chosen, like in a lottery, and given the opportunity to participate in job-training. A treatment group formed from the individuals that participated, although not everyone decided to go through with the training. Most of the variables in the dataset describe individual characteristics such as sex, race, marital status, and education; these can be considered categorical variables. Ethnicity is made of up of Black and Hispanic; the category that is neither Black or Hispanic is left ambiguous. The category of neither Black or Hispanic can be thought of as White, but that would leave out the possibility of Asian or even

Jason Yoon
3/13/17

Native American ethnicity; creating a new race variable risks the dummy variable trap. Because of this misspecification of data, there will be some correlation between the categorical race variables and the residual. The educational level that is recorded in the study is a binary variable, only counting if the observed individual has a High School or G.E.D. degree. This should show the scope of the study since it only measures academic degrees of the High School level. The study is likely to observing individuals with low to medium income. The last two variables show if the individual worked less than 13 weeks in the past year or if they are receiving AFDC (Aid to Families with Dependent Children) at the time of the study. There are numerical variables such as earnings 30 weeks after the assignment, birthdate of the individuals, and age. For the study, earnings are transformed and put into log form for symmetricity and to help show more linear relationships between variables. Figure 1 shows the histogram of log earnings of people who did not receive training; Figure 2 shows log earnings of participants who received training.

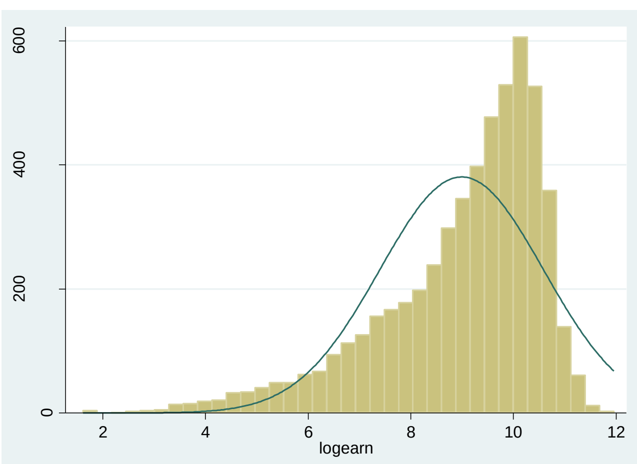

Figure 1: Log Earnings without training

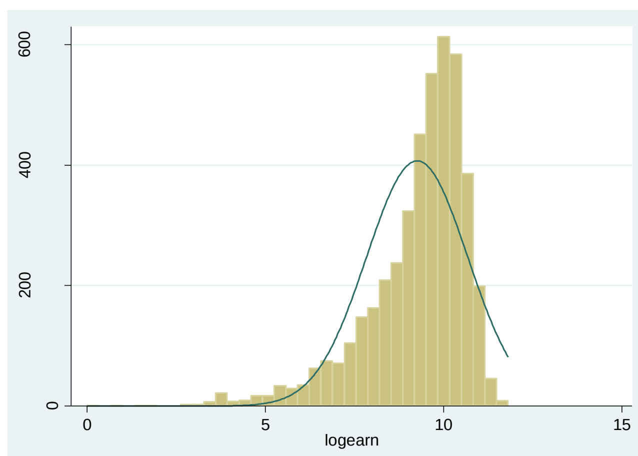

Figure 2: Log Earnings with training

Figure 1 omits earnings of people who underwent trainings, while Figure 2 does the same for no training. Figure 1 has a slightly lower mean and higher standard deviation than Figure 2 and does not seem to be as normally distributed. These exogenous variables determine the individual and determine earnings with or without training programs; they can be used to see the effect of training. Some of the data was manipulated; the variables *married*, *hsorged*, and *wkless13* had values that did not correspond to their binary nature. 764 observations in *married* were changed to 0 since the had a value less than 0.5. In *hsorged* 695 observations were changed to 1 since their original value exceed 0.5. In *wkless13*, 454 were changed to 0 and 724 were changed to 1 since they were lower and exceeded 0.5 respectively. This error could be either human or format error; either way observations were changed to fit their most likely values.

Table 1: Categorical Values

Categorical Variable	**Description**	**Dummy Variable**	**Description**	**Mean**
Assignment	If the individual was randomly chosen for the training program	assignmt	=1 if randomly chosen	0.6682
Training	If the individual chose to do the training program	training	=1 if received training or not	0.4336
Gender	Gender of the individual	sex	=1 if male	0.4554
Married	If the individual is married	married	=1 if married	0.2629
Race	The ethnic group the individual belongs to	black	=1 if race is Black	0.2596
		hispanic	=1 if race is Hispanic	0.1093
Education	If the individual has a High School or G.E.D. degree	hsorged	=1 if individual has the degrees	0.7265
AFDC	If the individual is receiving aid for their family	afdc	=1 if receiving aid	0.1869
Worked Less	If the individual has worked less than 13 weeks in the past year	Wrkless13	=1 if worked less than 13 weeks	0.4780

Table 2: Numerical Values

Numerical Variable	Description	Mean	Standard Deviation	Min	Max
earn	Total earnings 30 months after the assignment takes place	15,815.29	16,767.05	0	155760
logearn	Log of earn	9.111	1.527	0	11.96
bdate	Date of Birth	- 1,788.38	3,522.10	- 18,133	2,740
age	Age at the time of assignment	33.14	9.64	22	78

3. Methodology

There will be problems if only the OLS model is used to predict the effect of training on earnings. The basic OLS model used is as follows:

$$\boldsymbol{logearn = \beta_0 + \beta_1 training + \beta_2 sex + \beta_3 married + \beta_4 (sex * married) + \beta_5 hsorged + \beta_6 black + \beta_7 hispanic + \beta_8 (wkless13 * afdc) + \mu} \quad (1)$$

Using this model there are various classical linear model assumptions that are violated. Using the Breusch-Pagan Lagrange multiplier test, heteroskedasticity is present violating assumption of homoskedasticity. Intuitively this seems correct, since the model includes age as a predictor on earnings (confirmation from economic theory and practice that earnings on age produce heteroskedastic errors). Although there is no omitted variable bias in the model, there is a strong probability of selection bias within the model in the form of voluntary response bias. Training may increase earnings, but we are not sure if this increase is only due to training. Since accepting training after assignment was an active decision, the individuals that chose to do training when assigned could be possibly be more passionate, skilled, or driven to increase their earnings for a personal reason. The uncaptured information and bias will violate the assumptions of random sampling and the zero-conditional mean alongside with homoskedasticity. A new model is created to fight these violations; an instrumental variable should be used while controlling for robust errors. The instrumental variable is chosen to be *assignmt*, therefore *assignmt* must be proven to be random and uncorrelated with the residual or affect *logearn* directly. An interaction term is used between *wrkless13* and *afdc* to help capture the effect of individuals who cannot work due to their children. This is an assumption, but it seems like an effect that can be observed to better form the model. If an individual is receiving aid for dependent children and not working, it seems more than likely they are taking care of their children. *Sex* and *married* also form an interaction term because the model might overstate the coefficient and significance because of double counting. There might be married couples within the sample so the interaction term is in place to ward against that bias. Married couples within the sample also affect earnings as the couples might count two earnings.

3a. Logit & Probit Model for Assignment Randomness

The Logit and Probit models can be used to confirm the randomness of assignment of training on the sample population. This means a null hypothesis is formed that the explanatory variables are not statistically different from 0 when trying to predict change of assignment. From the Figure 3 (next page) no variables are significant in predicting the probability of *assignmt*, and the percentage correctly predicted is the percentage of people randomly drawn from the sample. From using the logit and probit models, it is taken that *assignmt* is random and therefore not correlated with the residual. The statistical insignificance of the variables mean the variables cannot accurately predict the probability of assignment. In using the binary outcome models there was also no consideration in using robust standard errors to control for heteroskedasticity even though it was observed in the basic model. The topic and understanding of using robust errors in binary models is beyond the scope of this study.

.

	(1)	
	assignmt	
assignmt		
0.sex	0	(.)
1.sex	-0.0410	(-0.80)
0.married	0	(.)
1.married	0.104	(1.47)
0.sex#0.ma~d	0	(.)
0.sex#1.ma~d	0	(.)
1.sex#0.ma~d	0	(.)
1.sex#1.ma~d	0.0207	(0.22)
bdate	-0.0000809	(-0.77)
age	-0.0314	(-0.82)
hsorged	0.0629	(1.39)
black	0.0515	(1.06)
hispanic	-0.00110	(-0.02)
0.wkless13~c	0	(.)
0.wkless13~c	0.0708	(0.71)
1.wkless13~c	0.0651	(1.40)
1.wkless13~c	-0.0340	(-0.51)
_cons	1.506	(1.39)
N	11204	

z statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

	(1)	
	assignmt	
assignmt		
0.sex	0	(.)
1.sex	-0.0252	(-0.80)
0.married	0	(.)
1.married	0.0635	(1.47)
0.sex#0.ma~d	0	(.)
0.sex#1.ma~d	0	(.)
1.sex#0.ma~d	0	(.)
1.sex#1.ma~d	0.0127	(0.22)
bdate	-0.0000496	(-0.78)
age	-0.0193	(-0.82)
hsorged	0.0385	(1.39)
black	0.0315	(1.06)
hispanic	-0.000780	(-0.02)
0.wkless13~c	0	(.)
0.wkless13~c	0.0432	(0.71)
1.wkless13~c	0.0400	(1.41)
1.wkless13~c	-0.0206	(-0.50)
_cons	0.929	(1.41)
N	11204	

z statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 3: The Logit (left) and Probit (right) Tests

3b. IV Regression

For a *assignmt* to be a valid instrument, it must be relevant and exogenous to fulfill the assumptions of an IV. The opportunity for training is only received when an individual is randomly assigned and as such the variable of assignment is correlated with the endogenous variable (*training)*. In the previous section, *assignmt* is proved to be random using the variables in the study, as such it can be strongly assumed that *assignmt* is not correlated to the residual. With the two key assumptions fulfilled, *assignmt* can now confidently be used as an instrument for *training*.

<u>The structural equation:</u>

$$\begin{aligned} \boldsymbol{logearn} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 training} + \boldsymbol{\beta_2 sex} + \boldsymbol{\beta_3 married} + \boldsymbol{\beta_4}(\boldsymbol{sex} * \boldsymbol{married}) + \boldsymbol{\beta_5 age} \\ + \boldsymbol{\beta_6 hsorged} + \boldsymbol{\beta_7 black} + \boldsymbol{\beta_8 hispanic} + \boldsymbol{\beta_9}(\boldsymbol{wkless13} * \boldsymbol{afdc}) \\ + \boldsymbol{\mu} \end{aligned} \quad (1)$$

<u>The reduced-form model:</u>

$$\begin{aligned} \boldsymbol{training} = \boldsymbol{\pi_0} + \boldsymbol{\pi_1 assignmt} + \boldsymbol{\pi_2 sex} + \boldsymbol{\pi_3 married} + \boldsymbol{\pi_4}(\boldsymbol{sex} * \boldsymbol{married}) + \boldsymbol{\pi_5 age} \\ + \boldsymbol{\pi_6 hsorged} + \boldsymbol{\pi_7 black} + \boldsymbol{\pi_8 hispanic} + \boldsymbol{\pi_9}(\boldsymbol{wkless13} * \boldsymbol{afdc}) \\ + \boldsymbol{v_2} \end{aligned} \quad (2)$$

The reduced-form model is also known as the first-stage equation of the 2SLS regression. There should be some consideration that a linear probability model is used to predict the value of instead of the binary value of training. This regression to predict values for *training* shows the variation of individuals who underwent training when they were randomly assigned. Using *assignmt* as a IV for *training* gives us the second stage in the 2SLS model.

<u>The second-stage model:</u>

$$\begin{aligned} \boldsymbol{logearn} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 \widehat{training}} + \boldsymbol{\beta_2 sex} + \boldsymbol{\beta_3 married} + + \boldsymbol{\beta_4}(\boldsymbol{sex} * \boldsymbol{married}) + \boldsymbol{\beta_5 age} \\ + \boldsymbol{\beta_6 hsorged} + \boldsymbol{\beta_7 black} + \boldsymbol{\beta_8 hispanic} + \boldsymbol{\beta_9}(\boldsymbol{wkless13} * \boldsymbol{afdc}) \\ + \boldsymbol{\mu} \end{aligned} \quad (3)$$

The reason of using the log of earnings was mentioned for symmetricity for more linear relationships, but another benefit is to change effects into a semi-elastic interpretation. It is important to note that when regressing the final model heteroskedasticity needs to be controlled again. Figure 3 shows the regression results when *assignmt* is used as an IV for *training*.

	(1)	
	logearn	
training	0.116*	(2.39)
0.sex	0	(.)
1.sex	0.0183	(0.49)
0.married	0	(.)
1.married	-0.0302	(-0.58)
0.sex#0.ma~d	0	(.)
0.sex#1.ma~d	0	(.)
1.sex#0.ma~d	0	(.)
1.sex#1.ma~d	0.502***	(7.30)
age	-0.00628***	(-3.91)
hsorged	0.249***	(7.31)
black	-0.0925**	(-2.58)
hispanic	-0.0299	(-0.61)
0.wkless13~c	0	(.)
0.wkless13~c	-0.357***	(-5.16)
1.wkless13~c	-0.499***	(-14.56)
1.wkless13~c	-0.795***	(-14.58)
_cons	9.309***	(133.01)
N	9872	

Z statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

Figure 3: Regression results using the final model

4. Analytical Results

From the regression and using Figure 3, *training* has the coefficient of 0.116 and is significant at the 5% level. The magnitude of the effect is quite large even when controlling for the variable for education. The positive effect of training and increase earnings by about 11.7%, but it should be put into comparison. The effect of having a high school degree or G.E.D. (*hsorged*, z stat = 7.31) is more valued than job-training at face value. Still it can be interpreted that those with a High School education or G.E.D. still benefit from job-training. This should show policy-makers that it is very worth investing into applying job-training to increase earnings of workers even if they have a High School education.

The interaction between *wkless13* and *afdc* has very big magnitudes of statistical significance. The negative correlation is expected since working less has a direct negative effect on earnings and applying for aid already means earnings are low. The other interaction term between *sex* and *marriage* is very significant. Although *sex* or *marriage* alone are not significant, it seems the base group of a married male is most indicative of high earnings with a coefficient of 0.502. This means if the individual is married and male, their earnings will go up by about 50% ceteris paribus. The variable for age has a negative effect with a very small magnitude, although it is still statistically significant. Interpretation says that as individuals get older they earn slightly less; this goes against intuition that experience increases wage, but it should be in the scope of the study. The job could include physical labor or something of that nature; Age would have negative effect since the minimum value of age is already 22. The variable *hispanic* is the only variable that is not significant with a coefficient of -0.0282 and z-statistic of 0.569. This may be because of the small Hispanic population within the sample; the total number of

Hispanic people is 1,225 or about 10.93% of the sample. To put this in comparison, Black individuals make up about 2,909 or 25.96% of the sample so Hispanics seem underrepresented. The statistical insignificance of *hispanic* maybe also be because of omitted *logearn*. This happens because there were many reported earnings of $0, so while log of earnings gives us a practical semi-elasticity it omits those observations.

As a post-estimation measure, the Durbin-Wu-Hausman (DWH) test can be used to show endogeneity of the instrumented variable (*training*). The DWH test uses the augmented regressors so it produces a robust test statistic. Under the null hypothesis of this test *training* is exogenous, but because the p-value of the test is 0.0003 we can reject to null; we are correct in assuming *training* is endogenous as shown in Figure 4. The model is also just-identified, since there are the same number of IVs and endogenous variables. There is also a strong correlation between *training* and *assignmt* as seen in Figure 5 so there is confirmation that *assignmt* continues to be a strong instrument.

```
Tests of endogeneity
Ho: variables are exogenous

Robust score chi2(1)            =  13.2663  (p = 0.0003)
Robust regression F(1,9859)     =  13.2758  (p = 0.0003)
```

Figure 4: DWH test for endogeneity

	training	assignmt
training	1.0000	
assignmt	0.5958	1.0000

Figure 5: Correlation between *training* and *assignment*

5. Concluding Remarks

Analysis of the data shows that participating in job-training increases earnings for all randomly assigned individuals. The effect of training remains statistically significant when instrumented with random assignment and supports popular economic theory. This means the increased earnings from training are statistically significant as well as economically significant

Jason Yoon
3/13/17

even when controlling for other exogenous variables such as education. Although the effect of education on earnings is greater than the effect of job-training, both variables should be actively desired to improve earnings. Improved earnings imply improved efficiency therefore policy-makers should actively try to invest in more job-training for individuals. Job-training would also be much easier, take less time, and be more practical for most individuals to complete as well, although education should still be advocated. Completion of job-training will most definitely increase the individual's earnings.

Something that is not captured in the paper's models or the data is the individual behavior and decision to undergo the training program. There should be more studies into what individuals consider going through training programs like proximity to the training site, transportation, amount of extra-time available, and much more. These extra variables would help show the difference in the individual that accepts the training. Increased earnings with job-training is already established to be statistically and economically significant, so future research should focus on getting maximum participation from the training programs.