The Effect of Employee Training Programs on Employee Income

Siyi Liu

2022/04/30

Contents

Code and data supporting this analysis is available at: https://github.com/siyiliu1202/Employment-training-program-study

1 Abstract

Good-quality worker training can help workers secure good jobs, increase the efficiency of businesses and corporations, and enhance productivity in the economy. While worker training programs are constantly encouraged in almost every industry to promote worker engagement and growth, its effect on salary or earnings remains debatable across industries. This analysis examines whether there is an effect of worker training programs on worker's earnings. We analyzed the Lalonde dataset and used propensity score matching to balance out several variables so we can have a more causal conclusion. We used a linear regression model to examine the effect of interest. Our results illustrated that worker training program does not have a statistically significant effect on worker's earnings (P-value = 0.169). We conclude that worker training program does not lead to higher earnings, after controlling for relevant variables.

Key words: employee training programs, Lalonde dataset, regression analysis, propensity score matching

2 Introduction

Employee or worker training program is a program that helps employees learn specific knowledge or skills to improve performance in their current roles or future roles (Deutsch 1987). Good-quality worker training can help workers secure good jobs, increase the efficiency of businesses and corporations, and enhance productivity in the economy (Bartel 1994). Unfortunately, the United States and Canada support very little worker training, and the training it does support frequently fails to lead to favourable jobs or boost productivity (Deutsch 1987; Bartel 1994; Kluve, Rother, and Puerta 2012). Government policy is not currently up to the challenge, and neither businesses nor employees can solve these problems on their own (Deutsch 1987). For this reason, people have been urging for a new kind of policy to ensure that employee or worker training improves the productivity of the workforce and leads to more well-paid jobs (Deutsch 1987). However, this is all based on conjecture. While worker training programs are constantly encouraged in almost every industry to promote worker engagement and growth, its effect on salary or earnings remains debatable across industries (Kluve, Rother, and Puerta 2012).

Current there are a lot of gaps to fill in this particular area from both scientific research and practical implementation points of view. The demand to conduct randomized experiments in the context of manpower training programs, and in analyzing effects based on causality in general, has been a topic of much debate among academic researchers and government policy makers (Deutsch 1987; Bartel 1994).

The research question of this analysis is to examine whether there is an impact of worker training programs on worker's earnings. The hypothesis is that worker training programs do not have any impact on worker earnings. We answered the research question and tested the hypothesis by analyzing the Lalonde dataset (LaLonde 1986). The Lalonde dataset contains demographic and socioeconomic variables such as age, marital status, race, and income on the sample (LaLonde 1986). This is the first study to evaluate the effect of employment training programs on earnings (LaLonde 1986; Dehejia and Wahba 1999). In this paper, we first describe the data used for the analysis, then we describe the statistical model built for the analysis in detail and present analysis results, and finally discuss our findings and end with concluding remarks.

3 Data

Data collection

The Lalonde dataset contains data collected from the National Supported Work (NSW) Demonstration and comparison groups drawn from the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID) (LaLonde 1986; Dehejia and Wahba 1999). The dataset contains demographic and socioeconomic variables such as age, marital status, race, and income on sampled individuals in the American workforce between 1974 and 1978 (LaLonde 1986; Dehejia and Wahba 1999). The sampled individuals were divided into two groups: one group completed a worker training program designed and sponsored by the 614 government and the other group did not. We accessed and collected the data from package `cobalt` in statistical programming software `R` (R Core Team 2021; Greifer 2021; LaLonde 1986; Dehejia and Wahba 1999).

Data cleaning

We first checked if there were missing values in the dataset and found no missing values. We then checked if each variable is of the appropriate type and found all variables were of the correct type. We re-ordered the levels of race so that "white" is the reference level. This is so that we can interpret effect sizes in terms of black and Hispanic workers relative to white workers in the analysis later.

Variable descriptions

The following describes each variable used in the analysis:

- **age**: age of the sampled individual, rounded to the nearest integer. It is of continuous type.
- **education**: years of education of the sampled individual, rounded to the nearest integer. It is of continuous type.
- **race**: race of the sampled individual - white, Hispanic or black. It is of categorical (three categories) type.
- **married**: indicates whether the sampled individual was married. It is of binary type (1 if married, 0 if not married).
- **degree**: indicates whether the sampled individual has a college degree. It is of binary type (1 if no degree, 0 if degree).
- **income**: annual income (earnings) in USD of the sampled individual after the worker training program. It is of continuous type. This is the outcome in our analysis.
- **treatment**: indicates whether the sampled individual completed the worker training program. It is of binary type (1 if treated in the National Supported Work Demonstration, 0 if from the Current Population Survey). This is the treatment/ exposure variable in our analysis.

A glimpse of the processed dataset used for the analysis (first 10 rows) is presented in Table 3 in the Appendix.

Variable summaries - numerical

Table 1 displays numerical summaries of annual income. The minimum income is 0, the maximum income is 60308, the mean income is 6793, the median income is 4759, the standard deviation is 7471, the interquartile range (IQR) is 10655. The range based on the maximum and the minimum is very big and the standard deviation and IQR are also very big, implying that the spread and variability of the distribution of annual income is very big.

Variable summaries - graphical

Table 1: Nummerical summaries of employee annual income.

Minimum	Maximum	Mean	Median	SD	IQR
0	60308	6793	4759	7471	10655

Figure 1 displays the distribution of annual income by whether the worker completed the training program. The income distribution for those who completed the program is simular to those who did not complete the program. We do see that both distributions are heavily right skewed because of large outliers (high income earners). These outliers and the skewness they induced contribute to the large spread and variability in the income distribution overall and for both groups.

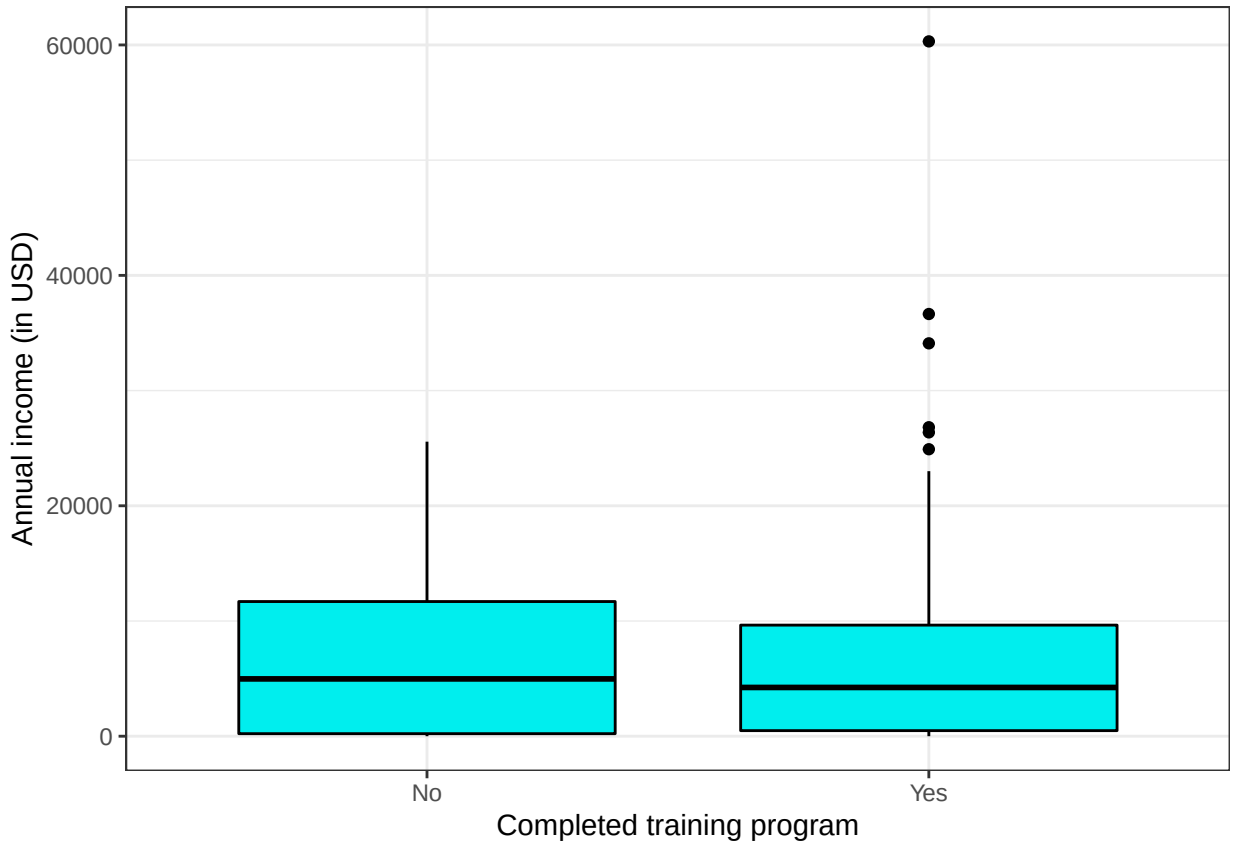

Figure 1: Distribution of annual income by training program.

All analyses in this report were conducted with `R version 4.0.3` (R Core Team 2021).

4 Model

To establish a causal effect of worker training program on worker's earnings, we had to ensure some variables that could affect the relationship of these variables were accounted for by balancing them in the treatment and control groups. We achieved this balance via propensity scores and propensity score matching.

Propensity scores are helpful when trying to draw causal conclusions from observational studies where the treatment was not randomly assigned. Propensity scores are the probabilities of subjects getting assigned to treatment (Imai and Van Dyk 2004). In a typical observational study, the propensity score is not known, because the treatments were not assigned by the experimenter (Imai and Van Dyk 2004). In such situation, the propensity scores are often estimated by the fitted values from a logistic regression on treatment using variables that the we wish to control for (Imai and Van Dyk 2004).

In an observational study, the treated and untreated groups are not directly comparable, because they may systematically differ in many variables, especially ones that correlate with the treatment and outcome of interest (Imai and Van Dyk 2004). The propensity score plays an crucial role in balancing the treatment and control groups to make them comparable (Imai and Van Dyk 2004; Dehejia and Wahba 2002). It has been demonstrated that treated and untreated subjects with the same propensity scores have identical distributions for all variables used to estimate the propensity score (Dehejia and Wahba 2002). This "balancing property" means that, if we account for the propensity score when we compare the groups, we have effectively transformed the observational study into a randomized block experiment, where "blocks" are groups of subjects with the same propensity scores (Dehejia and Wahba 2002). This in turn allows us to draw causal conclusions on the treatment on the outcome (Imai and Van Dyk 2004).

In our study, we wished to balance age, years of education, race, marital status and education between the group of subjects that went through the worker training program and the group that did not. This way, any difference in earnings from these two groups would be almost entirely attributable to the worker training program, not other variables. We used a logistic regression model to calculate the propensity scores with the aforementioned variables and fitted a linear regression model to study the relationship between earnings and worker training program using the same variables. To make sure our linear regression model was valid, we ran model diagnostic checks to make sure model assumptions were satisfied and checked if multicollinearity existed among the predictors using the variance inflation factor (VIF) of each predictor (Aiken et al. 2012; Tranmer and Elliot 2008). VIFs greater than 5 signifies multicollinearity existed in the model (Aiken et al. 2012; Tranmer and Elliot 2008).

First we show the equation of the logistic regression model for propensity score matching.

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_{\text{age}} + \beta_2 X_{\text{years of education}} + \beta_3 X_{\text{race: white}} + \beta_4 X_{\text{married - yes}} + \beta_5 X_{\text{degree - no}}$$

where

- p is the probability of getting assigned to go through the worker training program.
- The $X's$ denote the covariates in the model. They are what they are named in the above equations.
- β_0 is the model intercept and is not meaningful in our case since individual 0 years of age does not make sense in this analysis.
- β_1 is the change in log odds of getting assigned to the worker training program when age increases by one year.
- β_2 is the change in log odds of getting assigned to the worker training program when years of education increases by one year.

- β_3 is the difference in log odds of getting assigned to the worker training program for white vs. non-white workers.
- β_4 is the difference in log odds of getting assigned to the worker training program for married vs. non-married workers.
- β_5 is the difference in log odds of getting assigned to the worker training program for non-degree vs. degree workers.

Then we show the equation of the linear regression model for the outcome (earnings).

$$Y = \beta_0 + \beta_1 X_{\text{age}} + \beta_2 X_{\text{years of education}} + \beta_3 X_{\text{race: white}} + \beta_4 X_{\text{married - yes}} + \beta_5 X_{\text{degree - no}} + \beta_6 X_{\text{worker training program - yes}} + \epsilon$$

where

- Y is the worker's annual income in USD.
- The $X's$ denote the covariates in the model. They are what they are named in the above equations.
- β_0 is the model intercept and is not meaningful in our case since individual 0 years of age does not make sense in this analysis.
- β_1 is the change in average worker's income when age increases by one year.
- β_2 is the change in average worker's income when years of education increases by one year.
- β_3 is the difference in average worker's income for white vs. non-white workers.
- β_4 is the difference in average worker's income for married vs. non-married workers.
- β_5 is the difference in average worker's income for non-degree vs. degree workers.
- β_6 is the difference in average worker's income for those who completed the worker's training program vs. those who did not.
- ϵ is the random variation in income unexplained by the model.

5 Results

Table 2 displays results from the multiple linear regression model. Only years of education was statistically significant in predicting employee annual income. The rest of the variables, including the worker training program, were not statistically significant. Thus, we did not have sufficient evidence to reject our hypothesis that working training program has no impact on earnings. The results did make sense to us since years of education reflect the level knowledge and skill set a worker possesses and can bring to the role. And higher-paying roles tend to require higher level of education in most industries. The substantial edge higher level and more years of education can provide certainly trump any immediate advantage and benefits any short-term training program can provide. Propensity score matching matched 370 out of 614 individuals and the multiple linear regression model was fitted on these 370 matched individuals.

Table 2: Linear regression model results.

	Coefficient	Standard error	P-value
(Intercept)	333.364	3551.281	0.925
age	26.247	45.446	0.564
educ	501.148	228.240	0.029
raceblack	-1131.893	1023.866	0.270
racehispan	1016.327	1308.020	0.438
married	344.470	979.900	0.725
nodegree	-179.874	1121.446	0.873
treat	1109.558	804.995	0.169

In Figure 2 displays, from the linear regression model, a plot of standardized residuals vs. fitted values and a Normal QQ plot of standardized residuals. The standardized residuals vs. fitted values plot shows that the variability of the residuals increases slightly as fitted value increases, but nothing too serious. We observe some departure from the line on the Normal QQ plot, but the degree of departure from normality was not substantial. In summary, we conclude that the assumptions of the multiple linear regression model were sufficiently satisfied so that model results were reliable and not subject to a substantial amount of bias. The VIFs of all predictors were less than 5, indicating absence of multicollinearity. The VIFs are presented in Table 4 in the Appendix.

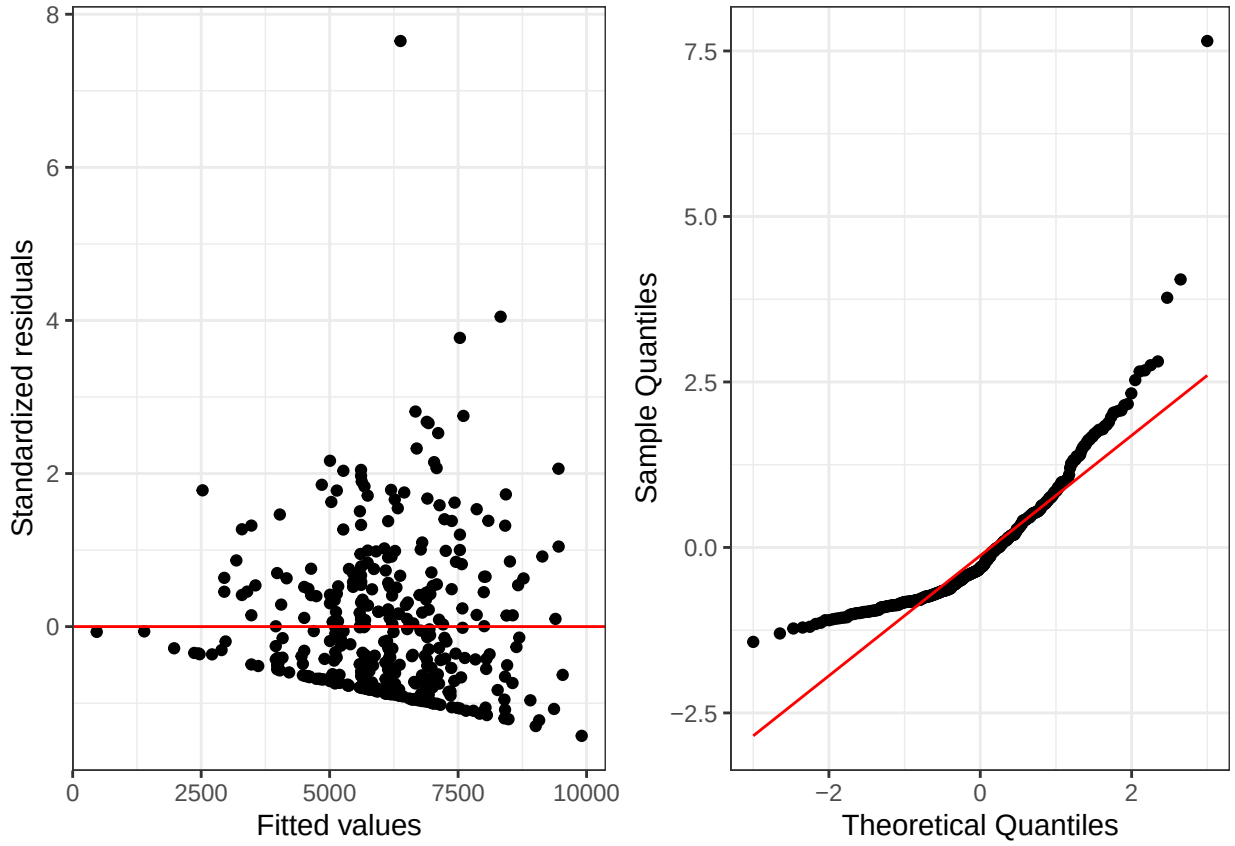

Figure 2: Plot of standardized residuals vs. fitted values (left) and Normal Q-Q plot of standardized residuals (right).

All analyses in this report were conducted with `R version 4.0.3` (R Core Team 2021).

6 Discussion

This report contains an analysis that examines the association between worker training program and worker's earnings, specifically it answers the research question of whether working training program has an impact on worker's earnings. We devised a hypothesis that working training program has no impact on worker earnings.

We found worker income program did not have a statistically significant relationship with worker's earnings. Thus, we did not have sufficient evidence from our data analysis and model results to reject the hypothesis that working training program. We conclude that completing a worker income program would not lead to higher earnings.

There were some limitations to our study and analysis. Through propensity score matching, observed variables in the dataset were balanced across treatment and control groups so these would not have any confounding effect on our results and conclusions, that is we could attribute differences in the outcomes entirely to differences in treatment received (completing or not completing the worker income program). But, our causality was not perfect in the sense that other hidden confounding variables not in the analysis dataset, that our propensity score matching did not consider, could have influenced our results and causal inference. Furthermore, propensity score matching only matched 370 out of 614 individuals, so we lost a significant amount of data and the information they brought.

Next steps would be collecting and combining more datasets so that we could account for more confounders and have a greater sample size after matching. We could also expand our analysis into looking at more than one worker income program to reduce variation in these programs and the impact of such variation on worker's earnings. This would in turn increase the generalizability of our results and conclusions.

7 References

Aiken, Leona S, Stephen G West, Steven C Pitts, Amanda N Baraldi, and Ingrid C Wurpts. 2012. "Multiple Linear Regression." *Handbook of Psychology, Second Edition* 2.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://CRAN.R-project.org/package=gridExtra.

Bartel, Ann P. 1994. "Productivity Gains from the Implementation of Employee Training Programs." *Industrial Relations: A Journal of Economy and Society* 33 (4): 411–25.

Dehejia, Rajeev H, and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs." *Journal of the American Statistical Association* 94 (448): 1053–62.

———. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84 (1): 151–61.

Deutsch, Steven. 1987. "Successful Worker Training Programs Help Ease Impact of Technology." *Monthly Lab. Rev.* 110: 14.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gelman, Andrew, and Yu-Sung Su. 2020. *Arm: Data Analysis Using Regression and Multilevel/Hierarchical Models*. https://CRAN.R-project.org/package=arm.

Greifer, Noah. 2021. *Cobalt: Covariate Balance Tables and Plots*. https://CRAN.R-project.org/package=cobalt.

Imai, Kosuke, and David A Van Dyk. 2004. "Causal Inference with General Treatment Regimes: Generalizing the Propensity Score." *Journal of the American Statistical Association* 99 (467): 854–66.

Kluve, Jochen, Friederike Rother, and Maria Laura Sanchez Puerta. 2012. "Training Programs for the Unemployed, Low-Income, and Low-Skilled Workers." *The Right Skills for the Job?*, 133.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*, 604–20.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2021. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Tranmer, Mark, and Mark Elliot. 2008. "Multiple Linear Regression." *The Cathie Marsh Centre for Census and Survey Research (CCSR)* 5 (5): 1–5.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain Francois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

8 Appendix

Glimpse of the analysis dataset:

Table 3: First 10 records of the dataset used in the analysis.

treatment	age	education	race	married	degree	income
0	50	0	white	1	1	220
0	19	3	white	1	1	0
0	32	4	white	1	1	0
0	47	3	white	1	1	6146
0	50	3	white	1	1	13976
0	25	5	white	1	1	14618
0	26	5	white	1	1	3700
0	39	5	white	1	1	18324
0	41	5	white	1	1	19451
0	19	12	white	1	0	1067

VIFs of the regression model:

Table 4: Variance inflation factors of variables in the multiple linear regression.

Variable	VIF
age	1.13
educ	2.02
race	1.28
married	1.12
nodegree	1.95
treat	1.19