Acta Technica *62* (2017), No. 7A, 1709–1718 © 2017 Institute of Thermomechanics CAS, v.v.i.

Heterogeneity of selection bias and training return—based on psm analysis of treatment effect

Yahui Liu[1], Hongzhi Cui[2]

Abstract. This paper adopts treatment effect and propensity score matching to separate the sample selection bias, correct the matching deviation, and estimate the average treatment effect of training on income. The study shows that participating in training could increase the income of sample individuals by 11.5% to 13.6%, indicating that training has the room for expansion, and it is necessary to make the policies to expand training coverage and encourage training participation. At the same time, there is a difference in the income effect of different groups, which is the heterogeneity of the training rate of return. The results of grouping treatment show that the income effect of training on rural population is higher than that on urban population. With the increase of educational level, the return rate of training decreases. The training has no significant income effect on the population with the educational degree of university or above. Training has no significant impact on the population aged 60 or above, but the return rate of training under the age of 60 increases with age group. From the perspective of the return rate of different groups, training should be targeted and focus on specific objects.

Key words. Training, Treatment effect, Propensity score matching, Selection bias.

1. Introduction

The research on the impact of training on income is usually the study of the return rate of training, and the consistent conclusion at home is that training has a significantly positive effect on income. Overseas research object for training return is mainly the on-the-job training of enterprise staff. Due to the differences of training individuals in such aspects as age, gender, level of education, the return rate of training may also have differences, namely training returns have heterogeneity. Bassanini (2006) found that training only has a significantly effect on the income increase of young and highly educated employees, but does not have great influence

[1] Graduate School of Chinese Academy of Social Sciences, Beijing, 102488, China

[2] Rural Development Institute Chinese Academy of Social Sciences, Beijing, 100732, China

on the individuals with higher age and lower education. However, the author thinks that the difference of return rate is due to a downward rigid salary, and the people with higher age and lower education have no chance to experience the falling wages, but are more frequently and directly laid-off, so as to be left out in the job market, lack of training. Budria & Pereira (2007) studied Portugal's training situation, finding that women, and the people with low level of education and long working years gain higher returns from training, and the ones that is older with low level of education that might get high return less attend training. Although the two articles work out the contrary results, they have a common focus: training returns is different for different groups, and older people, and people with low degree have low participation rate.

2. Data and research method

(1) data source

Data adopted in this paper are the survey data of the third phase of the Chinese women's social status by the all-china women's federation and the national bureau of statistics, as the standard point of investigation on December 1, 2010. In accordance with the regional development level, the stratified sampling is carried out, and the PPS sampling method is used to select sample to take interview survey on Chinese citizens reaching the age of 18 in 31 provinces except Hong Kong, Macao and Taiwan. The sampling unit in the first stage is county, district and county and municipal level (village, town, street in Beijing, Tianjin and Shanghai). The number of primary sampling unit in national sample is 460, and the number of primary sampling unit in in provincial independent sample is about 40. The second-stage sampling unit is village and neighborhood committee. Each primary sampling unit randomly selects 5 villages and neighborhood committees, and the structure of the village and neighborhood committee samples is determined according to the urbanization level. In the third stage, the sampling unit is households, and 15 households are randomly selected from each sample village and neighborhood committee. Finally, the interviewees of various individual questionnaires are determined by a specific random method in each household. The third-phase survey covers nine aspects: health, education, economy, social security, politics, marriage and family, lifestyle, legal interests and cognition, gender concept and attitude.

There are 26171 individual survey questionnaires in the third phase of the survey data. In this paper, based on the research question, the sample selection is conducted to eliminate income missing sample, students in reading, household people, retired people from labor market, people losing labor ability, as well as problematic data. Then the remaining 20699 samples are left, and the proportion to participate in training is 20.2%.

(2) Definition of variable

Dependent variable: labor income, including salary bonus, work allowance subsidy, operating income, etc., and excluding property income and transfer income.

Independent variable: for "whether you have participated in training in the last three years", the binary variable as the core independent variable of this paper is set; the covariates include the duration of education, seniority, skill and individual characteristic variables, as well as rural and regional virtual variables. In order to examine the heterogeneity of the training return rate of different groups, this paper groups the cultural level and age. (see table 1).

Table 1. Description and statistics of variables

	Group that attends training			Group that does not attend training		
variable name	sample size	mean value	Variance	sample size	mean value	Variance
Annual income logarithm	4176	9.81	0.89	16523	9.07	1.05
education period (year)	4170	11.06	3.16	16390	7.70	3.74
length of service (year)	4128	18.62	10.59	15895	24.37	12.20
quadratic component of length of service	4128	458.98	443.90	15895	742.97	616.21
Skill (having skill =1)	4176	0.55	0.50	16523	0.42	0.49
Times of job change	4176	1.17	1.57	16008	0.76	1.26
Male (male=1)	4176	0.55	0.50	16523	0.53	0.50
Party members and cadres (yes=1)	4176	0.48	0.50	16523	0.16	0.37
Rural (rural=1)	4176	0.30	0.46	16523	0.60	0.50
11 eastern provinces	4176	0.46	0.50	16523	0.39	0.49
8 central provinces	4176	0.28	0.45	16523	0.32	0.47
state-owned sector	4176	0.37	0.48	16523	0.13	0.33
Collective group	4176	0.07	0.26	16523	0.02	0.16
Private sector	4176	0.14	0.35	16523	0.12	0.33
Individual farming	4176	0.14	0.35	16523	0.47	0.50

(3) empirical model

Rubin (1974) proposed the “counterfactual framework”, in which dummy variable Di ={0,1} represents whether the individual participates in a project, 1 for participating, and 0 for not participating. Di is the processing variable. The participating group is the treatment group, and the group that does not participate is the control group. For a rational individual, when he expects the income from participating in a project is greater than the income that he does not participate in, namely $(y_{1i} - y_{0i}) > 0$, he will choose to participate in it.

To be specific, the equation of individual choice is as follows:

$$y_i = D_i y_{1i} + (1 - D_i)\, y_{0i} = y_{0i} + D_i(y_{1i} - y_{0i})\,.$$

Wherein, $(y_{1i} - y_{0i})$ is the treatment effect or causal effect of individual i participating in training. Assuming that whether the individual i attends the training has no effect on other individuals, treatment effect $(y_{1i} - y_{0i})$ is a random variable, and the expectation value is the average treatment effect that we pay close attention to: $ATE \equiv E(y_{1i} - y_{0i})$, in which ATE means the expected treatment effect of randomly selected individual from the overall, no matter whether the individual participates in the training or not.

For the individual that have participated in training, the average treatment effect should be shown by ATT, and the average processing effect of the individuals who did not attend the training should be shown by ATU:

$$ATT \equiv E(y_{1i} - y_{0i}|D_i = 1)\,,$$
$$ATU \equiv E(y_{1i} - y_{0i}|D_i = 0,\,.$$

Because individuals can be only in a state, to attend or not to attend, y_{1i} and y_{0i} cannot be observed at the same time. If the current income of participants and non-participants is simply compared, it will lead to selection bias, because the average difference between participants and non-participants is $E(y_{1i}|D_i = 1) - E(y_{0i}|D_i = 0)$, and the average difference can be decomposed into two parts:

$$E\left(y_{1i}|D_i = 1\right) - E\left(y_{0i}|D_i = 0\right) = E\left(y_{1i}|D_i = 1\right) - E\left(y_{0i}|D_i = 1\right) +$$
$$E(y_{0i}|D_i = 1) - E(y_{0i}|D_i = 0) = ATT + E(y_{0i}|D_i = 1) - E(y_{0i}|D_i = 0)\,.$$

Wherein $E\left(y_{0i}|D_i = 1\right) - E\left(y_{0i}|D_i = 0\right)$ *is selection bias of participants.* It can be also decomposed to get the selection bias of non-participants $E\left(y_{1i}|D_i = 1\right) - E\left(y_{1i}|D_i = 0\right)$.

Rosenbaum & Rubin (1983) proposed Propensity Score Matching to separate treatment effect and selection bias. Under the condition of meeting the negligible assumption, the people participating in training belong to the treatment group, and the non-participants belong to the control group. The basic idea is to find the individual j from the control group that is very similar to the observable variable x_i value of the individual i from the treatment group, namely $x_i \approx x_j$. Since individual i and individual j are very similar. The probability of them into the treatment group is similar. y_j can be taken as the estimator of the treatment group that is unable to observe y_{0i}, which can estimate the size of individual treatment effect. To deal with the matching for each individual in the treatment group, take similar measure for each individual in the control group, and make the concomitant variable used for the matching of treatment group and control group have no systematic difference or tiny difference, basically reaching the effect of similar random test, the sample matched is applied to calculate the average treatment effect.

$$\widehat{A\dot{T}T} = \frac{1}{N_1} \sum_{i:D_i=1} \left(y_i - \dot{y_{0i}}\right).$$

Wherein, $N_1 = \sum_i \left(D_i\right)$ is the quantity of individuals in the treatment group.

Logit regression is generally used to calculate propensity score and then matching is conducted according to the propensity score. There are many kinds of matching method. This paper uses one-to-four matching inside calipers, radius matching, kernel matching, and Markov matching to calculate respectively, as a robustness test of the model, and uses the deviation correction to correct estimator so as to correct bias occurring in the process of calibration matching.

3. Model Estimation and Results

(1) Test for the balance of data match

Propensity score matching requires the concomitant variable of On-Support) samples have no significant difference or their difference is very tiny; otherwise, the income difference between the two individuals might come from these explaining variables with significant differences. If the propensity score estimation is accurate, the mean of the matched treatment group should be closer to the mean of the control group. This process is called data balance. It can be seen from the balance test (table 2) that there are significant differences between the two sample groups of the training and the non-training before the matching. The standard deviation of all matched variables is less than 5%. The difference between two sample groups is significantly narrowed. In addition to the skills, the t test results of all concomitant variables do not refuse the null hypothesis of non-system difference between treatment group and control group. Joint inspection results show no significant difference between the two groups after the matching, which means reaching the balance of the data, and also indicates that concomitant variables used for the matching between the treatment group and control group have no system differences or the differences are quite small, basically similar to the random test result.

Table 2. Test for the matching results

Variable	Variable	Bias after matching	Rate of biased error decrease	t value before matching	t value after matching
education period	97.4%	-1.2%	98.8%	52.73***	-0.58
Skill	25.5%	4.3%	83.1%	14.47***	1.92*
length of service	-50.5%	-1.0%	98.0%	-27.53***	-0.50
square of length of service	-53.2%	-1.6%	97.0%	-27.69***	-0.88
Times of job change	28.3%	1.2%	95.9%	17.16***	0.46
Male	4.3%	0.3%	93.2%	2.43***	0.13
Party members and cadres	72.7%	-2.0%	97.2%	45.17***	-0.81
Rural	-63.1%	2.5%	96.0%	-35.10***	1.17
11 eastern provinces	15.0%	1.3%	91.1%	8.54***	0.60
8 central provinces	-10.8%	2.4%	78.0%	-6.03***	1.10
state-owned sector	59.5%	-1.6%	97.3%	37.85***	-0.62
Collective group	21.6%	2.4%	89.1%	14.33***	0.90
Private sector	6.1%	-0.9%	85.0%	3.55***	-0.40
Individual farming	-76.5%	1.4%	98.2%	-39.53***	0.78
joint survey	R2		LR chi2	p > chi2	
Before matching	0.217		4329.411	0.000	
After matching	0.001		13.73	0.470	

Note: ***, **, * respectively means it is significant in the level of 1%, 5% and 10%

(2) Treatment effect of training

In this paper, a variety of matching methods are used for robustness test. In addition to the sample selection bias, there may be a new bias in the matching process: first, there is uncertainty in the first stage when calculating the propensity score with Logit or Probit; second, inaccurate matching can cause bias. Therefore, this paper further makes the deviation correction estimate. Deviation correction estimator is based on Markov matching, using the method of regression, and take secondary matching in the control group and treatment group to get a robust standard error established under the condition of different variance. According to table 3, the effect coefficients are different under different matching methods, which are caused by the different matching methods. No matter which matching method is adopted, the effect of training on income is significant and stable. An increase of 11.5% to 13.6% in the income of individuals participating in the training could increase the income of non-participants by 11.9% to 21.2%. Compared with the Markov matching results without bias correction, the post-processing effect of deviation correction is reduced, but it is still significant at the 1% level.

Table 3. average treatment effect by using different matching method estimates

	Average treatment effect of treatment group		Average treatment effect of control group		Average treatment effect	
	ATT	standard deviation	ATU	standard deviation	ATE	standard deviation
one-to-four matching inside calipers	0.122***	0.0217	0.119***	0.0288	0.112***	0.0246
radius matching	0.115***	0.0204	0.130***	0.0260	0.126***	0.0224
kernel matching	0.129***	0.0205	0.172***	0.0237	0.163***	0.0207
Markov matching	0.136***	0.0157	0.212***	0.0308	0.186***	0.0258
Deviation correction matching estimation	0.118***	0.0159	0.152***	0.0208	0.143***	0.0256

Note: ***, **, * respectively means it is significant in the level of 1%, 5% and 10%

No matter which kind of matching method it is, ATU>ATE> ATT, which means the current average returns of training individuals is less than that of randomly selected individuals to participate in the training, and even less than those who did not attend the training (if they attend training), namely, once there is the opportunity, those non-participants could obtain higher returns from training. With the calculation for the sample selection bias E(y_0i |D_i=1)-E(y_0i |D_i=0), the participant's selection bias is 1.7%, 1.4%, 1.0%, 0.3% and 2.1%, which are positive. It shows that even if the participants did not attend, their income level higher is than those that actually did not attend, that is, the actual participants may already have more human capital and social capital advantage. However, the non-participant selection bias is negative, indicating that the expected return from the training of non-participant is higher than that of the participants.

The contrast of the statistics between the participants and the non-participants may explain the difference indirectly. Relative to the non-participant group, the participant group is five years younger than the non-participant group, its education period is three to four years more than the non-participant group, its proportion of party members and cadres reaches 48%, the state sector is 37%, and the urban population proportion is closer to 70%. Survey data show that in the non-participant sample, 22.3% lack information and opportunities, 9.5% have poor educational foundation. For the rural population, both are higher. Among non-participants, the proportion of farmers accounts for 60%. The population mainly engaging in agricultural accounts for 47%. For young individuals with high education, and party cadres, this part of the individuals originally has a comparative advantage. For the non-participants who are older with low education, less social capital or personal ability, they may have the economic situation constraints, may not have the education basis for training, or lack training opportunities. Especially in the countryside, when there is a free training opportunity provided by the government, in order to complete the training task and reduce the cost of mobilization, the village cadres will mainly encourage party members and cadres to participate in the training, and for those villagers who really lack skills, have the low degree of education, and are older may not the object firstly encouraged by the cadres.

(3) grouping treatment effect of training

Table 4. Grouping treatment effect

	K neighbor matching		radius matching		kernel matching	
Grouping	ATT	standard deviation deviation	ATT	standard deviation	ATT	standard deviation
cities and towns	0.125***	0.0231	0.122***	0.0279	0.124***	0.0224
cities and towns	0.146***	0.0364	0.134***	0.0332	0.153***	0.0423
Primary School or Below	0.203***	0.0808	0.159***	0.0802	0.200***	0.0779
junior high school	0.133***	0.0417	0.105***	0.0412	0.131***	0.0387
senior high school	0.110***	0.0297	0.110***	0.0292	0.106***	0.0249
University degree and above	0.043	0.0439	0.037	0.0315	0.043	0.0386
Below the age of 30	0.068**	0.0461	0.060**	0.0498	0.073*	0.0424
Aged from 30 to 40	0.109***	0.0382	0.108***	0.0361	0.115***	0.0353
Aged from 40 to 50	0.152***	0.0379	0.140***	0.0349	0.153***	0.0426
Aged from 50 to 60	0.179***	0.0592	0.191***	0.0548	0.234***	0.0759
Over the age of 60	0.129	0.1831	0.130	0.2161	0.198	0.1664

Note: ***, **, * respectively means it is significant in the level of 1%, 5% and 10%

Based on the urban and rural, educational level and age group, this paper puts forward the estimation of group treatment effect (see table 4). There is a significant difference in the average treatment effect between urban and rural areas. Training can increase the income of rural population by 13.4% to 15.3%, which is higher than the training rate of urban population. The rate of return on training decreases with the improvement of the educational level, and the influence on the income of the population with the education degree of university and above is not significant. In terms of age group, the training has no significant effect on the income effect of people aged 60 or above. For the population below the age of 60, the return rate of training increases with the age group. According to the results of group treatment effect, if the degree of education and age are regarded as a kind of human capital, the return rate of training seems to show a rule that it declines with the increase of human capital. If this is real, then from the angle of training investment returns, it should conform to the principles of efficiency. Under the limited training resources, it should select training object, and give priority to the group with weak human capital and social capital.

4. Conclusion and discussion

In this paper, the propensity score matching method (PSM) is adopted to analyze the sample deviation and estimate the average treatment effect of the training by empirical analysis of 20,699 samples in 31 provinces and municipalities nationwide.

Conclusion 1: training can increase the individual income of the sample by 11.5% to 13.6%. Those who do not attend training may receive a return of 11.9% to 21.2% through training. This indicates that there is scope for training, but the participation rate of training is still low, so it is necessary to expand training coverage and make policies to encourage training participation.

Conclusion 2: training has different effect on income of different groups. The income effect of rural population is higher than that of urban population. With the decrease in the return rate of training, the training has no significant effect on the population of the university or above. Training has no significant impact on the population aged 60 or above, but the return rate of training for the people under 60 increases with age group. From the perspective of different return rate of different groups, training should be targeted and focused on specific objects.

Since the current training population is generally younger, with higher level of education, and most of them are the urban population, with a high proportion of party members and cadres, this might reveal the current training object's choice is not optimal. In addition, in terms of the income gap, education and on-the-job training reflect the human capital is the main reason for widening difference in income (Goa Mengtao, Yao Yang, 2006). If human capital return varies between different groups, the difference between human capital returns will increase the income distribution effects of human capital, and, in turn, it has the effect of intensified unequal income distribution. In this case, the human capital investment must be more inclined to the poor, which can effectively narrow the income gap (Zhang Chewei, 2006). Through increasing the skills of the people with lowest income level, it can effectively curb long-term income inequality trend. Training as an important channel of human capital investment has more rapid and direct effect than normal education, and is the important opportunity to improve one's disadvantage in the labor market. In addition to expanding coverage and increasing the targeted groups, it also should pay attention to the quality of the training.

References

[1] Barron et. al.: *Job Matching and On-The-Job Training*, *Journal of Labor Economics 7* (1989), No. 1,

[2] Bartel, A. P: *Training, Wage Growth and Job Performance: Evidence from A Company Database*, National Bureau of Economic Research (1992) No.,w4027.

[3] A. Bassanini: *Training, Wages and Employment Security: An Empirical Analysis on European Data*, Applied Economics Letters *13* (2006), No. 8, 523–527.

[4] S. Budría, P. T. Pereira: *The Wage Effects of Training in Portugal: Differences across Skill Groups, Genders, Sectors and Training Types*, Applied Economics *39* (2007) No. 6, 787–807.

[5] L. M. Lynch: *Private Sector Training and Its Impact On The Earning of Young Workers*, National Bureau of Economic Research (1989) No. w2872.

[6] M. Jacob: *Job Training, Wage Growth and Labor Turnover*, National Bureau of Economic Research (1988), No. w2690.

[7] D. B. Rubin: *Estimating Causal Effects of Treatments in Randomized and Nonrandomized studies*, Journal of Educational Psychology *66* (1974), No. 5, 688–701.

[8] P. R. Rosenbaum, D. B. Rubin: *The Central Role of the Propensity Score in Observatianal Studies for Causal Effects*, Biometrika *70* (1983), No. 1, 41–55.

[9] C. Jobin, P. Duquette: *The Impact of Agricultural Extension on Farmer Nutrient Management Behavior in Chinese Rice Production: A Household-Level Analysis*[J]. Sustainability *6* (2014), No. 10, 6644–6665.

[10] M. D. Grady, D,Ėdwards , C. Pettus-Davis, et al.: *Does volunteering for sex offender treatment matter? Using propensity score analysis to understand the effects of volunteerism and treatment on recidivism*[J]. Sex Abuse *25* (2013), No. 4, 319–346.

[11] A. Berger, P. S. Mckinnon, K. Larson, et al.: *SB4 Propensity-Score Matching (PSM) to Control for Selection Bias in "Real-World" Treatment Comparisons: A Cautionary Tale Concerning Antibiotic Therapy for Infectious Disease*[J]. Value in Health, *15* (2012), No. 4, A4–A5.

[12] M. Caliendo, R. Hujer, S. L. Thomsen: *The employment effects of job-creation schemes in Germany: A microeconometric evaluation*[J]. Advances in Econometrics *21* (2008), No. 07, 381–428.

[13] S. Mcguinness, P. J. Sloane: *Labour market mismatch among UK graduates: An analysis using REFLEX data*[J]. Economics of Education Review *30* (2011), No. 1, 130–145.

[14] M. Li: *Using the Propensity Score Method to Estimate Causal Effects A Review and Practical Guide*[J]. Organizational Research Methods *16* (2013) No. 2, 188–226.

[15] D. Chun, Y. Chung, C. Woo, et al.: *Labor Union Effects on Innovation and Commercialization Productivity: An Integrated Propensity Score Matching and Two-Stage Data Envelopment Analysis*[J]. Sustainability *7* (2015), No. 5, 5120–5138.

[16] D. Gelo, E. Muchapondwa, S. F. Koch, et al.: *Decentralization, market integration and efficiency-equity trade-offs: Evidence from Joint Forest Management in Ethiopian villages*[J]. Journal of Forest Economics *22* (2016), 1–23.

[17] C. Carlini, M. D. Girolamo, A. Macinai, et al.: *Selective synthesis of isobutanol by means of the Guerbet reaction : Part 2. Reaction of methanol/ethanol and methanol/ethanol/ n -propanol mixtures over copper based/MeONa catalytic systems*[J]. Journal of Molecular Catalysis A Chemical *200* (2003) No. 1, 137–146.

Received May 7, 2017