The Effect of Job Training on Wages

Sangseok Lee, Sang-Jyh Lin, Julio Portella, Abdur Rehman, Vanessa Tang

10/4/2019

Introduction

In this study, we investigate two questions on the impact of job training: first, does job training increase real wages?; and second, does job training increase the likelihood of getting positive wages?

We fit a linear model to understand the impact of job training on wages and find out whether job training leads to higher wages. For the response variable, we use the change in wages between 1978 and 1974. This helps us quantify the impact of job training on real changes in wages rather than merely its impact on the post-training wage in 1978. With a model fit to predict the change in wage, a range for the effect of job training on change in wages can be determined. In addition, interactions with demographic variables are considered as are any other possible associations with wage changes.

To understand the impact of job training on the likelihood of positive wages or getting a job, we fit a logistic model. Our response variable is whether workers who completed the job training earned a positive or zero wage (in effect, no job) in 1978. We want to address whether workers who receive job training tend to be more likely to have jobs than those who do not receive job training.

Data

The data is taken from a previous study (Dehejia and Wahba, 1999) which is a subset of the data used in the landmark study that investigated the impact of job training on wages for the National Supported Work (NSW) program (Lalonde, 1986). The data includes 614 male subjects, 185 of whom where in treatment, and 429 in control. Unlike the original study, only male subjects are included and their demographic characteristics and job propsects are not randomly distributed between control and treatment conditions.

To determine whether workers who receive job training earn higher wages than workers who do not receive job training, a new variable called `delta` was calculated by subtracting wages in 1974 from those in 1978. This measures the difference in wages after completing job training.

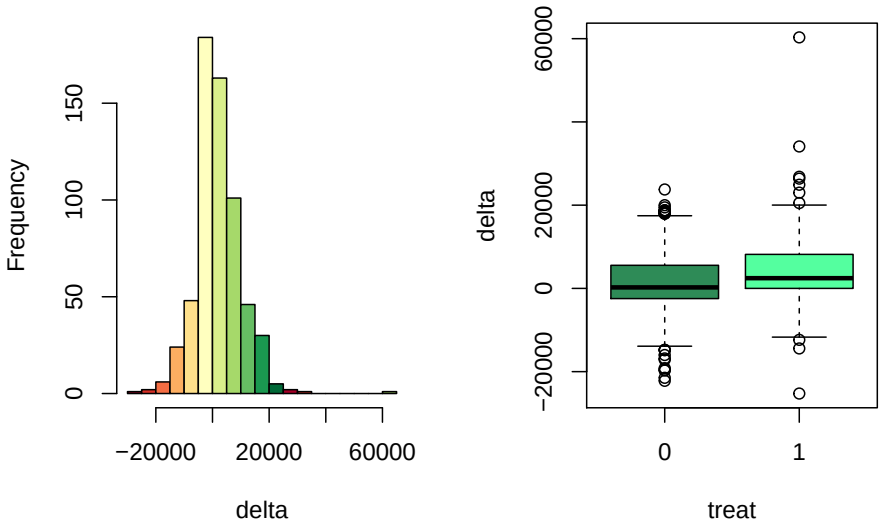
In addition to the 1974 and 1978 wages, the 1975 wage is also present in the dataset. The 1975 wage is difficult to interpret because we do not know whether workers were paid during job training, or how many months of job training they received annually. As a result, we decided ignore the 1975 wage for the purposes of this study.

Exploratory Data Analysis

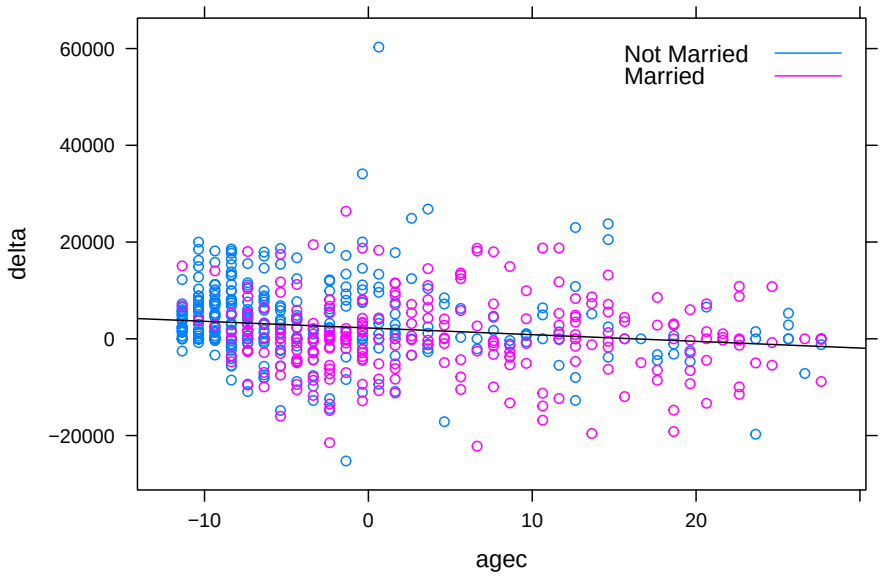
Linear Model

In our initial exploratory data analysis, the histogram for `delta`, the change in wages from 1974 to 1978, showed a relatively normal distribution. While there is one significant outlier, the overall distribution is relatively normal. Thus, no further transformations were needed on the response variable, `delta`.

Looking at the relationships between `delta` and the numeric variables such as `age` and `education`, we did not find any difference between each variable, and there are no unusual relationships to consider. Among the cateogrical variables, we noted that `treat` shows a slight difference in wage change between the treat and control conditions.

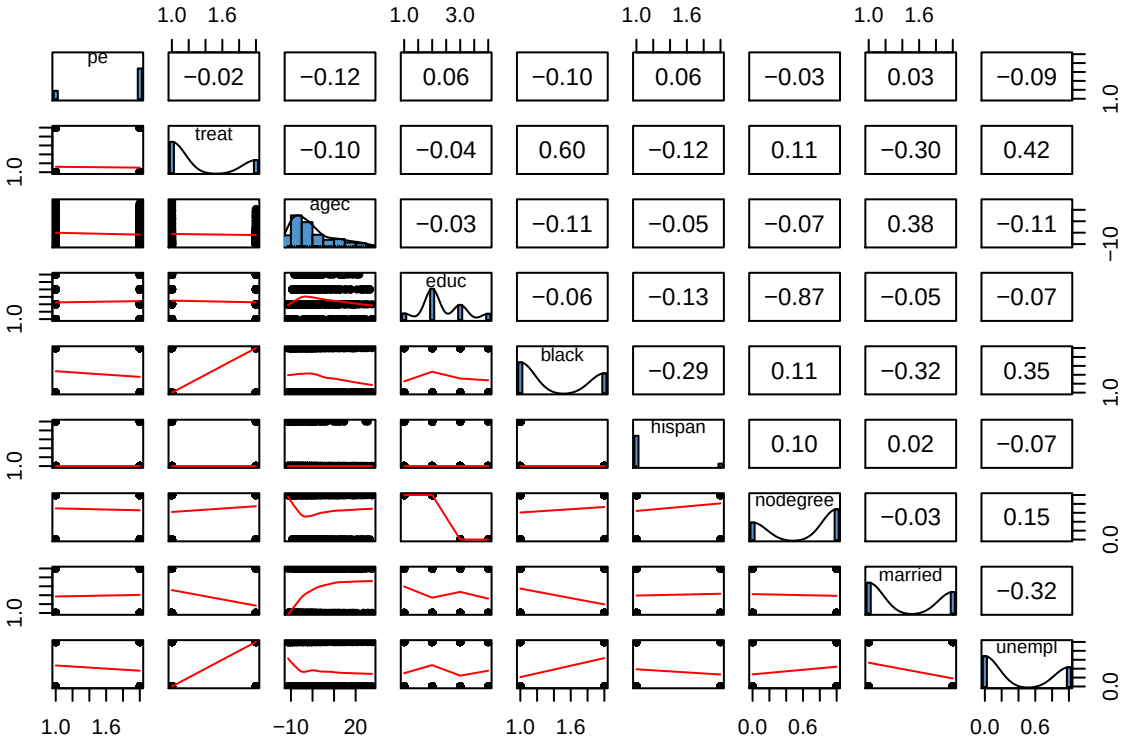


Based on the plot of age centered (`agec`) and change in wage (`delta`) where color corresponds to the two levels of `marriage`, it appears that there may be an interactive effect of `agec:marriage`. There is a higher density of married (blue) individuals on the left side where the wage change is higher. This indicates that we can expect an interaction term between `marriage` and `age` in the model. Our data exploration did not suggest any other strong associations between `delta` and any combination of the other predictors.



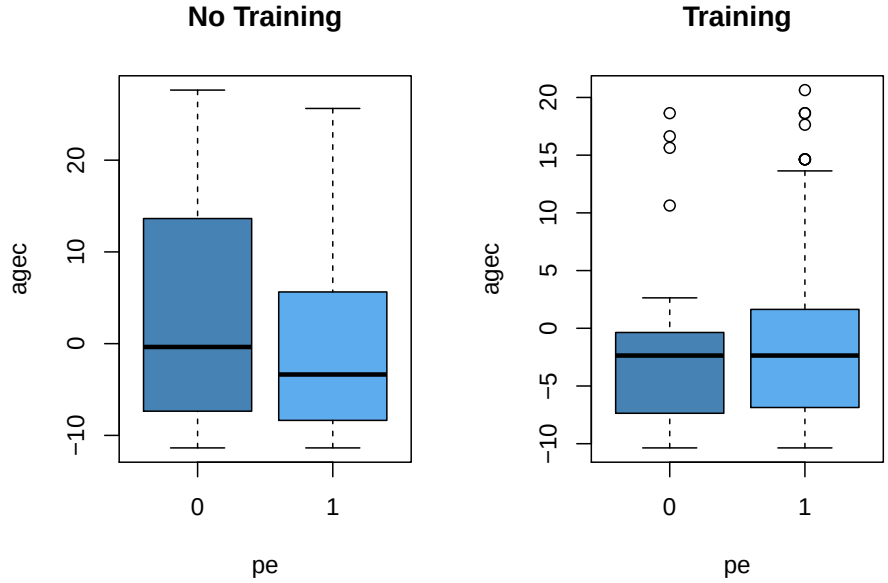
Logistic Model

For the logistic model, we converted `educ` into a categorical variable with 4 levels (See Appendix I). By looking at workers that did not have a positive wage before the program began, we created `unempl`. We felt this was important because unlike the data used in the original study (Lalonde, 1986), our dataset is not randomized between the control and test groups across the pre-experimental attributes of the workers (e.g. employment status). As with the linear model, we centered `age` into `agec` to make the intercept more meaningful.



The pairwise plots show that `pe` shares a weak relationship with all predictors. This relationship is positive for `educ`, `hispan`, and `married`, and negative for `treat`, `agec`, `black`, `nodegree` and `unempl`. We noted that `nodegree` is highly correlated with `educ`, but `educ` adds more detail to our model. In addition, both `nodegree` and `married` seem to share the weaknest relationship with `pe`. However, as with `treat`, we included them in the initial iterations of the model to understand whether this relationship is statistically significant at the 95% significance level.

It also appeared that there could be interactions between `treat` and `agec`, `treat` and `black`, and `treat` and `unempl`.



Model and Results

Linear Model

To find the best fitted model for our dataset, we manually fit different models and then tested our final specification against common selection techniques such as backward, forward, and stepwise selection. We also performed different ANOVA tests to compare models with different interactive terms, in which a single interaction term was added to one model and compared to the same model without the interaction term. We added interactive terms that were statistically significant at the 95% significance level, and relevant to our research question. We ended up with the folowing model:

`delta` ~ `married` + `treat` + `agec` + `treat:agec` + `married:agec`

In this model, change in real wage between 1978 and 1974 (`delta`) is predicted by `married`, `treat`, and `agec`; and the interaction terms `treat:agec` and `married:agec`. This model performed relatively well with an adjusted R^2 of 7.16% and an RSE of 7741. While these numbers are not very high, a model with all predictors and no interaction terms had an adjusted R^2 of 5.29% and a RSE of 7818. Through rigorous model selection and ANOVA testing, we were able to deem that this model provided the best fit when predicting the change in wage.

We included `re74` as a predictor in some of our model specifications, but ultimately decided to exclude it from our final model. There are three reasons we did this. First, we already had a proxy for 1974 wage through `delta`. Second, `re74` has many zero values that are open to interpretation. We do not know whether zero wages in 1974 include workers who were volunteering, or doing unpaid jobs. Third, there is also limited information on whether all workers supplied their pre-training income in 1974. Since it is not a randomized sample, we can speculate that including `re74` would likely bias the coefficients. For future iterations of our model, we recommend controling for previous experience with a variable that is more robust than `re74`.

Linear Regression Coefficient Estimates

term	estimate	std.error	statistic	p.value
(Intercept)	2122.2759	533.37519	3.978955	0.0000776
married1	-1833.7230	715.58871	-2.562538	0.0106307
treat1	2572.9746	727.97966	3.534404	0.0004397
agec	-201.9440	51.76956	-3.900826	0.0001066
treat1:agec	294.9896	89.63183	3.291126	0.0010556
married1:agec	129.4405	70.76420	1.829181	0.0678621

95% Confidence Intervals for Coefficient Estimates

Variable	2.5%	97.5%
Intercept	1074.79	3169.76
married1	-3239.05	-428.40
treat1	1143.31	4002.63
agec	-303.61	-100.28
treat1:agec	118.96	471.02
married1:agec	-9.53	268.41

Explanation

After we applied stepwise methods to build our model and analyzed the VIF results, we found that every predictor in our model is weakly correlated to one another. Hence, we do not need to worry about the variance inflation for our model. We found that there is no significant evidence that the wage difference is different among different demographic groups. This again is confirmed in running an ANOVA test comparing

two nested models where `black` and `hispan` are added individually to determine significance. Neither of these ANOVA tests for the two race variables yielded significant p-values at the 95% level. Therefore, there is no difference in the impact of job training on change in wages based on demographics.

There is significant evidence that workers who receive job training had an average wage change of $2,572.97 more than those who did not receive job training. Based on the 95% confidence interval, the likely range for wage change is from $1143.31 to $4002.63. This suggests that the average wage change between workers who recieve job training and those who did not ranges from $1143.31 to $4002.63.

Lastly, the effect of other predictors are listed from strongest to weakest predictive value based on their absolute t-values:

1. For every year older than the average age (27 years), wage change will decrease by an average of approximately $200.
2. Based on the interaction of `age` and `treatment`, receiving job training can increase wage change by approximately $120 for every year over the mean age of approximately 27 years.
3. Based on the coefficient estimate for `married`, being married decreases wage change by an average of approximately $1830.
4. Lastly, being married in addition to a 1 year increase in age from the mean age of 27 years can increase wage change by about $130 relative to unmarried workers.

After checking the residual plots for each of the predictors, no patterns were immediately evident. Hence, our model is appropriate based on residuals.

In order to determine if our final linear model is sufficient, model assumptions of linearity, independence, equal variance, normality were checked. The residuals vs fitted-value plot looks randomly scattered, and there's no discernable pattern, and the LOESS curve is flat and close to 0. Therefore, the assumption of linearity is met. In examining the Normal Q-Q plot, there's a little skew in both tails that deviates from the center line. However, this deviance is minimal and acceptable, and hence the normality assumption is met. Furthermore, there are not many points far above or below the fitted-value plot and no pattern is clearly visible. Therefore, the assumptions of equal variance and independence are met. As stated previously, after a thorough examination of residual plots, it was determined that the final linear model met the assumptions of equal variance as well as linearity, normality and independence (See Appendix III).

Logistic Model

After performing stepwise model selection methods to minimize the AIC, interaction terms were individually added to the stepwise model, and a deviance test was performed to determine signficance of one interaction term at a time. Interaction terms yielding a signficant p-value less than 0.05 that were relevant to our question were added to the model. The stepwise model was robust to manual forward and backward selection, with the coefficient estimates and the p-values staying relatively constant between different iterations.

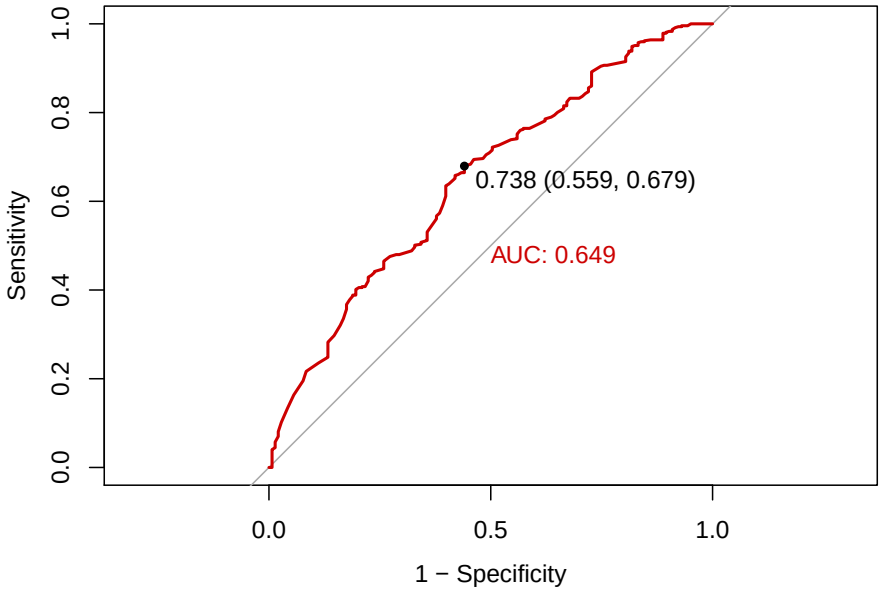
As a result of this model building process, we ended with a logistic model with `pe` as the response variable and `agec`, `black`, `unempl` and `treat` as the main predictors. We also included the interaction terms: `agec:treat`, `black:unempl`, and `black:treat` in our model. While we noticed `agec:treat` and `black:treat` in our data exploration, we were surprised by the statistical significance of `black:unempl` and decided to include it to modulate the effects of other race-related coefficients (`black`, `black:treat`).

`pe` ~ `agec` + `black` + `unempl` + `treat` + `black:unempl` + `agec:treat` + `black:treat`

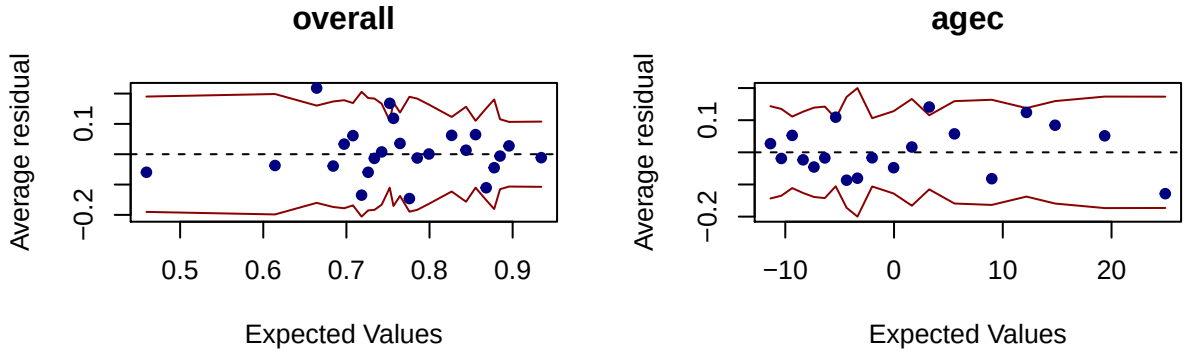
term	estimate	std.error	statistic	p.value
(Intercept)	1.6885725	0.1707052	9.891748	0.0000000
agec	-0.0445139	0.0105436	-4.221888	0.0000242
black1	-0.8356302	0.3230253	-2.586888	0.0096847
unempl	-1.0253003	0.3057583	-3.353303	0.0007985
treat1	1.7527874	0.7685536	2.280631	0.0225703
black1:unempl	0.9807539	0.4344519	2.257451	0.0239799

term	estimate	std.error	statistic	p.value
agec:treat1	0.0678339	0.0275298	2.464016	0.0137390
black1:treat1	-1.5691116	0.8276358	-1.895896	0.0579738

This model has an AIC of 646.38 and AUC of 0.649. At the optimal decision boundary of 0.738, this model has an accuracy, sensitivity, and specificity of 65.1%, 67.9%, and 59.9%, respectively.



Furthermore, exmaining the binned residuals showed no more than two points falling outside the 95% confidence intervals. While these outliers are not ideal, there are very few, and thus, the model is sufficient in terms of residuals.



Explanation

The model results show that `agec`, `black`, `unempl`, `treat`, `black:unempl`, `agec:treat`, and `black:treat` are statistically significant at the 95% significance level or higher. Since `treat` is highly significant, we can see that there is evidence that workers who receive job training have greater odds of obtaining positive wages after the program. In fact, job training appears to increase the odds of positive wages by 477% or about 4.8 times. This suggests that job training is successful in helping workers obtain jobs. A different specification of the model may help quantify the effect of this impact, and whether it has practical significance.

The 95% confidence intervals suggest that `treat` will have an effect between approximately 58% and 3640% on the odds of workers obtaining a positive wage. However, there is evidence that this effect differs by race. `black:treat` is marginally significant with a miltiplicative effect of -80%. This suggests that being a black

worker decreases the odds of a positive wage by about 80%. Taken together, job training may negatively affect the odds of a black worker obtaining positive wages (58% - 80% = -22%), which is a remarkable finding given that job training programs like the NSW are targeted to serving workers from a minority background. However, we should modulate this finding by adding that `hispan` was not statistically significant as a main effect, and did not have any interactive effect with `treat`. Hence, while the NSW program may have worked for hispanic workers, it seems to be failing black workers.

There was also two other statistically significant interactions, `black:unempl` and `agec:treat`. For `black:unempl`, the multiplicative effect is 167% which suggests that the job training program serves black workers who were unemployed before the program. This futher modulates the negative effect of `black:treat`, suggesting that the training program doesn't work for black workers who had jobs before the program but does work for those who were unemployed.

Similarly, the multiplicative effect of `agec:treat` is 7%, suggesting that older workers who obtain training perform better than younger workers who receive job training. Since `agec` has a multiplicative effect of -4%, the job training does not serve older workers in general.

`unempl` has the largest negative impact on the odds of obtaining positive wages with a multiplative effect of -64%. This suggests being unemployed before the program reduced a worker's odds of obtaining positive wages by about 64%. The job training does not appear to work for previously unemployed workers unless they are black.

Predictors and their Multiplicative Effect in Percentages

Variable	Exponentiated Coefficients in Percentages
Intercept	541.17%
agec	95.65%
black1	43.36%
umempl	35.87%
treat1	577.07%
black1:unempl	266.65%
agec:treat1	107.02%
black1:treat1	20.82%

95% Confidence Intervals of the Multiplicative Effect of Predictors in Percentages

Variable	2.5%	97.5%
(Intercept)	391.75	766.04
agec	93.67	97.63
black1	23.12	82.38
unempl	19.69	65.55
treat1	157.73	3741.79
black1:unempl	113.30	623.92
agec:treat1	101.55	113.20
black1:treat1	2.97	88.243

Conclusion

Based on our analysis of the effect of job training on wages, it can be determined that job training is beneficial. Those who received job training had larger (positive) changes in wages than those who did not receive job training. Likewise, those who received job training had higher odds of positive wages after the training than those who did not. Therefore, job training generally has positive effects on increases in wages and the odds of being employed.

While this study aimed to find the best fitted models to determine the effects of job training on wages, it has several shortcomings. First, as stated previously, our dataset is not randomized between the control and test groups across the pre-experimental attributes of the workers (e.g. gender, race, employment status), making it difficult to account for differences between the two tests groups. As a result, our linear and logistic models may not be considered "good" but are the best models based on certain metrics available in this dataset. Furthermore, we did not include wages in 1975 because it is unclear whether these wages are from workers getting paid for training, or how long they have been training, or what other factors could account for this wage. A future study should include months of job training during 1975 to help account for this, and this variable could be used to determine if being paid during job training has an effect on wages after training completion.

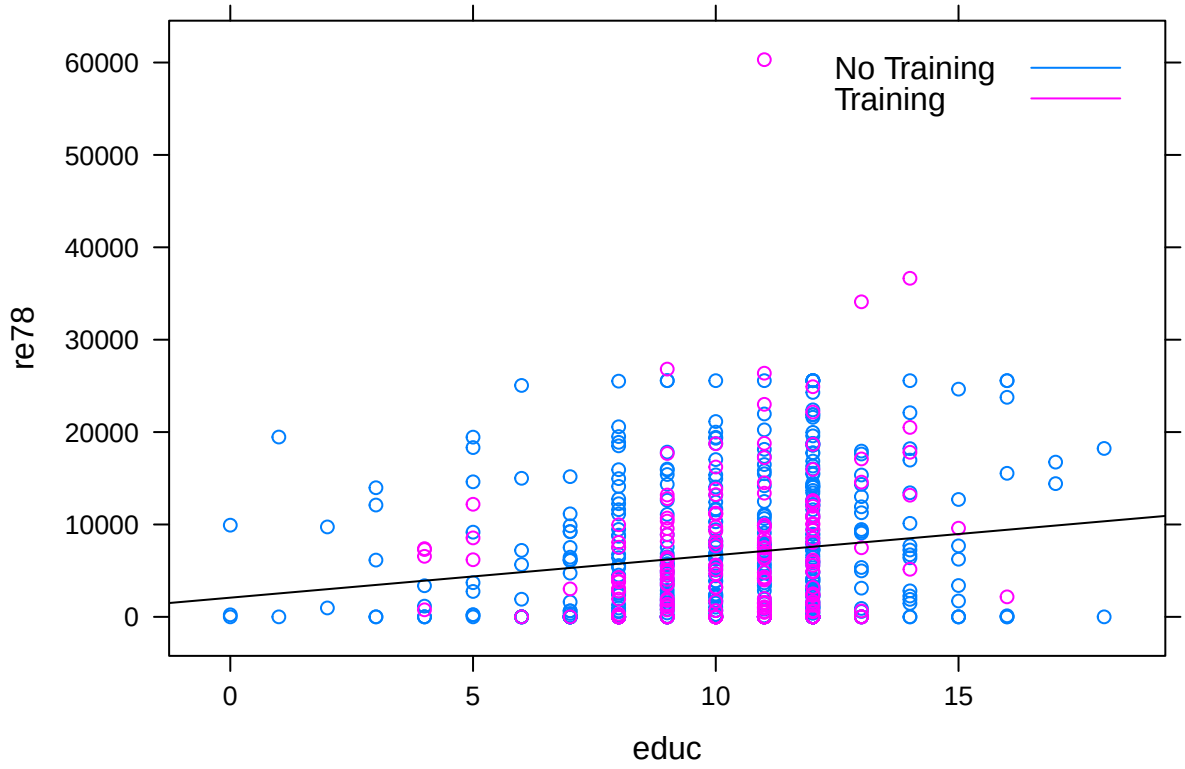
Despite these issues, this study successfully determines that job training leads to higher positive change in wages and increased odds of being employed.

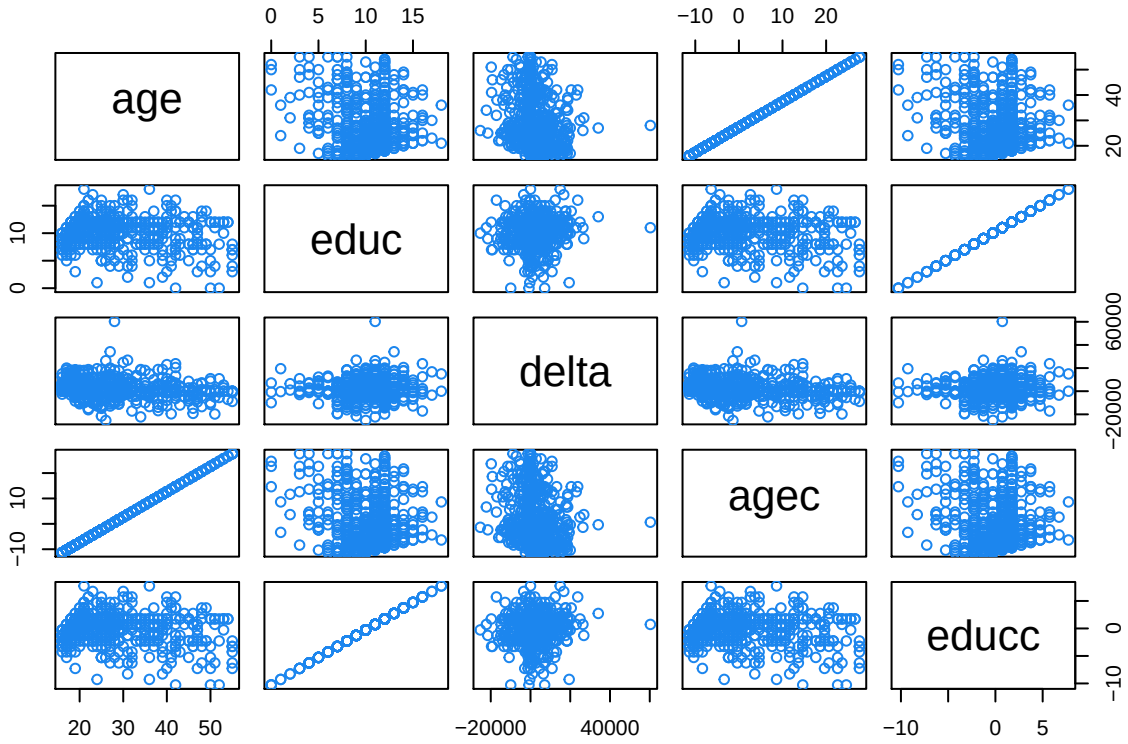
Appendix

I: Data Definitions

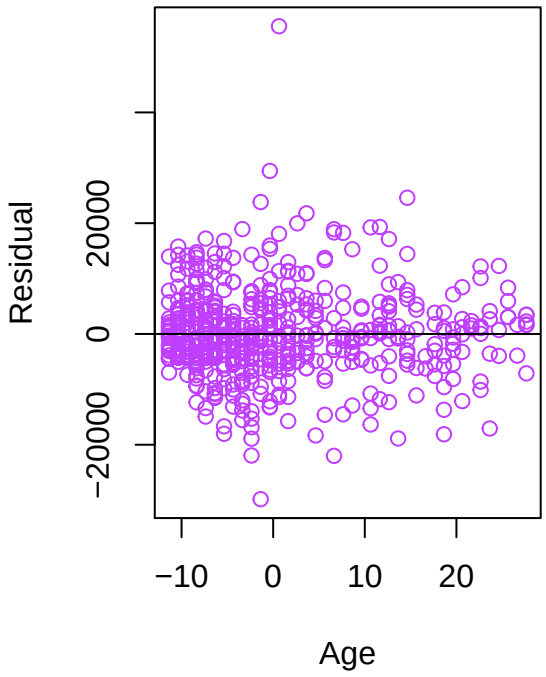
Variable	Meaning	Definition
treat	condition	1 if participant received job training, 0 if participant did not receive job training
age	age	age in years
educ	education	years of education
black	black	1 if race is black, 0 otherwise
hisp	hispanic	1 if Hispanic ethnicity, 0 otherwise
married	married	1 if married, 0 otherwise
nodegree	no degree	1 if participant dropped out of high school, 0 otherwise
re74	real annual earnings in 1974	-
re75	real annual earnings in 1975	-
re78	real annual earnings in 1978	-
unempl	unemployed	1 if re74 = 0, 0 otherwise
delta	difference in wage	re78 - re74
pe	positive earnings	1 if re78 > 0, 0 otherwise

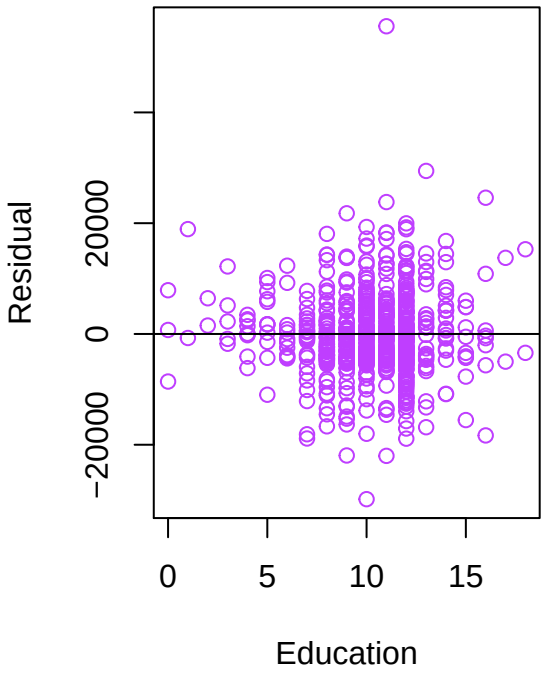
II: Exploratory Data Analysis: Linear Model

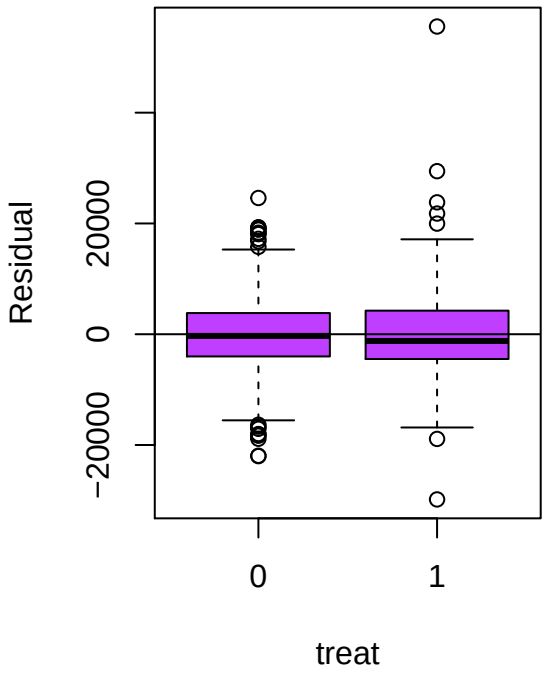

No Training
Training
re78
educ
60000
50000
40000
30000
20000
10000
0
0
5
10
15

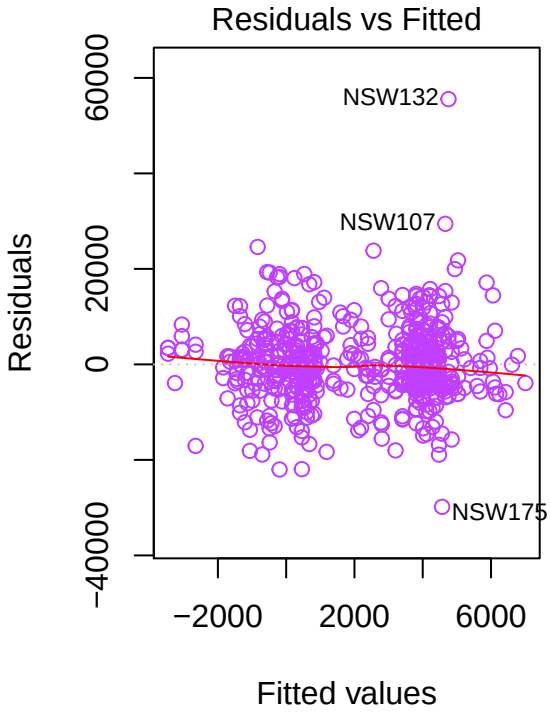

age
educ
delta
agec
educc

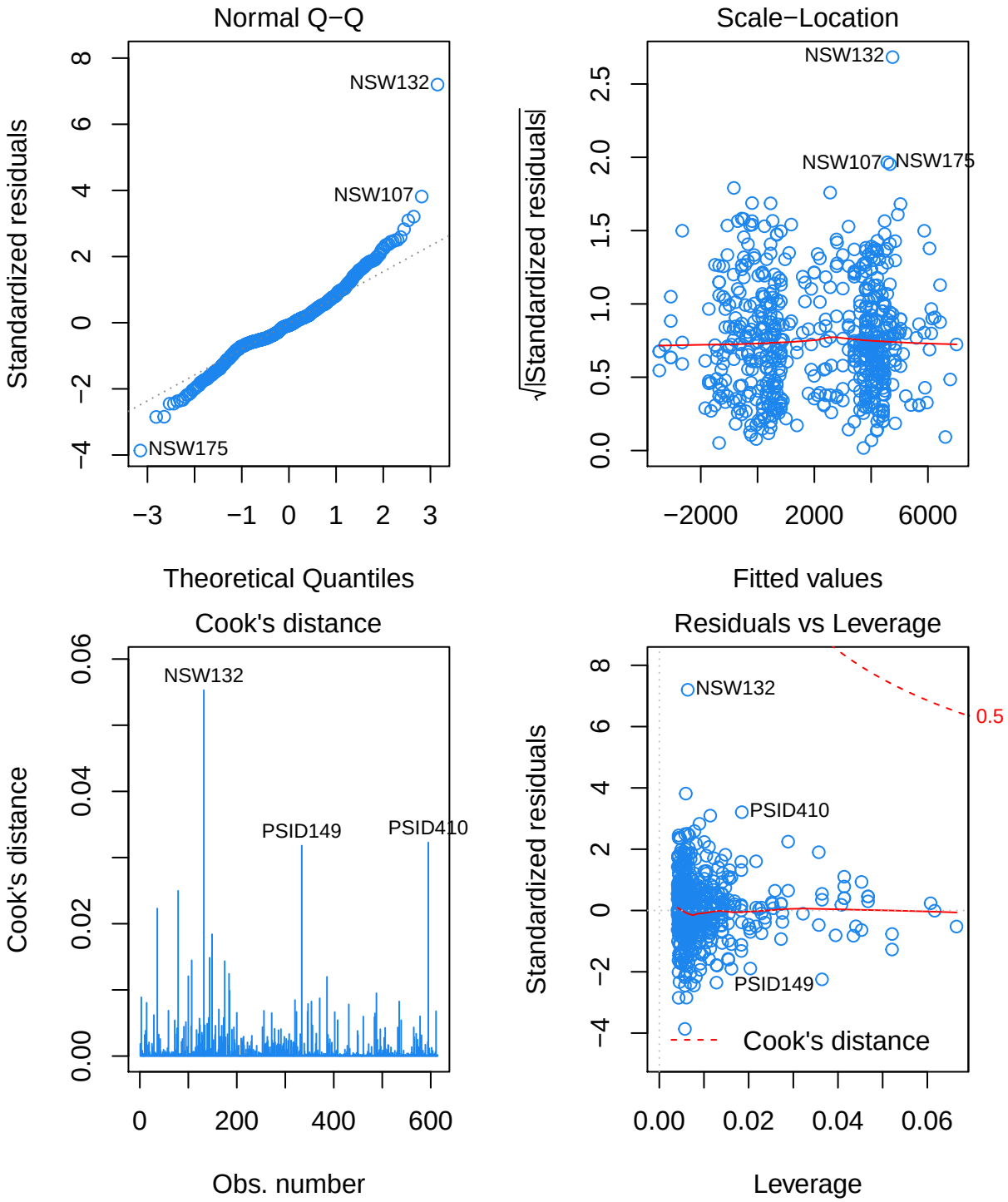
III: Model Selection and Testing: Linear Model










Normal Q−Q
Standardized residuals
Theoretical Quantiles
NSW132
NSW107
NSW175
Scale−Location
√|Standardized residuals|
Fitted values
Cook's distance
Obs. number
PSID149
PSID410
Residuals vs Leverage
Leverage
0.5

IV: Model Selection: Logistic Model

```
## Analysis of Deviance Table
##
## Model 1: pe ~ treat + agec + educ + black + hispan + unempl + treat *
##     agec + treat * educ + treat * black + treat * hispan + treat *
##     unempl + agec * educ + agec * black + agec * hispan + agec *
##     unempl + educ * black + educ * hispan + educ * unempl + black *
##     unempl + hispan * unempl
## Model 2: pe ~ agec + black + unempl + treat + black:unempl + agec:treat +
##     black:treat
##   Resid. Df Resid. Dev  Df Deviance
## 1       581     595.76
## 2       606     630.38 -25  -34.625
```

```
##           Reference
## Prediction   0   1
##          0  80 151
##          1  63 320
```

```
##  Accuracy
## 0.6514658
```

```
## Specificity
##   0.5594406
```