Modeling and Representation of Data - Team Project 1

John Owusu Duah, Michelle Van, Peining Yang, Sarwari Das, Satvik Kishore

Introduction

In this project, we analyse data derived from an experiment conducted by the National Supported Work (NSW) Demonstration in which researchers wanted to assess whether or not job training for disadvantaged workers had an effect on their wages. Specifically, we seek to understand if workers who receive job training tend to earn higher wages than workers who do not receive job training. We also investigate if this effect differs across demographic groups. Through the course of this analysis, we perform exploratory data analysis to identify the different socioeconomic characteristics that are associated with real annual earnings, and then model them using a linear regression. We find that treatment status is a significant predictor of income growth, and that this effect does not differ across demographic groups.

Data

We study a subset of the data originally organized in Lalonde, R. J. (1986), which is our main reference for this analysis. Here, the treatment group includes male participants for which 1974 earnings can be obtained, and the control group includes unemployed males whose income in 1975 was below the poverty level.

Refer to appendix for the full data dictionary. A summary of the continuous variables can be found in Table 1. The following transformations are applied to the data:

- Combined 'black' and 'Hispanic' into a 'race' variable, where 0 indicates 'other' races, 1 indicates black men and 2 indicates Hispanic men.
- Created a new variable 'growth', which is the difference between the real earnings in 1978 and 1974.
- Drop the variable re75; we see that participants were paid during the experiment, and hence the wages recorded in 1975 act as a confounding variable to our outcome (growth). To capture the unique effect of the treatment, we choose to drop re75.
- Center our age and education variables, for easier interpretation of the intercept.

We have 614 records in our final data, with 429 participants in the control group and 185 in the treatment group. We checked for missing data and found none. Large number of zero values are seen in each of the income columns, but given that the experiment was run on disadvantaged workers lacking job skills, we conclude that these are meaningful to the inferential questions asked and are not being treated like missing values. Outliers are also seen in the data: in re78 for example, the maximum value is more than 5 times the value of the third quartile. We make note of this, and choose to deal with it during our regression modelling.

Table 1: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	614	27.363	9.881	16	20	32	55
educ	614	10.269	2.628	0	9	12	18
re74	614	4,557.547	6,477.964	0	0	7,888.5	35,040
re75	614	2,184.938	3,295.679	0	0	3,249.0	25,142
re78	614	6,792.834	7,470.731	0.000	238.283	10,893.590	60,307.930
growth	614	2,235.288	8,033.464	−25,256.800	−1,240.175	6,459.640	60,307.930

Exploratory Data Analysis

To quantify the effect of a treatment in a randomized experiment, ideally, the treatment and control groups should be balanced. In our data, we see that this is not the case. On running a t-test on baseline income levels (re74) across treat, we see that a significant difference ($p<0.001$) exists across means: for the control group, baseline income is 5619.24 dollars on average, and for the treatment group, it is 2095.57 dollars on average. This provides justification for using growth in income (re78-re74) instead of final income as our outcome variable. Fig 1 shows the underlying distribution of our response variable, growth. It is fairly normal, so we don't feel the need to perform any transformations on it. A significant difference ($p<0.001$) also exists in mean age across the two groups (Fig 2). Average age for the control group is 28.03 years, while it is 25.82 years for the treatment group. As age can impact income growth, we take this as a limitation in our analysis. Large differences also exist in the proportion of married people across treatment groups: in the treatment group, only 18.91% of the participants were married, compared to the 51.28% of people in the control group. For education, 70.91% of the treatment group were degree holders, compared to 59.67% of the control group.

Figure 1: Distribution of Earnings Growth

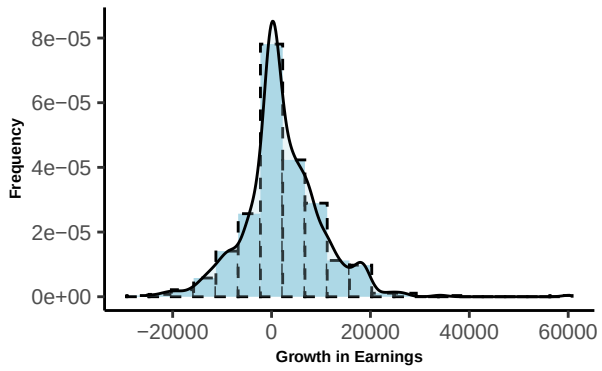
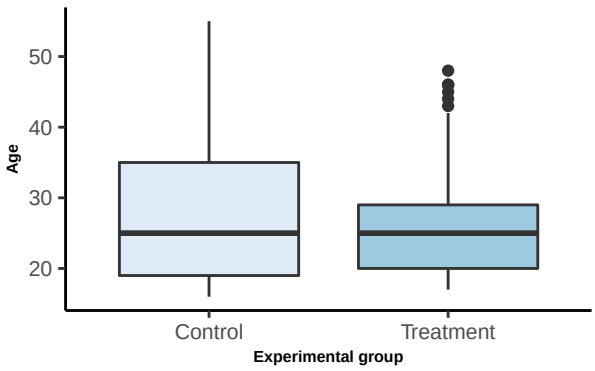


Figure 2: Age across Treatment groups



Before the modelling process, we perform an exploratory analysis to understand the plausible relationships between the predictors.Comparing income growth across predictors, boxplots show us that median growth was higher for participants in the treatment group, compared to the control group. (Fig 3) However, on performing a t-test, we see that difference in means were not significant. Median growth was also higher for unmarried people, compared to married people. No significant relationships were seen for degree holders and years of education. Across races, it was seen that median growth was highest for hispanic people, followed by black people and other races. Across age, a decline was noticed in earnings growth as a participant ages.

We are further interested in knowing whether the treatment effect varies across any demographic groups. On exploring interaction effects across our variables, we notice that trend of earnings growth against age changes across treatment groups: in the control group, earnings growth reduces as people age, while in the treatment group older people see more income growth. Trend changes are also noticed in the interaction between age and married: married people in the treatment group show lower median income growth than unmarried people, but the trend reverses in the control group. We also see some potential interactions for treat with race, and treat with degreeholders.

Figure 3: Growth across Treatment groups

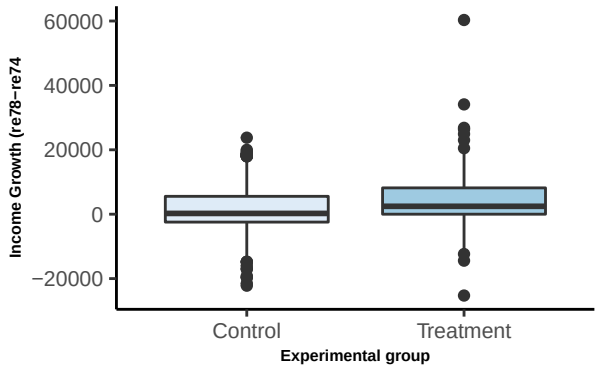
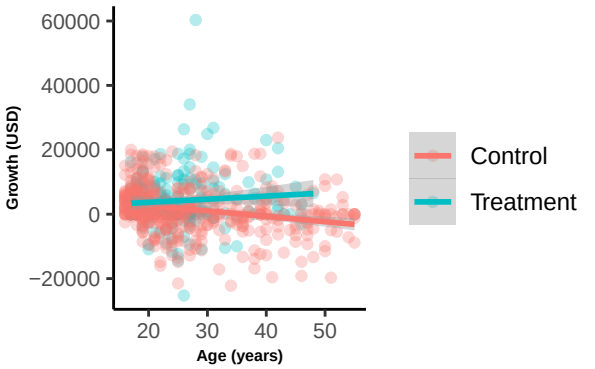


Figure 4: Change of growth with age across treat



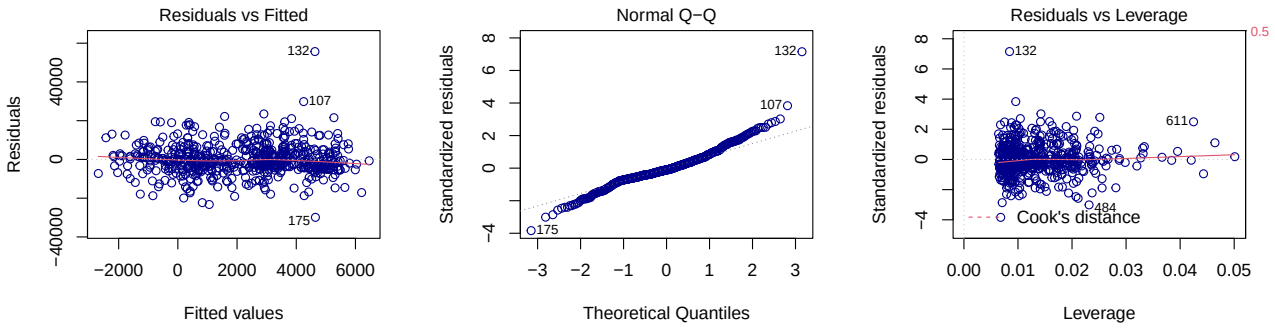
Model Building: Baseline (Main effects)

To begin, we construct a simple regression model with income growth as the response variable, and the main effects of all variables as our response variables. Age and education are centered.

$$Growth \sim \beta_0 + \beta_1\ treat + \beta_2\ age + \beta_3\ married + \beta_4\ educ + \beta_5\ race + \beta_6\ nodegree + \epsilon$$

A summary for the baseline model can be found in the Appendix. Our main takeaways were: (1) whether or not the participant was in the treatment group was a highly significant predictor ($p<0.001$) of their income growth (2) age and married are also significant predictors of income growth (3) the model explains about 5% of the variability in income growth.

On plotting the continuous variables 'age' and 'educ' against growth, we see somewhat linear trends. Transformations like log- transformations and polynomial forms do not improve the residual plots, so we decide to go with the variables as they are. On checking the residual plots below, we see that points in the Residuals vs Fitted plot are random, but look clustered in the center. They form a somewhat equal band about the axis, so we conclude that both independence and constant variance are not violated, but can be improved. In the Q-Q plot, points are on the 45 degree line, but we see deviations on both ends. We carried out several iterations of transformations of the variables to improve the condition of normality in the error term but no noticeable improvement was found. Since the points do not sharply deviate from the 45 degree line, we conclude that our model does not violate the normality assumption. On checking the Scale-Location and standardized residuals vs leverage plot, we see that we have outliers, but they are not influential. Hence we do not decide to remove them. To improve our model, we move on to the model selection process.



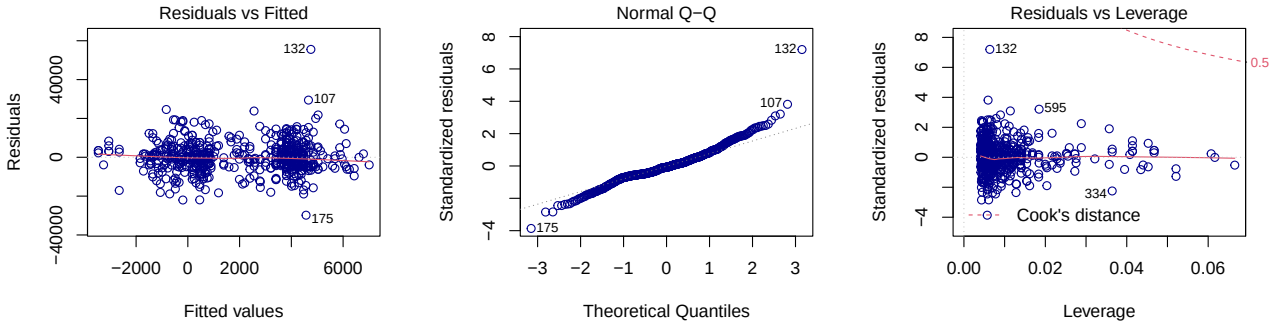
Model Selection: Final (Forward Selection with AIC)

For our model selection we construct a null model having just treat as the response variable, and a full model having all main effects as well as all possible interaction terms for treat, age and race and education. We did not want to be penalized for using a larger number of variables (since we observed interactions in EDA), so we use AIC with forward selection to get the final model given below. A nested F-test justifies the addition of these interactions ($p<0.05$), so we proceed with this model.

$$Growth \sim \beta_0 + \beta_1\ treat + \beta_2\ age + \beta_3\ married + \beta_4\ married : age + \beta_5\ treat : age + \epsilon; \epsilon \sim N(0, \sigma^2)$$

We assess linearity by plotting 'age' and 'educ' against growth and see a somewhat linear trend. Thus the linearity assumption is met. On checking the residual plots, we see that points in the Residuals vs Fitted plot are random, but look clustered in two groups. Since the independence assumption requires independence in the Y-axis and does not regard patterns on the X axis, we conclude that the independence assumption is not violated. However, the presence of the clusters indicate the presence of an omitted variable, which we count as a limitation in our analysis. The points seem to form an equal band around the axis, so we conclude that the constant variance assumption is met. Checking the Q-Q plot, we see that more points are on the 45 degree line than the baseline model, although deviations still exist. We conclude that normality of residuals have improved, and the assumption has been met.

On checking the Scale-Location and standardized residuals vs leverage plots, we see outliers (especially point 132). However, these are not influential so we choose to not remove them. Finally, on checking for multicollinearity, we see that the highest Variance Inflation Factor (VIF) was 1.23, which is substantially less than the threshold for concern.



Model Interpretation

Table 2 gives a summary for the model. According to our model, on controlling for whether the participant is married, when age is at the baseline (27 years, since it is centered), on average a person in the treatment group would see an increase in income that is \$2572.97 ($p<.001$) greater than someone in the control group. Age is also a highly significant predictor of income growth: for someone in the control group, controlling for all else, as age increases from the baseline by a year, income growth decreases by \$201.94 ($p<0.001$) on average. For someone in the control group, when age is at baseline, a married person would see income growth that is \$1833.72 ($p<0.01$) lower than someone in the treatment group on average. Finally, on average, unmarried 27 year olds in the control group would see an income growth of \$2122.28. Further, as age increases by a year, a person in the treatment group would see an additional increase in income growth by \$294.99 ($p<0.001$) on average. Further, for an increase in age by a year, a married person would see an additional increase in income growth by \$129.44 ($p<0.05$) on average. All the predictor variables in the final model have a significant effect on real annual earnings. This model explains 7% of the variability in income growth from 1974 to 1978.

Table 2: Linear Regression Model Output

	Estimate	Std. Error	t-value	p-value	95% CI
(Intercept)	2122.276	533.375	3.979	0	(1074.795, 3169.757)
treat1	2572.975	727.98	3.534	0	(1143.315, 4002.634)
age_c	-201.944	51.77	-3.901	0	(-303.613, -100.275)
married1	-1833.723	715.589	-2.563	0.011	(-3239.049, -428.397)
treat1:age_c	294.99	89.632	3.291	0.001	(118.964, 471.015)
age_c:married1	129.441	70.764	1.829	0.068	(-9.531, 268.412)

Note:

R-Squared: 0.079. Adj. R-Squared: 0.072. p-value: 1.227e-09

Limitations

The goal of the analysis is to determine the associative effect of training on earnings of workers. However, the control group includes unemployed participants in the survey. Including these participants introduces noise and obfuscates the analysis. Ideally, to achieve the goal of the assignment, data used should be exclusive to employed workers. We considered removing participants who were unemployed but stopped when we discovered that unemployed workers constituted 26% of the control group.

During the EDA, we discovered a substantial disparity in key sample statistics between the control group and the treatment group. The treatment group has a mean age of 26 years old and the control has a mean age of 28 years old. Also, the treatment group has mean annual earnings of $2,096 in 1974 while the control group has mean annual earnings of $5,619 in 1974. Again, ideally, to determine the effect of training on earning of workers, it is imperative that the two groups be as closely balanced as possible, practically.

With an adjusted coefficient of determination of 7%, our final model can accurately explain 7% of the variability in the growth of annual earnings of participants with the predictor variables. We can infer from this that other predictor variables outside the scope of what is available in the dataset, can explain the growth of annual earnings better.

Conclusion

Question 1: There is statistically significant evidence that training has positive relationship with the growth of income from 1974 to 1978, accounting for other effects and interaction between job training and age. The difference in real annual earnings from 1974 to 1978 increases by $2,254.80 for participants who were trained compared to participants who were not trained, keeping all other predictor variables constant.

Question 2: After computing"two-sided" confidence intervals for the effect of training, accounting for all other predictors, we are 95% confident that the difference in real annual earnings from 1974 to 1978 increases by an amount between $953.20 and $3,556.41 for participants who were trained compared to participants who were not trained.

Question 3: After carrying out an F-test to determine if the effect of training on earnings differs by demographic groups, we had a p-value of 0.447. In a regression model, it was not a significant predictor. This means there is no statistically significant evidence that the effect of training on earnings differs by demographic groups.

Question 4: Apart from the training status of participants, the age, real annual earnings in 1975 and interactions between training status and age have a significant effect on the growth in real annual earnings from 1974 to 1978. The predictor variable that had the most effect on the growth in real annual earnings is the age of participants. Keeping all other predictor variables constant, the growth in real annual earnings decreased by $149.52 for every year older that a participant gets. Also, keeping all other predictor variables constant, the growth in real annual earnings decreased by $0.41 for every dollar increase in real annual earnings for 1975.

Appendix

1. Data dictionary

Variable	Description
treat	*1 if participant received training, 0 if participant did not*
age	*age in years*
educ	*years of education*
black	*1 if race is black, 0 otherwisee*
hisp	*if Hispanic ethnicity, 0 otherwise*
married	*1 if married, 0 otherwise*
nodegree	*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*

2. Summary for baseline model

Table 4: Linear Regression Model Output

	Estimate	Std. Error	t-value	p-value	95% CI
(Intercept)	1942.991	839.735	2.314	0.021	(293.848, 3592.135)
treat1	2369.293	876.025	2.705	0.007	(648.88, 4089.706)
age_c	-85.086	35.437	-2.401	0.017	(-154.68, -15.492)
married1	-1895.452	733.679	-2.583	0.01	(-3336.313, -454.591)
educ_c	154.001	177.426	0.868	0.386	(-194.445, 502.446)
race1	-512.706	862.26	-0.595	0.552	(-2206.087, 1180.676)
race2	692.004	1056.855	0.655	0.513	(-1383.539, 2767.546)
nodegree1	773.267	952.401	0.812	0.417	(-1097.14, 2643.674)

Note:

R-Squared: 0.064. Adj. R-Squared: 0.053. p-value: 1.227e-09