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. (Add Findings)

Data Pre-processing

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.

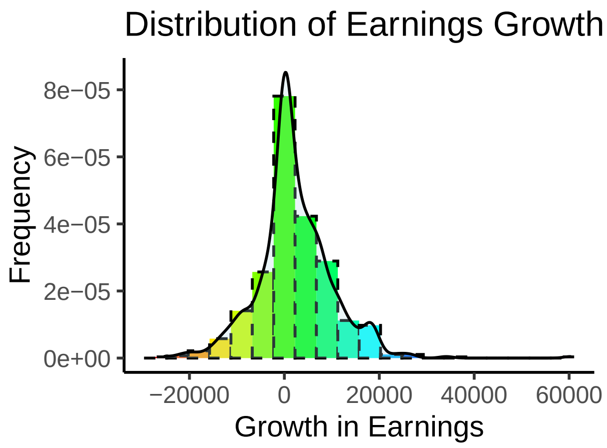
Table 1: Summary Statistics

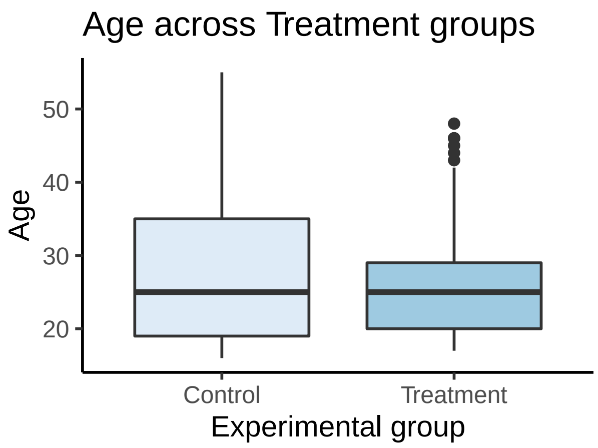
Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
age	614	27.363	9.881	16	20	32	55
educ	614	10.269	2.628	0	9	12	18
re74	614	4,557.547	6,477.964	0	0	7,888.5	35,040
re75	614	2,184.938	3,295.679	0	0	3,249.0	25,142
re78	614	6,792.834	7,470.731	0.000	238.283	10,893.590	60,307.930
growth	614	2,235.288	8,033.464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

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 (details in Appendix), 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.

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 X 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 X). 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.





Before the modelling process, we perform an exploratory analysis to understand the plausible relationships between the predictors (All plots in Appendix).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 X) 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.

Need to fit 2 EDA plots here: plot from our ppt for age vs treat + another

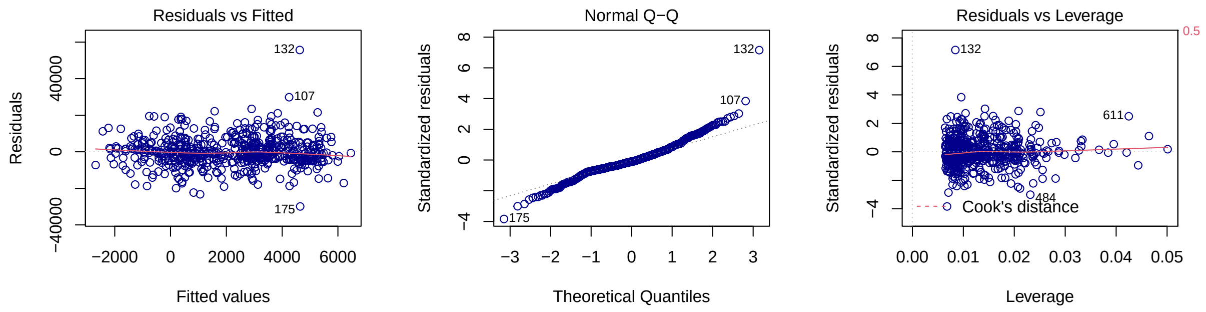
Model Building: Baseline (Main effects)

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. but 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, 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.

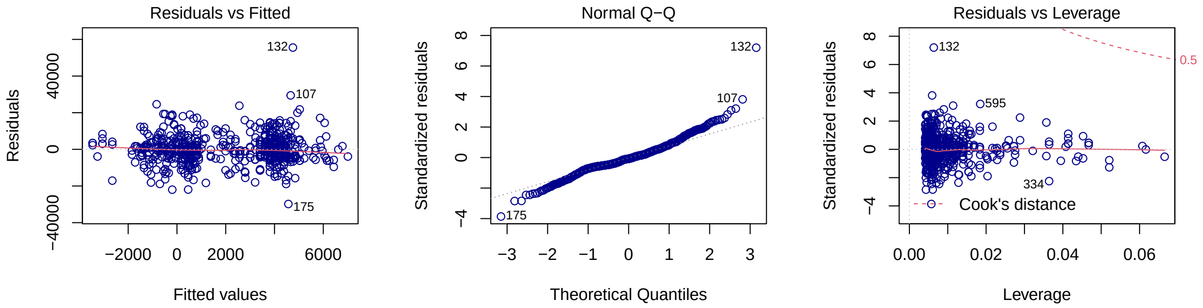


Model Selection: Final (Forward Model Selection with AIC)

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. Using AIC (**GIVE REASON WHY**), forward selection gives us the final model given below.

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

Table X gives a summary for the model. We assess linearity by plotting 'age' and 'educ' against growth and see a 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: Results

	Dependent variable:
	growth
treat1	2,572.97*** (727.98)
age_c	−201.94*** (51.77)
married1	−1,833.72** (715.59)
treat1:age_c	294.99*** (89.63)
age_c:married1	129.44* (70.76)
Constant	2,122.28*** (533.38)
Observations	614
R^2	0.08
Adjusted R^2	0.07
Residual Std. Error	7,740.62 (df = 608)
F Statistic	10.45*** (df = 5; 608)
Note:	*p<0.1; **p<0.05; ***p<0.01