The Impact of Education on Income

Andy Shi

Diana Zhu

Stat 186 Final Project

May 5, 2015

1 Introduction

The relationship between education and income has long been a topic of interest among economists and education has been widely perceived as an effective way of reducing income inequality. Our paper seeks to investigate their relationship using a cross-sectional dataset. Literature research shows that both theoretical and empirical studies have suggested ambiguous correlation between the two. For instance, the human capital model of income distribution from the works of Schultz and Becker and Chiswick indicates that the level and the distribution of schooling across the population largely determines the distribution of income [1, 2]. Although the model shows a clear positive relation between educational and income inequality, the impact of average schooling on income distribution is unclear, which may be positive or negative depending on the rates of return to education. Earlier empirical work also shows a close correlation between income and education in developed countries. Becker and Chiswick show, for example, shows that income inequality is positively correlated with inequality in schooling and negatively correlated with the average level of schooling across states in the United States [2]. Chiswick also suggests that earnings inequality increases with educational inequality using cross-sectional data from nine countries [3]. Subsequent studies likewise found a negative correlation between a higher level of schooling and income inequality, although Ram finds that mean schooling and schooling inequality have no statistically significant effects on income inequality.

2 Question of Interest

We are interested in exploring the connection between education and income. Specifically, we will investigate the causal relationship between these two variables using the Rubin Causal Model. We hypothesize that there is a significant positive association between education and income.

3 Methodology

3.1 Dataset

We use the dataset, Research on Early Life and Aging Trends and Effects (RELATE): A Cross-National Study [4], which compiles cross-national data that contain information on education, income, early childhood health conditions, and adult health conditions. The dataset contains information for 15 countries and includes indicators such as years of education, current income, and mother and father education levels and occupation.

Our dataset had high amounts of missingness, both in the response, the treatment, and the covariates. We investigated the nature of this missingness by examining distributions of covariates for individuals with observed and missing PPP (Figures 1–4). The distribution of these covariates differ between individuals with observed and missing PPP, indicating that our data is not missing completely at random (MCAR). Nonetheless, because we had no previous experience working with missing data, we decided to drop all individuals with missing response variables.

Additionally, we did also dropped individuals with missing values in the covariates. Because most of our covariates were categorial, we could not impute a mean value. Furthermore, because we were interested in the mean causal effect of education on PPP, we thought imputing the median value for these missing covariates would introduce bias into our estimation.

3.2 Model

There are 63632 observations after we drop 24641 observations with missing variables in PPP, education, or the covariates. Our variable of interest, PPP, is yearly per capita household income expressed in Purchasing Power Parity, which facilitates cross-national comparisons. Our treatment is whether the unit receives secondary school education. The control variables in our model are country fixed effects, survey year (aliased with country fixed effects), birthcohort (year of birth binned to 16 categories), gender, whether or not a person was born in a rural area, civil status (i.e. whether the unit is married or not), whether a person was native-born in their current country of residence, and household size. We include civil status is that a person can be more/less motivated depending on the marital status. Similarly, we also control for household size (hhsize), since intuitively a unit might be more motivated to work with a bigger family size; on the other hand, household size could also be negatively associated with income, as a person might have less time to spend on work with a bigger family. We also intended to control for childhood health, but unfortunately we had to eventually drop it because of high missingness. We also could not control for childhood disease indicators, like malaria and tuberculosis, or mother's education, because too many observations were missing these variables.

We first removed the outcome of interest, PPP. We then estimated a propensity score for each unit in the observational study using the fitted values from a logistic regression. In our logistic regression model, we use the covariates mentioned previously and an interaction term between country, rural area birthplace, and gender. Our basic hypothesis is that gender has different effects on education depending on the country. For example, we observe that females in China are a significantly less likely to receive education in China compared to other places. Also in rural areas, women might also be less likely to receive education than in the city. The effect of being in rural area and being a female is also likely to change with country. The full propensity score model is as follows:

$$\begin{aligned}\text{logit}(E(S_i|\text{Covariates})) &= \beta_0 + \beta_1 C_i + \beta_2 R_i + \beta_3 G_i + \beta_4 B_i + \beta_5 M_i \\ &\quad + \beta_6 H_i + \beta_7 N_i + \beta_8 C_i R_i + \beta_9 R_i G_i + \beta_{10} C_i G_i + \beta_{11} C_i R_i G_i\end{aligned} \tag{1}$$

where C_i, R_i, G_i, B_i, M_i, H_i, and N_i are unit i's country, rural birth indicator, gender indicator, birth cohort, civil status, household size, and native birth indicator, respectively. Histograms for predicted propensity scores for observations under control and active treatment are shown in Figure 5. We see a clear difference between the estimated propensity scores for the two groups, with units assigned to active treatment having higher estimated propensity scores, indicating our propensity score estimation is valid.

In the next step, we discarded control units with estimated propensity scores lower than the minimum of the active treated units' estimated propensity scores or higher than the maximum of the active treatment units' estimated propensity scores. These discarded units have the estimated propensity scores being either greater than the highest or less than the lowest estimated propensity scores within the active treatment group. This means these discarded units are very different from any treated units in our sample. Thus it is impossible to estimate the discarded samples' potential outcomes under active treatment using existing samples in the active treatment group.

Using the remaining data, we created five subclasses based on quintiles of the estimated propensity scores. We used two subclassification regimes. The first split units into subclasses based on quintiles of the propensity score. The second used the following quantiles $(0, 0.75, 0.85, 0.90, 0.95, 1)$. The distribution of control and active treatment units for each subclass under each regime are shown in Tables 1 and 2, and plots showing covariate balance within each subclass under the two subclassification regimes are shown in Figures 6–9 and 10–13.

Subclass	no secondary education	yes secondary education
1	13488	366
2	12418	1436
3	10723	3127
4	8286	5568
5	5116	8736

Table 1: Number of units in treatment and control group, subclassification regime 1.

Subclass	no secondary education	yes secondary education
1	43199	8779
2	3287	3723
3	1494	2118
4	1146	2100
5	905	2513

Table 2: Number of units in treatment and control group, subclassification regime 2.

Then we use a linear regression to estimate the treatment effect in each subclass. Because linear regression makes normality assumptions about the error term, we log-transformed our response, PPP. To account for observations where PPP is zero, the transformation we used was $\log(\text{PPP} + 1)$. A histogram of transformed PPP is shown in Figure 14. Our linear model included the main effects from the logistic regression for propensity score except for native birth (after discarding mismatched controls, only 1-2 units per subclass were of non-native birth), the treatment indicator, and a three-way interaction term for country, gender, and rural birth. The full model is as follows:

$$\begin{aligned} \log((PPP)+1) = \beta_0 + \beta_1 C_i + \beta_2 R_i + \beta_3 G_i + \beta_4 B_i + \beta_5 M_i \\ + \beta_6 H_i + \beta_7 C_i R_i + \beta_8 R_i G_i + \beta_9 C_i G_i + \beta_{10} C_i R_i G_i \end{aligned} \tag{2}$$

We examined plots of Cook's distances to identify and remove highly influential observations. We took a weighted average across all the classes to compute the overall treatment effect and its variance, weighting by number of treated units in each class. We then compare this result to naive regression without subclassification based on propensity score.

4 Results

Method	Estimate	Standard Error	95 Percent CI	exp(CI)
Naive Regression	0.11249	0.017134	(0.0789, 0.1461)	(1.0821, 1.1573)
Regression on subset, regime 1	-0.02815	0.03700	$(-0.10067, 0.04436)$	(0.90423, 1.04536)
Regression on subset, regime 2	-0.01960	0.041101	$(-0.10017, 0.06952)$	(0.90567, 1.06285)

Table 3: Results for estimated treatment effect of achieving secondary education on PPP. The confidence intervals are obtained using a normal approximation and represent linear change on the log scale for PPP. The exp(CI) represents change on the original scale.

Our results are shown in Table 3. We see a slight but significant and positive association between secondary education and PPP. Looking at the transformed confidence interval on the original scale, we see that, on average, controlling for country, rural status, gender, birth cohort, civil status, and household size, getting secondary education will increase your income by 1.08–1.15 times.

On the other hand, our results from subclassification show no significant association between secondary education and income.

5 Discussion

We do not detect a significant effect of education on PPP when regressing on subclasses. A significant association is detected when regressing on the entire dataset. However, we believe we cannot make causal inferences because SUTVA and unconfoundedness do not hold.

SUTVA is not plausible because it is possible for units to interfere with each other. Their potential outcomes are not independent, because, for example, high-achieving individuals could motivate their friends or family members.

Additionally, we do not think the covariates that we have in the dataset can support the unconfoundedness assumption. In selecting covariates we want to include all the information that can affect a sample's potential outcomes. However, a person's income is determined by extremely complicated mechanisms. For example, family background, which can be indicated by mother's and father's education/occupation, can affect both the person's education and future income. Another covariate that we would like to have is adult health, since that also affects both a person's education and future income, or childhood health status. However, in our dataset, such additional covariates were either not available or were missing in most of the units.

Because we cannot assume that SUTVA or unconfoundedness hold, we cannot draw strong causal conclusions from this study. However, we can compare the associations obtained from naively running linear regression on the entire dataset vs. running linear regression independently in each subclass. We are able to detect a slight but significant difference when regressing on the whole dataset, but this difference is no longer significant once we subclassify our units based on propensity score and run regressions within each subclass. Subclassification on propensity score puts similar units in classes and compares them directly, thus providing us with more direct comparisons than naive linear regression. When our comparisons are more fair, we do not detect an effect, which could indicate that the effect of education on income might be confounded by other variables.

6 Conclusion

It is often very difficult to find causal relationships from broad observational studies such as this one because there are not enough covariates to support the unconfoundedness assumption. However, we used subclassification by propensity score to obtain a revised estimate for the treatment effect of secondary education on PPP. We achieved different results from naively running linear regression without subclasses, indicating that the association between education and income might be explained by other factors. We hope to find different datasets with additional covariates and less missingness than the RELATE dataset, and provide more definitive evidence whether a causal link exists between education and income.

References

[1] Theodore W Schultz. Investment in human capital. *The American Economic Review*, pages 1–17, 1961.

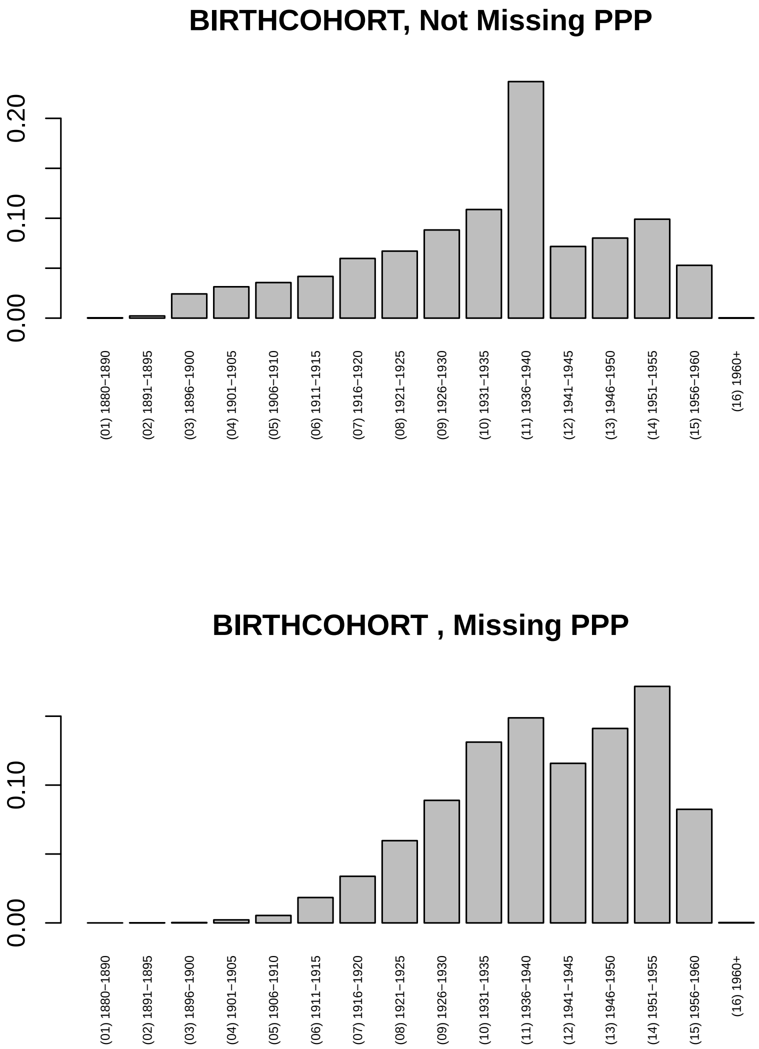
[2] Gary S Becker and Barry R Chiswick. Education and the distribution of earnings. *The American Economic Review*, pages 358–369, 1966.

[3] Barry R Chiswick. Earnings inequality and economic development. *The Quarterly Journal of Economics*, pages 21–39, 1971.

[4] Mary McEniry. Research on early life and aging trends and effects (relate): A cross-national study. `http://doi.org/10.3886/ICPSR34241.v1`, 2013-06-12.

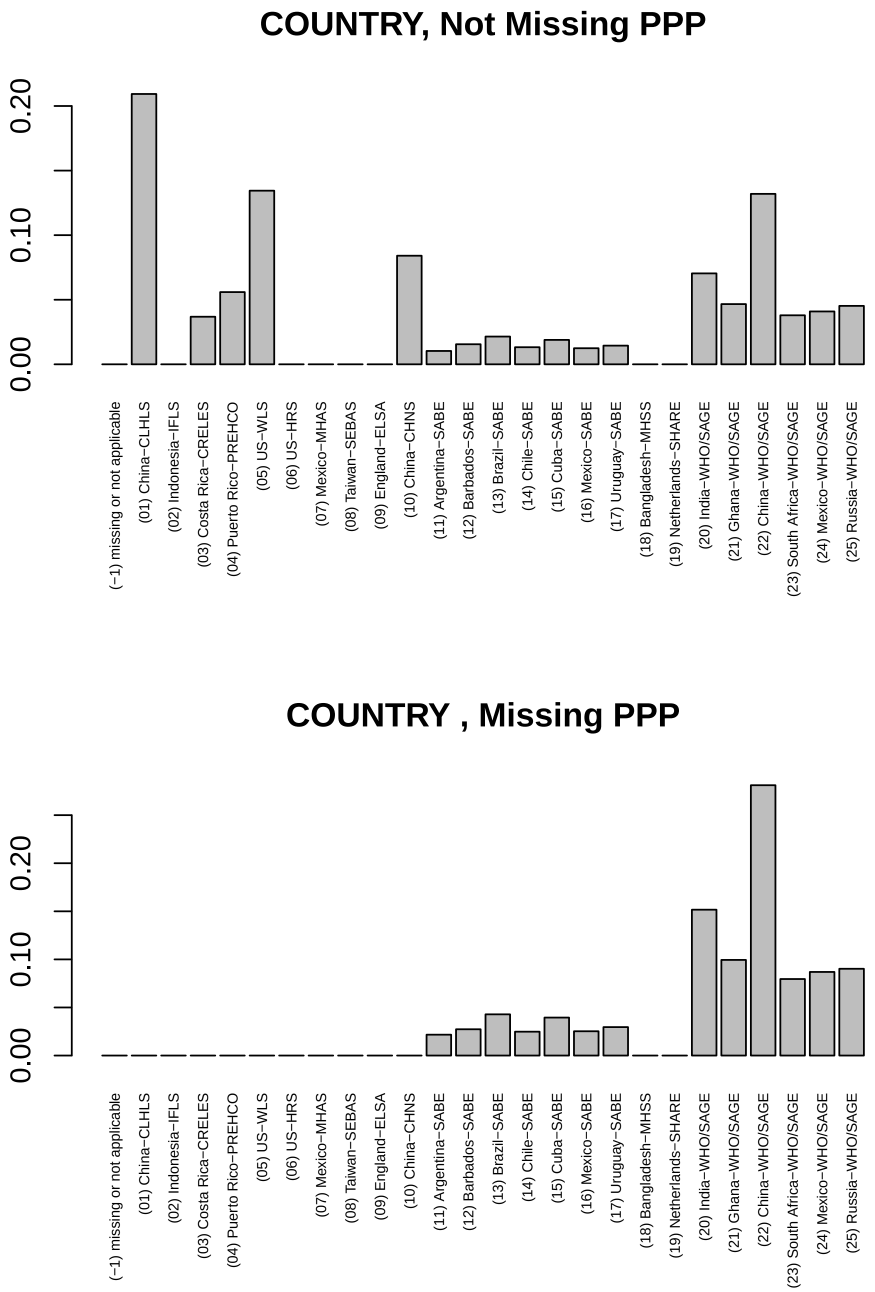
A Appendix of Figures

A.1 Missingness


BIRTHCOHORT, Not Missing PPP

BIRTHCOHORT , Missing PPP

Figure 1: Distribution of birth cohort for observations with non-missing and missing PPP.



COUNTRY, Not Missing PPP

COUNTRY , Missing PPP

Figure 2: Distribution of country for observations with non-missing and missing PPP.

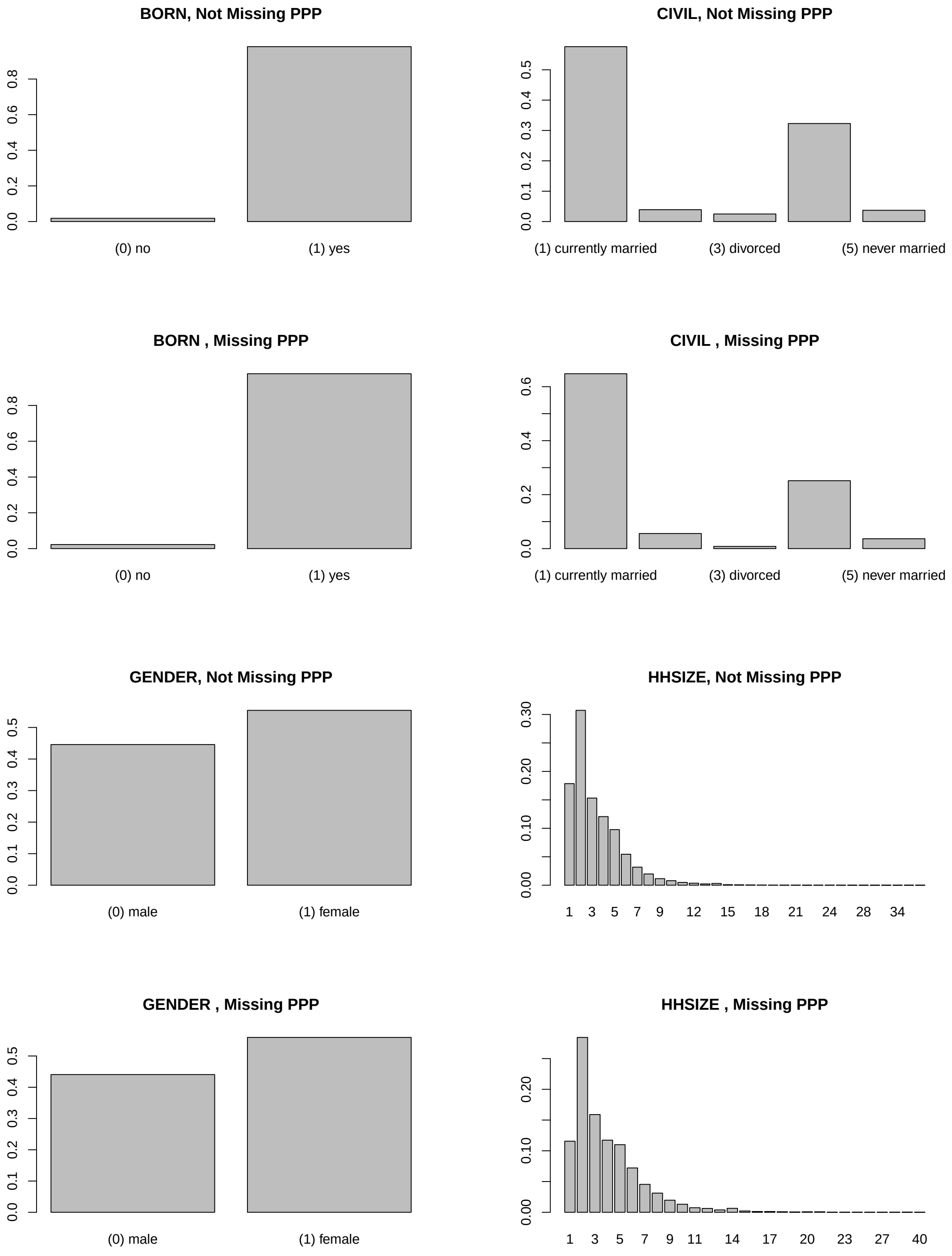


Figure 3: From left to right, top to bottom: Distribution of native birth indicator, civil status, gender, and household size for observations with non-missing and missing PPP.

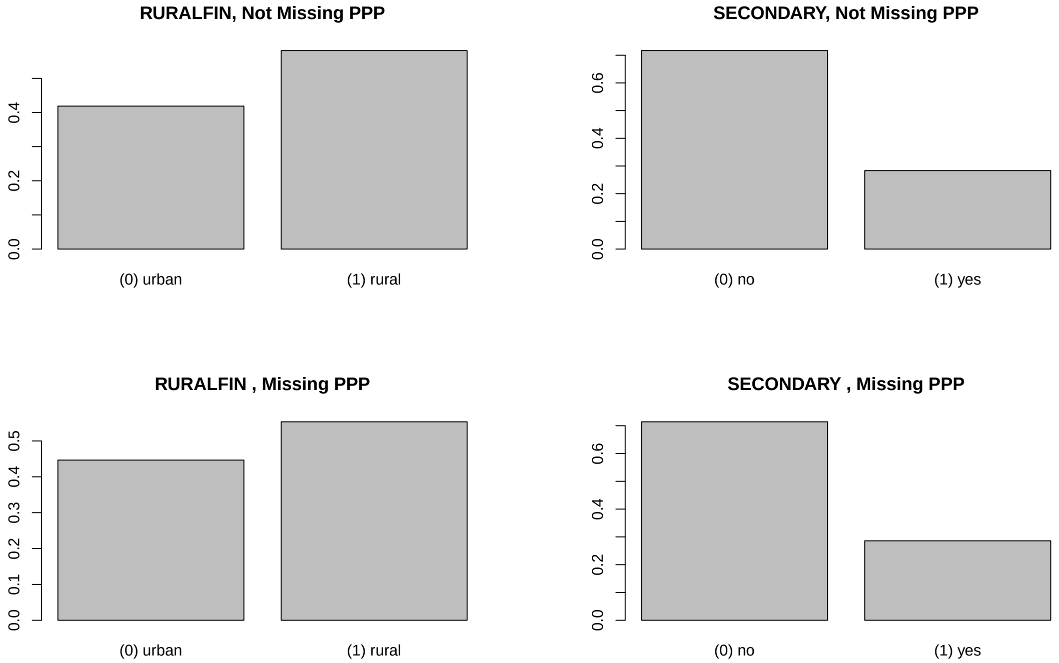


Figure 4: From left to right: Distribution of rural status and secondary education achievement for observations with non-missing and missing PPP.

A.2 Propensity Score Estimation

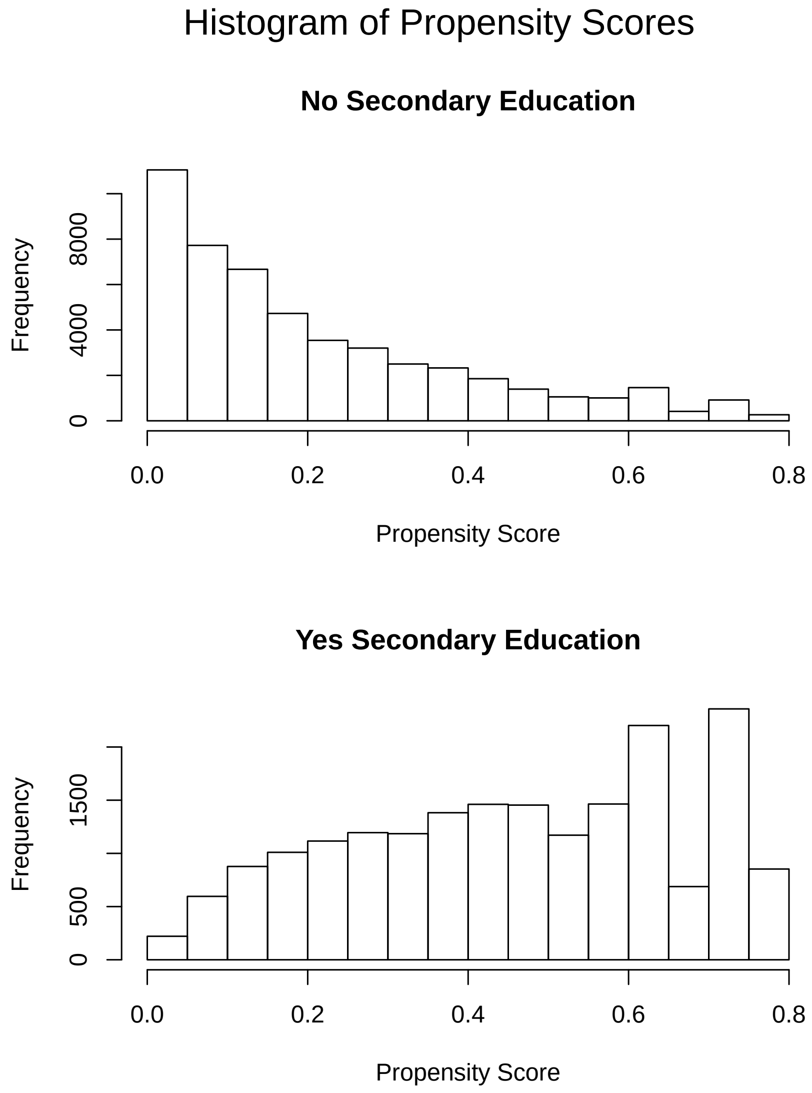


Figure 5: Distribution of estimated propensity scores for units under control (no secondary education) and treatment (yes secondary education).

A.3 Covariate Balance

A.3.1 Subclassification Regime 1

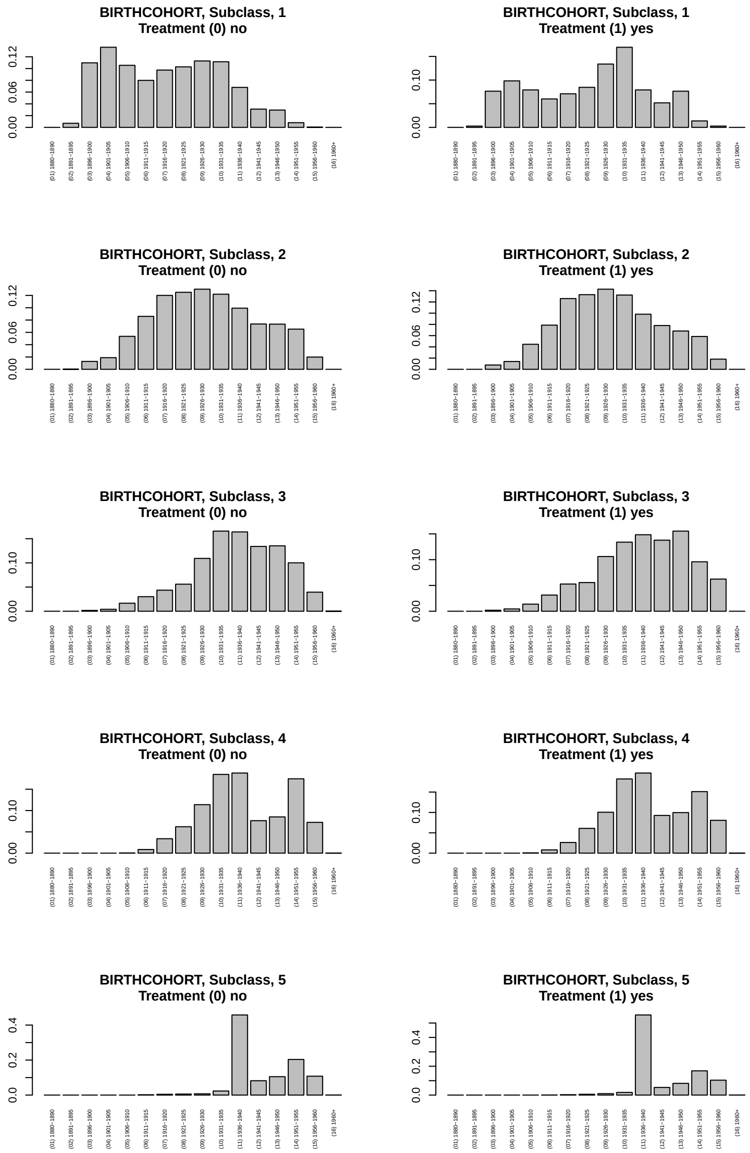


Figure 6: Distribution of birth cohort among treated and control units in each of the 5 subclasses.

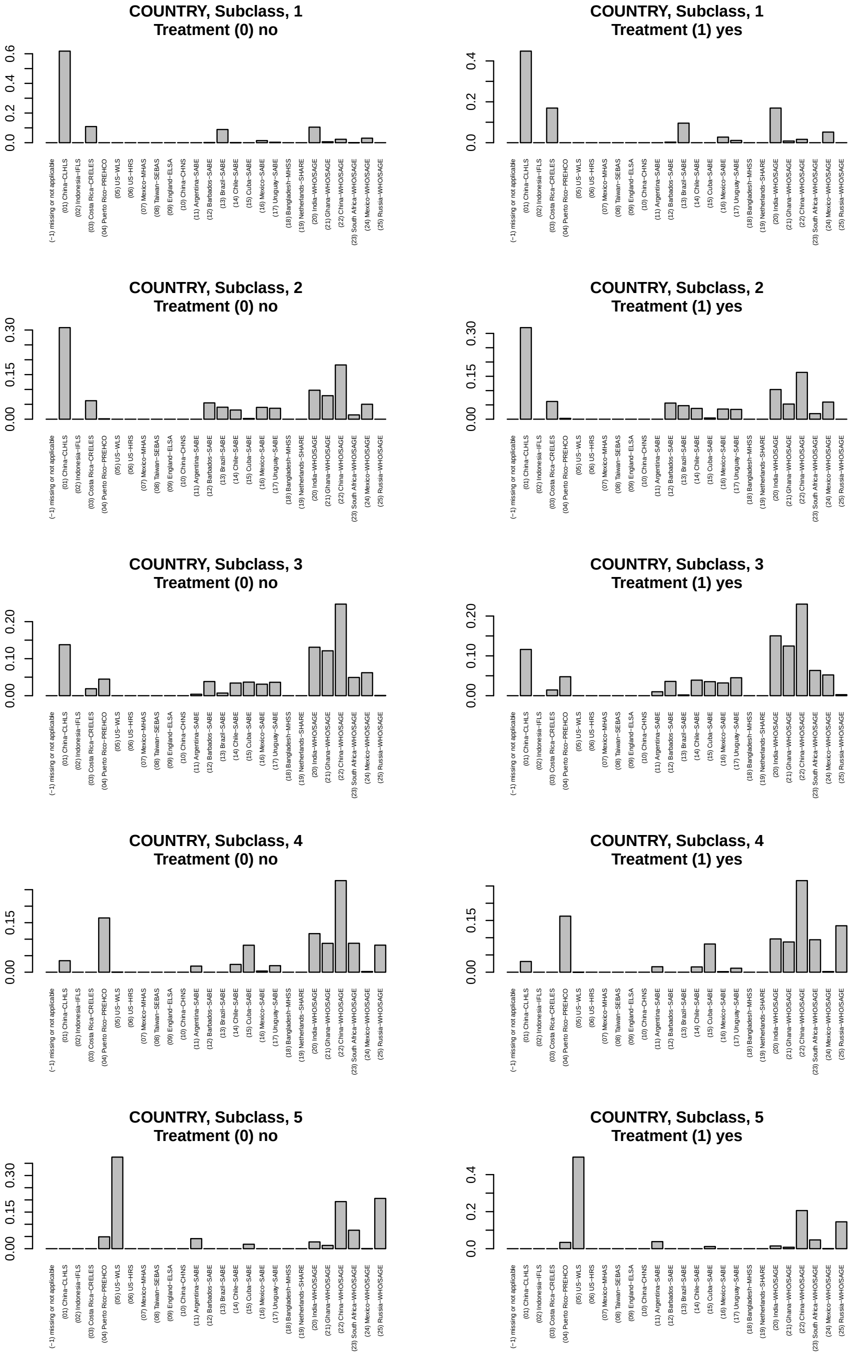


Figure 7: Distribution of country among treated and control units in each of the 5 subclasses.

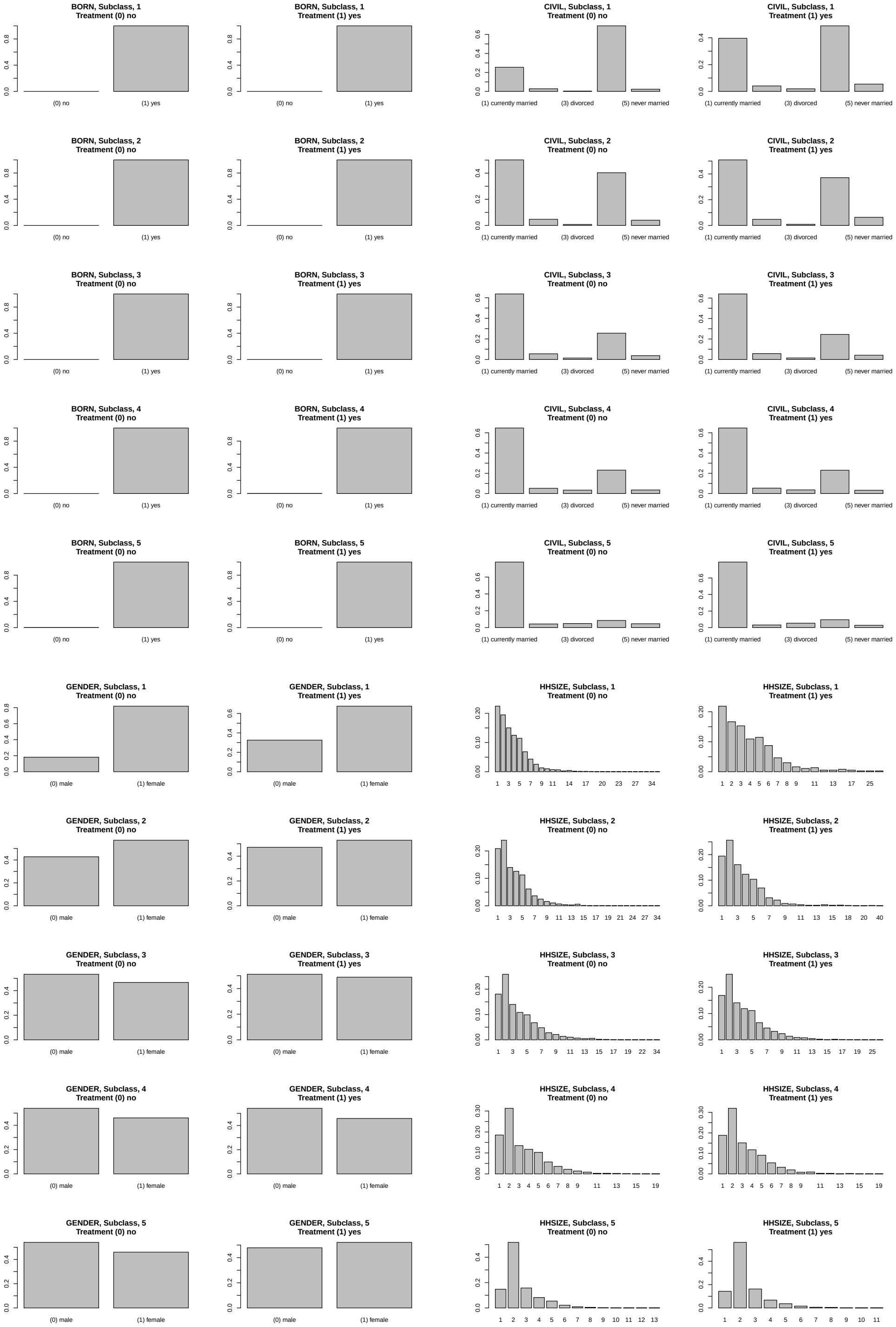


Figure 8: From left to right, top to bottom: Distribution of native birth indicator, civil status, gender, and household size among treated and control units in each of the 5 subclasses.

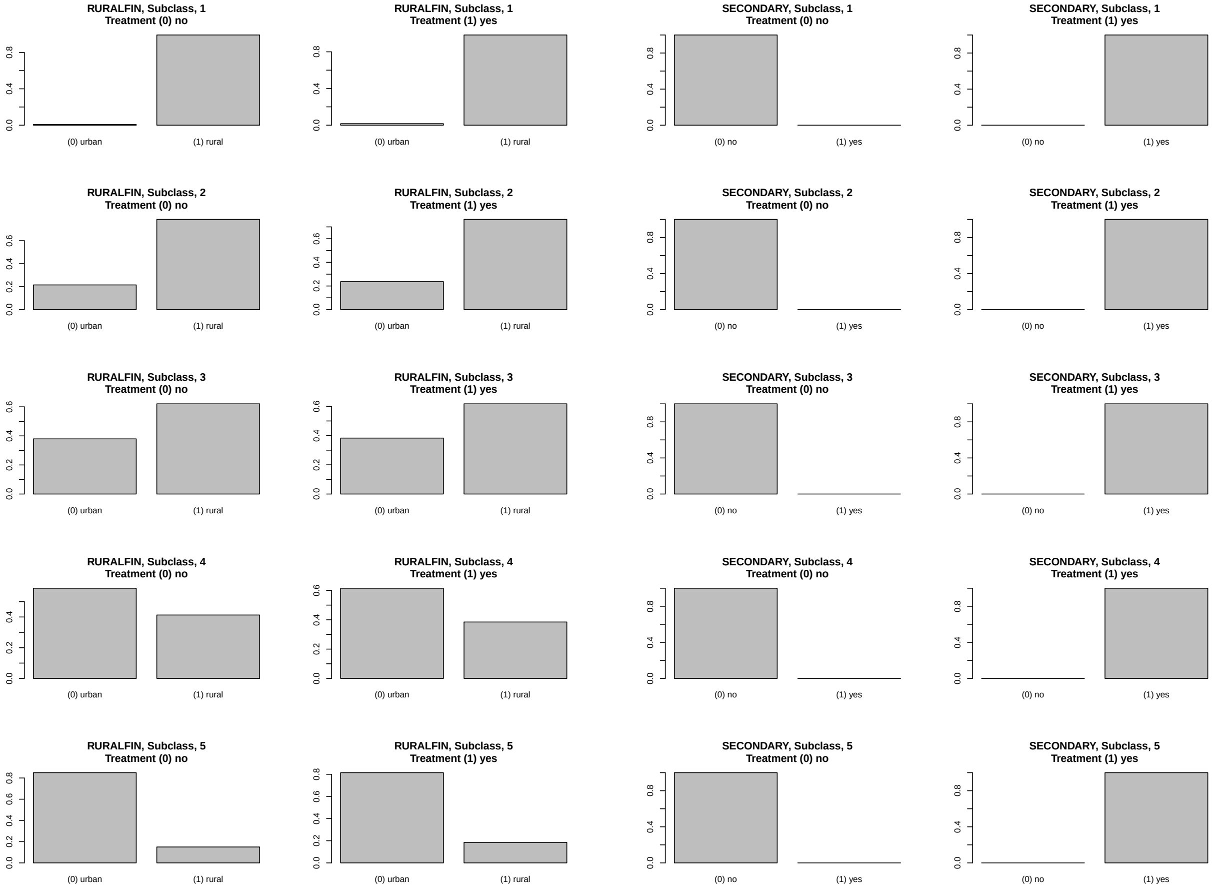


Figure 9: From left to right: Distribution of rural status and secondary education achievement among treated and control units in each of the 5 subclasses.

A.3.2 Subclassification Regime 2

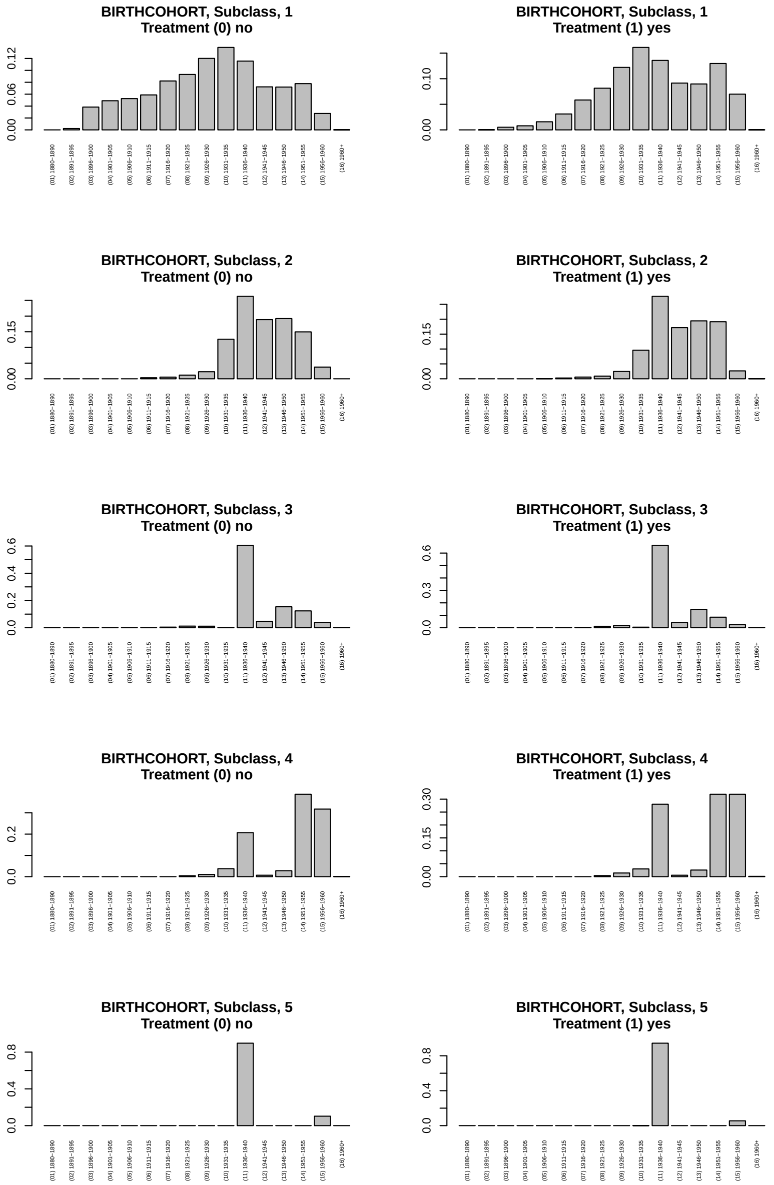


Figure 10: Distribution of birth cohort among treated and control units in each of the 5 subclasses.

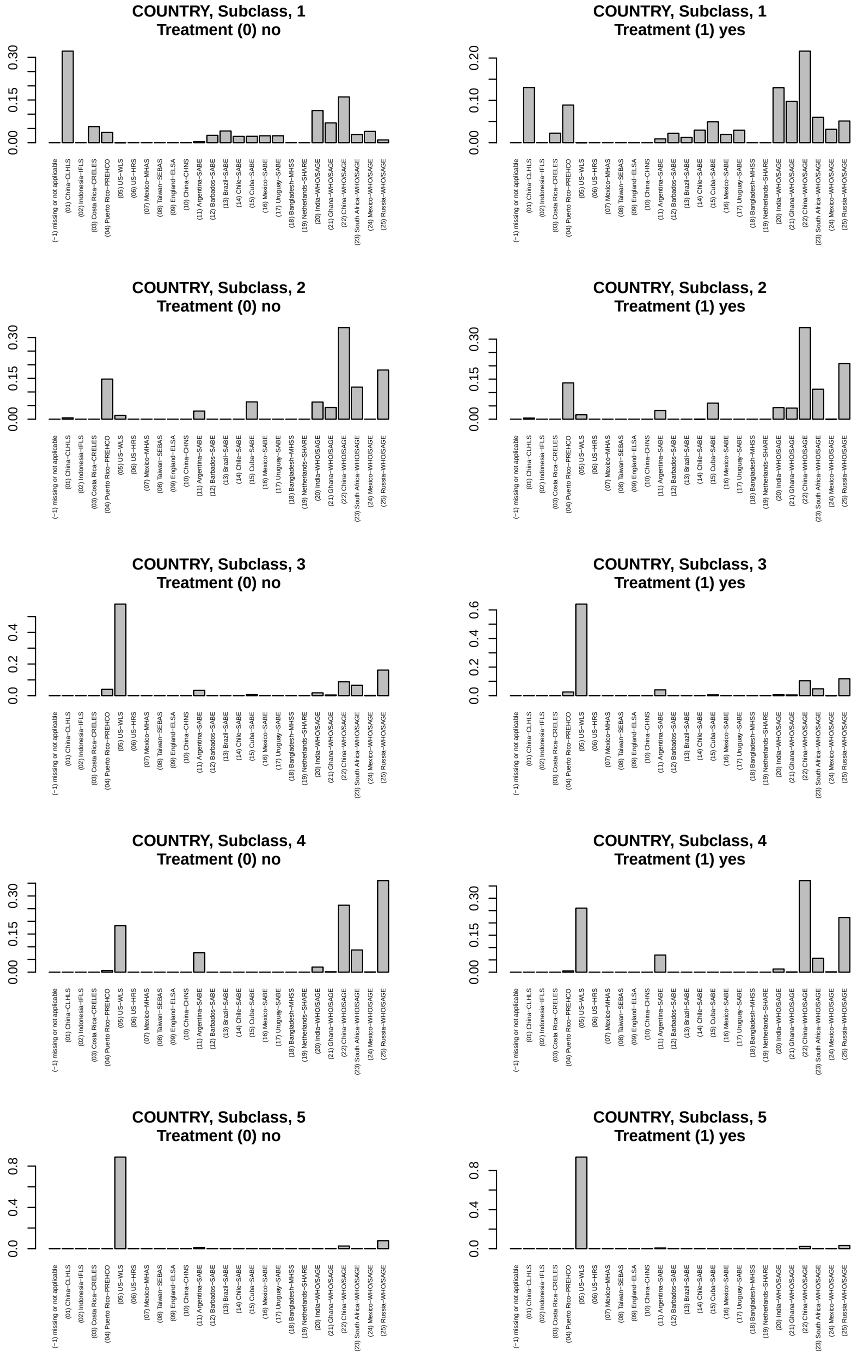


Figure 11: Distribution of country among treated and control units in each of the 5 subclasses.

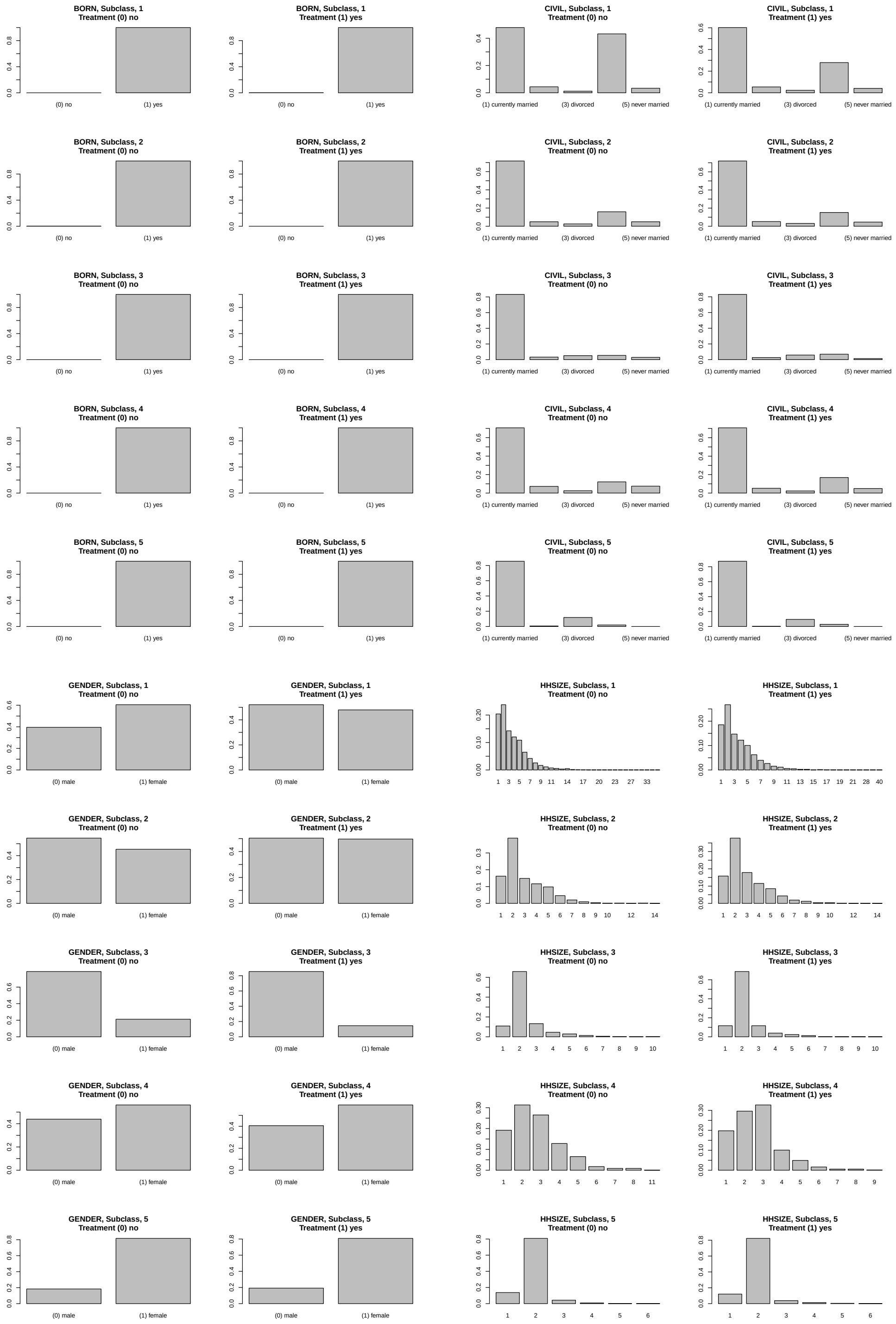


Figure 12: From left to right, top to bottom: Distribution of native birth indicator, civil status, gender, and household size among treated and control units in each of the 5 subclasses.

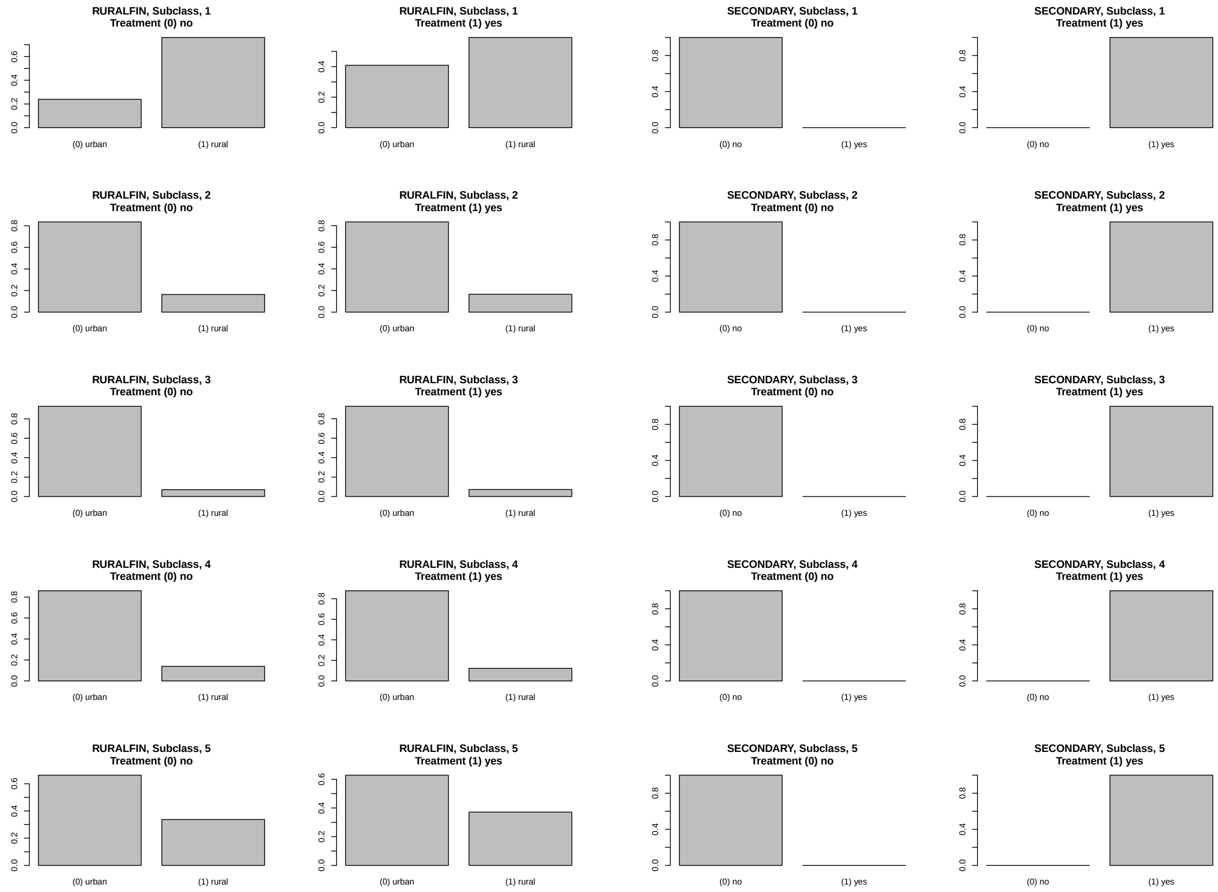


Figure 13: From left to right: Distribution of rural status and secondary education achievement among treated and control units in each of the 5 subclasses.

A.4 Modeling PPP

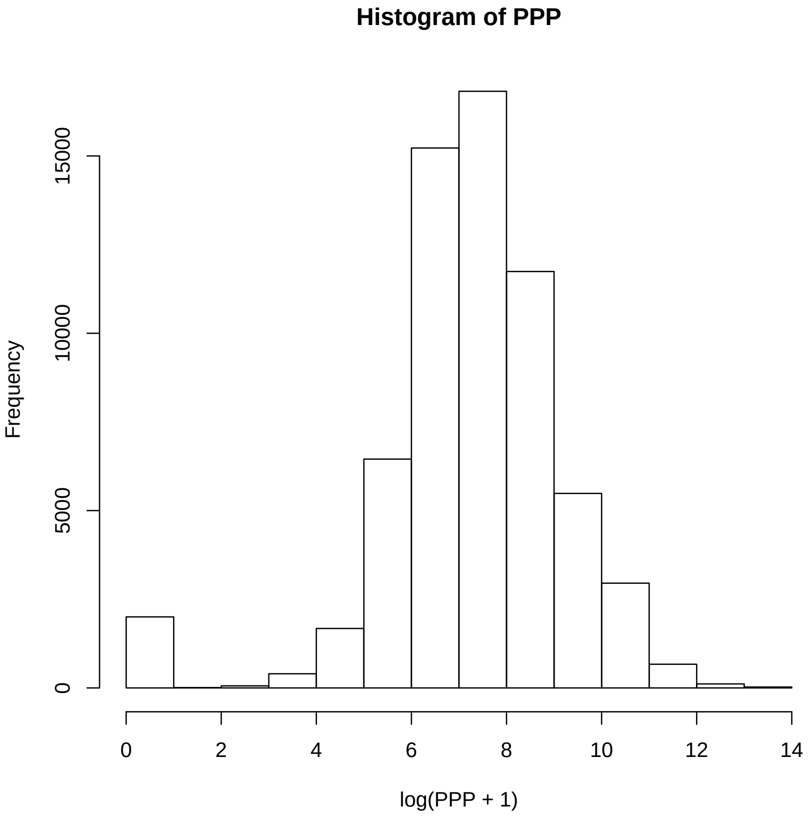


Figure 14: Histogram of log(PPP + 1)