**Question 4:** Focusing on the self-selection case, assess whether the following methods overcome the selection bias. In each case, provide a short discussion of why the method works or why it does not work:

The estimate of β suffers from selection bias if there is a difference in the no-treatment potential outcomes between the treated and the untreated. Here, we attempt to resolve overcome this selection bias using three approaches: instrumental variables, regression discontinuity design, and propensity score matching.

**Part A: Instrumental variables**

We want a variable that is correlated with an endogenous explanatory variable and uncorrelated with the error term. For this purpose, we create a voucher in our hypothetical intervention, represented by the variable “vouch”. This variable will be randomly assigned to farms in the study, and only farms with the voucher can receive the intervention.

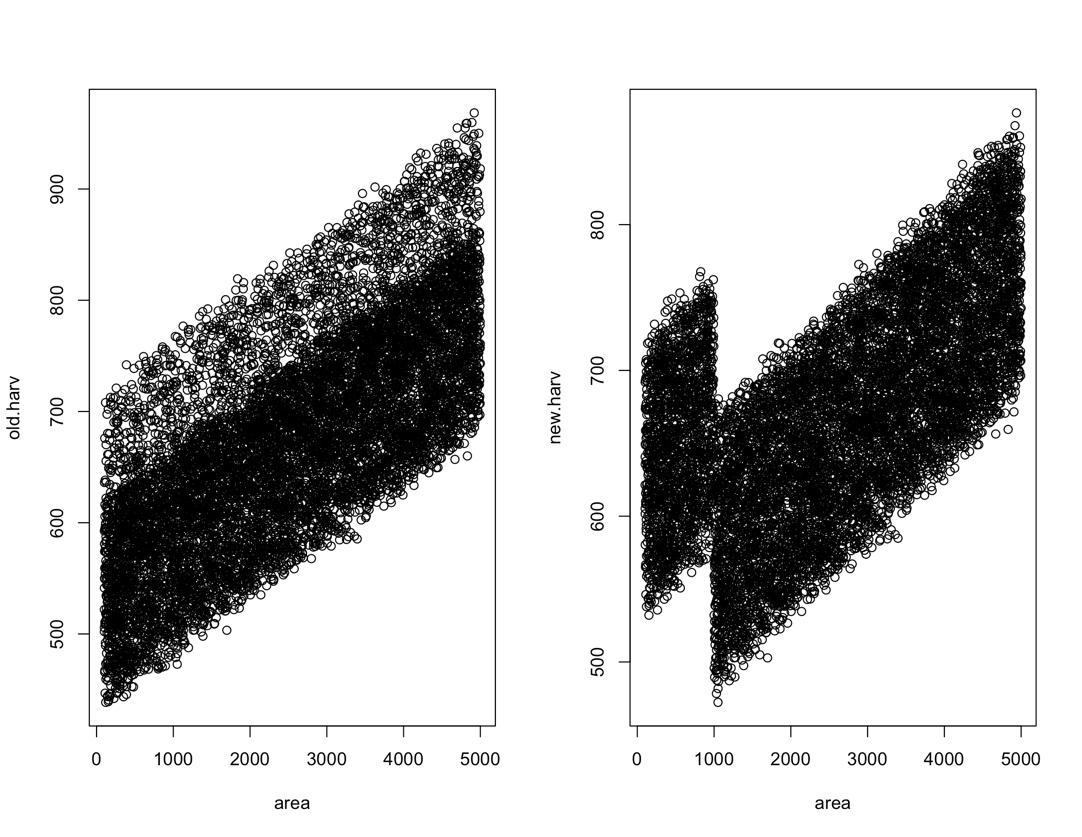
We designate our dependent variable (harvest) as Y1, our endogenous explanatory variable (treatment) as Y2, our exogenous variables (area, sun, rain, exp, dist, child) as X1, and our instrument (vouch) as X2. To compare, we run an ordinary least squares model (Y1 ~ Y2 + X1) and a two-stage least squares model with the instrumental variable (Y1 ~ Y2 + X1 | X1 + X2). The exogenous variables must be included on the instrumental variable side of the model as well as the endogenous variable side.

The results are not different between the two models, so perhaps our instrument is not as good as expected. This doesn’t make sense though, since in theory, it is a perfect instrument that is randomly assigned to half of the study population.

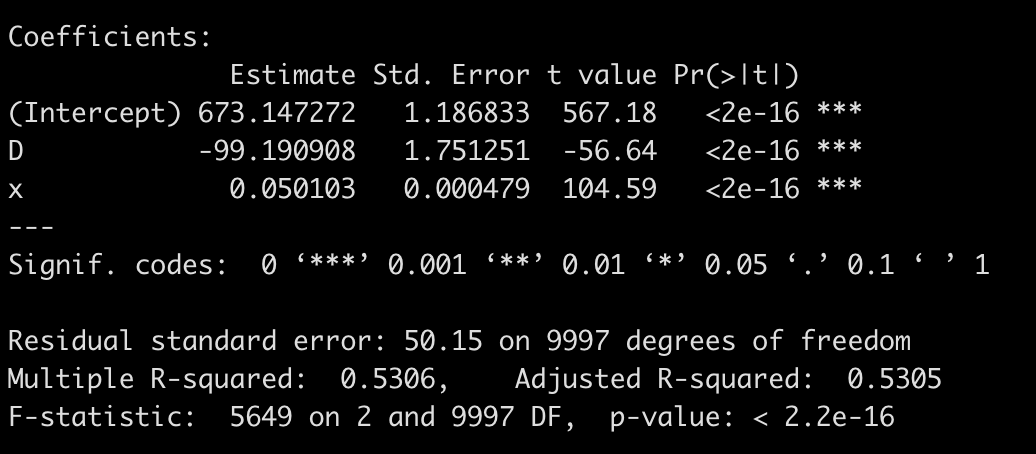
**Part B: Regression discontinuity design**

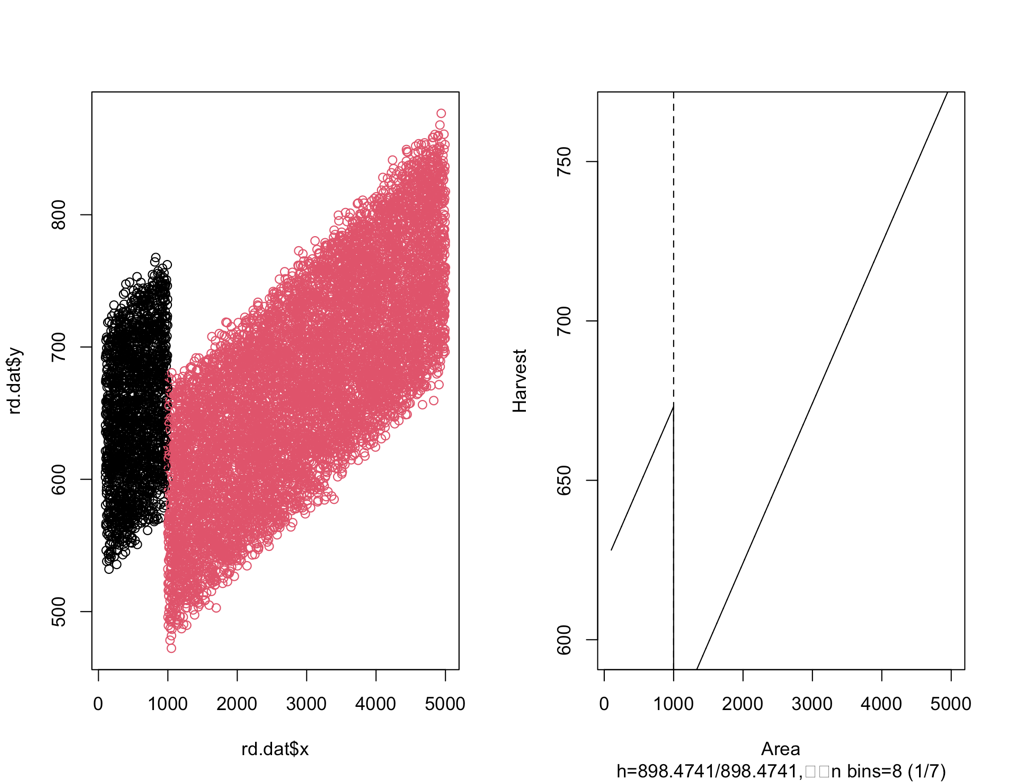
For regression discontinuity design (RDD), there must be some sort of arbritrary boundary or rule that divides the study population. This often happens in government programs, e.g. eligibility for senior health care (Medicare) in the United States at age 65. This would be an example of a sharp discontinuity, though fuzzy discontinuities (e.g. a buffer region) also exist.

For our intervention, we can implement an RDD by redefining the treatment variable such that it restricts access to the intervention based on a farm’s area. Only farms with area less than or equal to 1.000 square meters can participate. After restricting treatment in this way, we then recalculate the harvest values using the new treatment values. We can see that we now have a sharp discontinuity between farmers on either side of the area restriction.



Using a package for RDD in R, we are able to create a sharp RDD linear model based on the 1000 square meter restriction. The results show that there is a strongly significant effect of the treatment for those that were arbitrarily included via the size restriction.

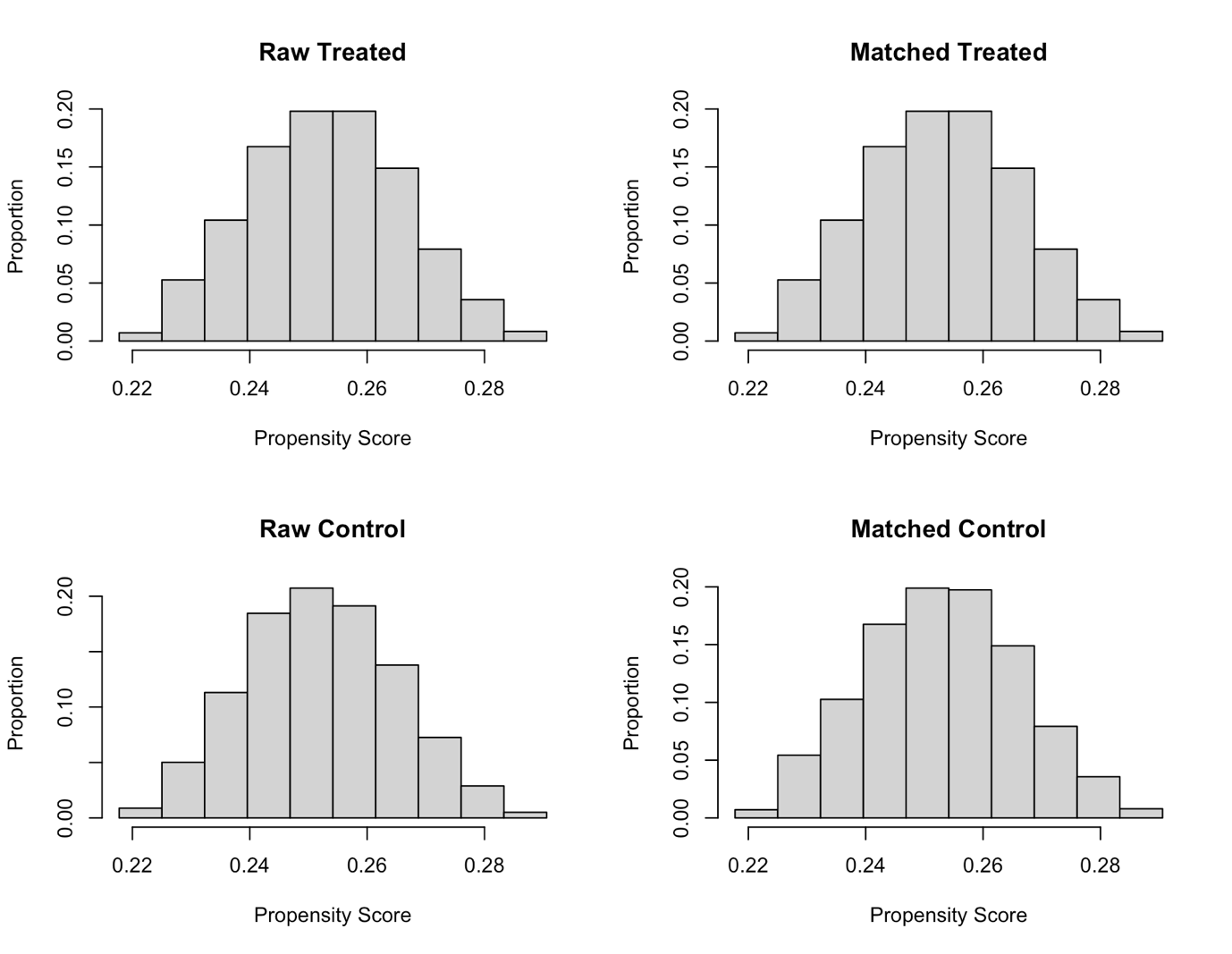




**Part C: Matching approach**

Here we try propensity score matching to see if we can eliminate selection bias by comparing only similar individuals in our model. W use nearest neighbor matching with a 1:1 ratio. This provides a matched dataset of 2524 individuals.

Though propensity score matching is successful in this case, perhaps due to the random data generation that created our dataset, there is already very little difference in the variable means between the treated and untreated groups in the study.



Though we obtain slightly different coefficients with a linear model using the matched dataset, the results do not appear to be very different overall.

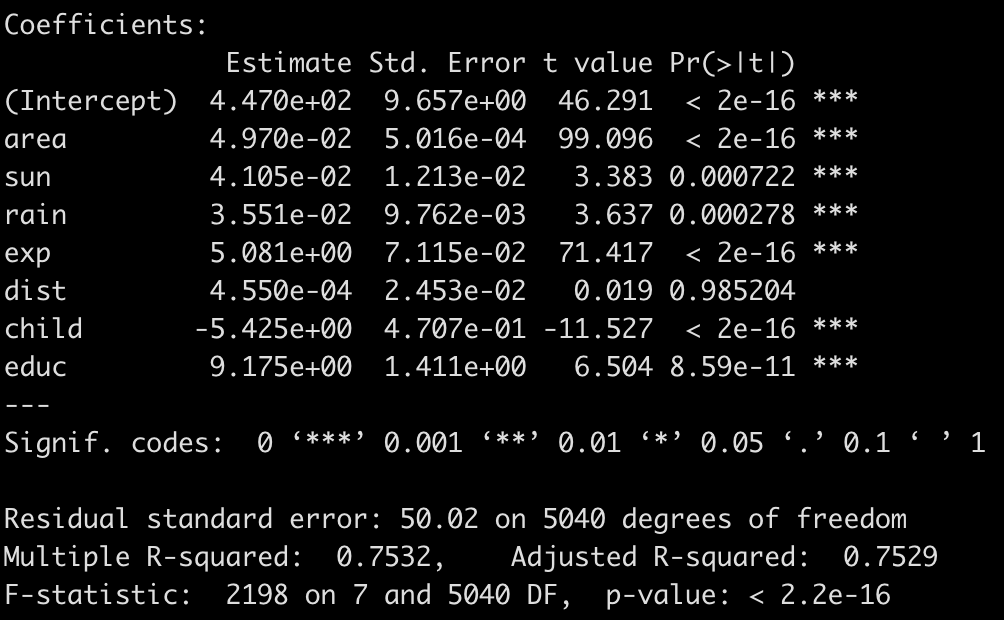


Figure - matched data linear model results

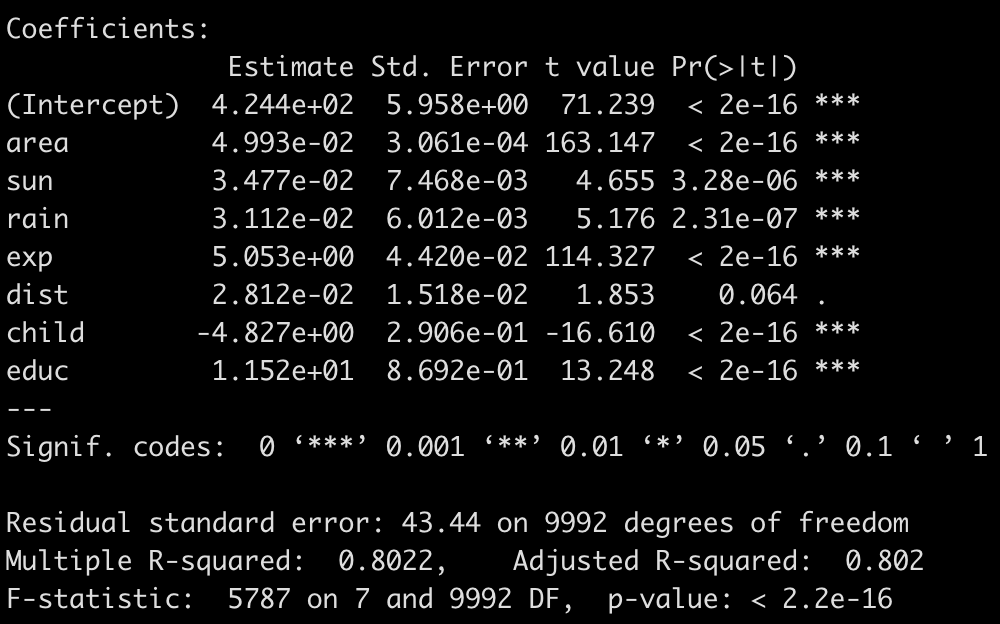


Figure - original linear model results (unmatched)

**Conclusion**

Between the three approaches, the RDD appears to have been the best at reducing selection bias, as it clearly showed a difference based on an arbitrary cut-off.