Section 5: Univariate matching and exact matching

Yotam Shem-Tov

Fall 2015

Motivation

- In RCT's subjects do not control the assignment of treatment. In observational data the assignment of treatment or control is usually controlled (or influenced) by the subjects. This can lead to selection problems, and dissimilarity between the treated units and the control units
- What is the motivation behind matching? It provides a procedure for overcoming problems of dissimilarities between treated units and control units in observational data
- OLS can also be used to adjust for differences between treated units and control units
- Can you think of other procedures to adjust for differences in X? Stratification

Assumptions

- The two main assumptions are:
  1. $Y_{i1}, Y_{i0} \perp T_i | X_i$ - CIA
  2. $0 < Pr(T_i = 1|) < 1$ - Overlap condition
- The CIA implies that conditional on X_i the potential outcomes are exchangeable, i.e

$$\mathbb{E}(Y_{i0}|T_i = 1, X_i) = \mathbb{E}(Y_{i0}|T_i = 0, X_i) = \mathbb{E}(Y_{i0}|X_i)$$

- The first assumption alone implies Ignorability. Both assumptions together imply Strong Ignorability
- Can we identify the *ATT* using matching? Yes
  Can we identify the *ATC* and *ATE* using matching? Yes
- Can we identify the median treatment effect? *No!* The median treatment effect requires stronger assumptions in order to be identified. In a randomized control trail we can also not identify the median treatment effect without stronger assumptions.

Bias in observational studies

- An observational study is biased if the treated and control groups differ prior to treatment in ways that matter for the outcomes under study.
- An overt bias is one that can be seen in the data at hand. For example, the differences between the treated and control groups on age and education are overt biases. Overt biases can be controlled by adjustments, such as matching or stratification. We create matched sets or strata of subjects with the same value of the covariates and then compare treated and control subjects within these strata.
- A hidden bias is similar to an overt bias but cannot be seen because the required information was not observed or recorded. For example, if IQ differs between the treated and control groups and IQ matters for earnings even after we control for education and the other variables, then the study has a hidden bias.

Matching: Mechanical vs Scientific Tasks

- Two tasks are required for inference using matching
  1. Constructing match pairs
  2. Diagnostics of the matched treatment and control
- The first task is how do we create matched pairs. This is a fairly mechanical task
- The second task is to decide whether or not those units that look comparable are comparable ... and this is not a trivial task
- Ultimately, we are asking ourselves if our mechanical operations are sufficient for identification of our treatment effect.
- As Rosenbaum says, "The second task is not a mechanical but rather a scientific task, one that can be controversial and difficult to bring to a rapid and definitive closure; this task is, therefore, more challenging, and hence more interesting."

The treatment assignment mechanism: Notation (Rosenbaum)

We consider a simple model of treatment assignment:

- N - the number of units in the study
- M - the number of treated units i the study
- X_i - the observed covariates of unit i
- u_i - unobserved covariate of unit i
- $T_i \in \{0, 1\}$ - the treatment assignment of unit i
- $Y_i = T_i Y_{i1} + (1 - T_i) Y_{i0}$ - the observed response
- π_i - the probability that subject i received treatment, independently of other subjects
- In this notation we allow π_i to very between units
- The assignment of treatment to one unit does not affect the probability of other units to be treated

A Model of Treatment Assignment

- In the population before matching, we imagine that each subject i received treatment with probability π_i.
- Is $\pi_i \perp \pi_j$? Is $Cov(\pi_i, \pi_j) = 0$?
- **Assumption: The allocation of treatment to unit i is independent of the allocation of treatment to unit j, i.e $\pi_i \perp \pi_j$**
- Is this assumption equivalent to SUTVA? *No* What is the differences? *SUTVA is independence of the potential outcomes of unit i and unit j, it does not refer to the probabilities of treatment assignment*

A Model of Treatment Assignment

- π_i may vary from one person to the next and is not known. More precisely:

$$\pi_i = Pr(T_i = 1 | Y_{i1}, Y_{i0}, \mathbf{x}_i, u_i)$$

$$Pr(T_1 = t_1, \ldots, T_N = t_n | Y_{11}, Y_{10}, \mathbf{x}_1, u_1, \ldots, Y_{n1}, Y_{n0}, \mathbf{x}_n, u_n)$$

$$= \prod_{i=1}^{N} \pi_i^{t_i} (1 - \pi_i)^{1 - t_i}$$

- Contrary to RCT the probability of being assigned to treatment is a function of the potential outcome, (Y_{i1}, Y_{i0})
- The ideal objective is to know exactly what are $(\pi_1, \ldots, \pi_n)$, however this quantities are not observed.
- Is (Y_{i1}, Y_{i0}) random variables? Is x_i random variables? What is random in this model?

The Ideal Match

- Suppose that we could find two subjects, say k and l, such that exactly one was treated, $T_k + T_l = 1$, but they had the same probability of treatment, $\pi_k = \pi_l$.
- We can pair these two subjects and call them a match pair. Note, though, that we are imposing an assumption because we now require that $0 < \pi_i < 1$, otherwise we wouldn't be able to find matches.
- It is difficult to create this matched pair because we don't observe u_k or u_l, and we either observe Y_{k1} or Y_{l1} (but not both) and either Y_{k0} or Y_{l0}.
- Supposing that we could create a matched pair with $\pi_k = \pi_l$ and $T_k + T_l = 1$, then what would this give us?

Treatment odds

Recall Bayse theorem: $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$.
Define,

$$A = T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l$$

$$B = T_k + T_l = 1 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l$$

Hence, $P(B|A) = 1$,

$$Pr(T_k + T_l = 1 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l, T_k = 1, T_l = 0) = 1$$

Therefore,

$$\begin{aligned} & Pr(T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l, T_k + T_l = 1) \\ &= \frac{Pr(T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)}{Pr(T_k + T_l = 1 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)} \\ &= \frac{P(A)}{P(B)} \end{aligned}$$

Treatment odds continue...

$$\frac{Pr(T_k = 1, T_l = 0 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)}{Pr(T_k + T_l = 1 | Y_{k1}, Y_{k0}, \mathbf{x}_k, u_k, Y_{l1}, Y_{l0}, \mathbf{x}_l, u_l)}$$

$$= \frac{\pi_l^{1+0}(1-\pi_l)^{(1-1)+(1-0)}}{\pi_l^{1+0}(1-\pi_l)^{(1-1)+(1-0)} + \pi_l^{0+1}(1-\pi_l)^{(1-0)+(1-1)}}$$

$$= \frac{\pi_l(1-\pi_l)}{\pi_l(1-\pi_l) + \pi_l(1-\pi_l)} = \frac{1}{2}$$

- It is important to note that this is always true if treatment is assigned by the fair flip of a fair coin
- Or independent flips of a group of biased coins where the same biased coin is used when subject i and subject j have the same observable characteristics (assuming no coins have a probability of 0 or 1)

Exact Matching: Motivation

- If the naive model is true, then it is clear that if we can exactly match on $\mathbf{x}$, then the model will follow and we can reconstruct the distribution of treatment assignments in a randomized paired experiment simply by matching based on the observed covariates
- If there is only one covariate that determines how treatment is assigned, then this is straight forward: we just matched on that covariate
- With a large enough sample, it might even be straight forward to exactly match on a couple of covariates, however it becomes very difficult to exactly match on many covariates, especially with finite samples (see *the curse of dimensionality*)
- With continuous covariates it can easily be impossible to perform exact matching

Starifying on **x**

(equivalent to exact exact matching, with multiple units in each strata)

- Stratification on x: From the M units, select $N \leq M$ units and group them into S non-overlapping strata with n_s units in stratum s. In selecting the N units and assigning them to strata, use only the x's.
- Renumber the units so the ith unit in stratum s has treatment assignment T_{si} and covariate x_{si}
- Let $T = (T_{11}, ..., T_{Sn_S})$ and $m = (m_1, ..., m_S)$, where $m_s = \sum_{i=1}^{n_s} T_{si}$
- An exact stratification on x has strata that are homogeneous in x, so $x_{si} = x_{sj}$, $\forall_{i,j}$ in strata s
- With exact stratification on x,

$$\pi_i = Pr(T_i = 1 | Y_{i1}, Y_{i0}, \mathbf{x}_i, u_i) = \frac{1}{|\Omega|}$$

where $\Omega = \prod_{s=1}^{S} \binom{n_s}{m_s}$. Is there any assumptions for the above equality to hold?

Matching on **x**

- Exact matching is a special form of stratification in which there are constraints on the number of observed treated and control units in each stratum. A Matching on x is a matched sample formed by:
  1. placing some restrictions on $S, \mathbf{m}, \mathbf{n}$
  2. picking a stratification that meets these restrictions based exclusively on the patterns of x
- Examples:
  1. Pair matching requires one treated and one control unit in each stratum.
  2. Matching with multiple controls requires one treated and at least one control unit in each stratum.

Propensity Score

- The propensity score is defined as the conditional probability of treatment, $T = 1$ given the observed covariates $\mathbf{x}$

$$e(\mathbf{x}) = Pr(T = 1|\mathbf{x})$$

- The balancing property is always true, regardless of if the naive model holds or not. The balancing property states that treated and control units with the same propensity score have the same distribution of the *observed* characteristics. This gives us that treatment and observed covariates are conditionally independent given the propensity score.

$$Pr\{\mathbf{x}|T = 1, e(\mathbf{x})\} = Pr\{\mathbf{x}|T = 0, e(\mathbf{x})\} \Leftrightarrow T \perp \mathbf{x}|e(\mathbf{x})$$

- It is important to see that within a given matched pair, it is not necessary that subject k and subject l have the same values of $\mathbf{x}$, only that they have the same propensity score, $e(\mathbf{x}_k) = e(\mathbf{x}_l)$.

Propensity Score

- We often estimate the propensity score, coming up with an estimate $\hat{e}(\mathbf{x})$ to produce balance on the observed covariates $\mathbf{x}$
- If the naive model *were* true, then from the propensity score we could get ignorable treatment assignment. We could produce the "ideal match" from the propensity score, since it just reduces our dimensionality of $\mathbf{x}$
- If the naive model holds, then $\pi_i = e(\mathbf{x})$, so matching on the propensity score is matching on π_i. In the naive model:

$$T \perp Y_{i1}, Y_{i0}, u_i | \mathbf{x} \Rightarrow T \perp Y_{i1}, Y_{i0}, u_i | e(\mathbf{x})$$

Estimating the propensity score

- How should we estimate $e(\mathbf{x})$? What model (method) should we use getting $\widehat{e}(\mathbf{x})$?
- Should we use a Logistic regression?Should we use a Probit model?
- If we choose a Logistic regression, should we use all the covariates? should we include interactions? how should we choose which covariates, interactions, and higher order terms to include?
- Should we use OLS? Is Logistic regression better than OLS? Logistic regression is **not** necessarily better than OLS
- I recommend using a general linear model such as Logit or Proit.
- What is our objective in choosing a method for estimating $e(\mathbf{x})$?

Example: Welders and DNA

From *Design of Observational Studies*

- *Welders get exposed to chromium and nickel, substances that can cause inappropriate links between DNA and proteins. Costa, Zhitkovich, and Toniolo measured DNA-protein cross-links in samples of white blood cells from 21 railroad arc welders exposed to chromium and nickel and from 26 unexposed controls. All 47 subjects were male. In their data there are three covariates, namely age, race and current smoking behavior. The response is a measure of DNA-protein cross-links*
- The covariate balance prior to matching is,

	Ave. Treat	Ave. control	T-test	Wilcoxon	KS
age	38.238	42.692	0.029	0.068	0.073
smoker	0.524	0.346	0.233	0.230	0.857
black	0.095	0.192	0.349	0.367	1.000

Example: Welders and DNA

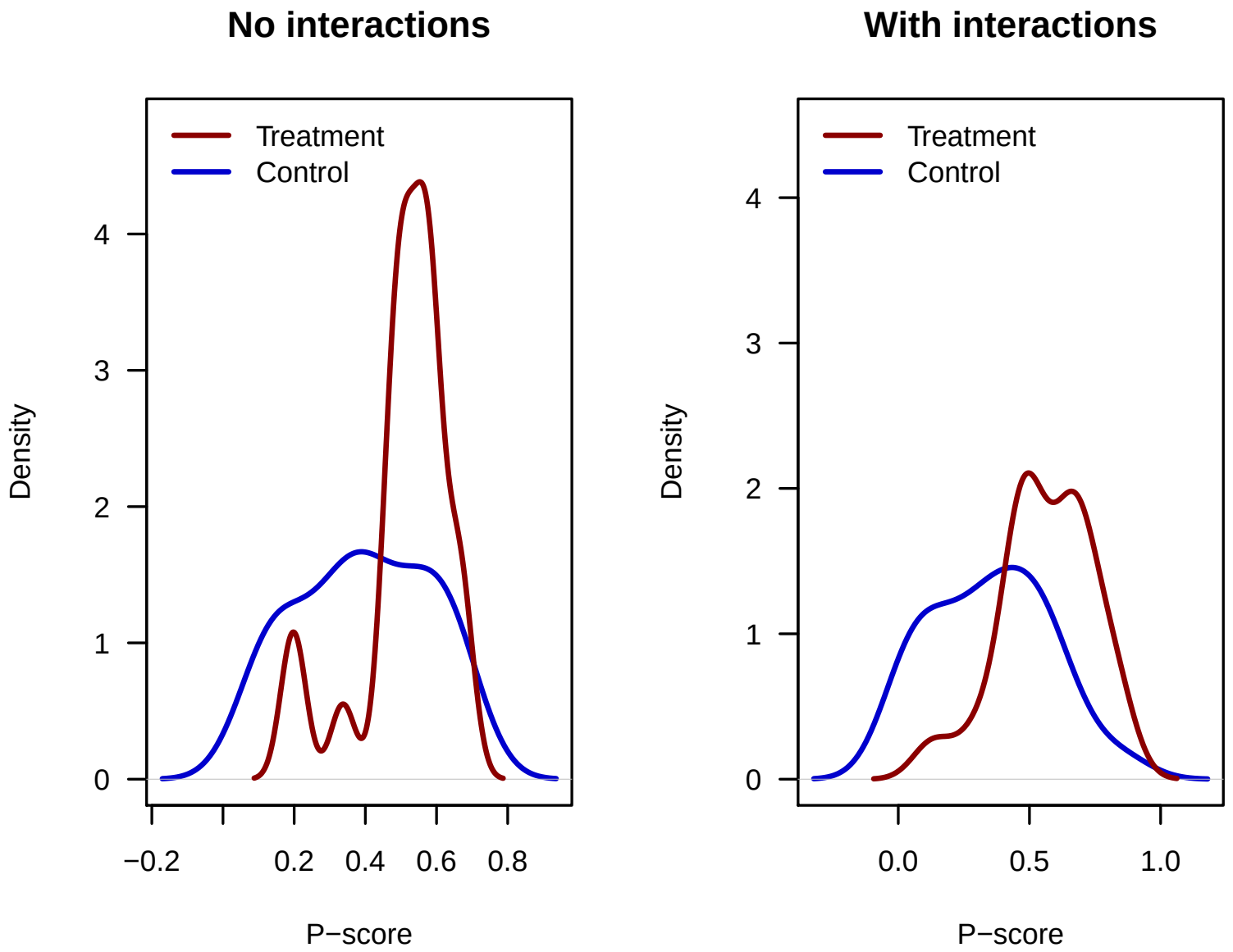
From *Design of Observational Studies*

- The balance looks not bad. Can we diagnose similarity in the joint distribution of covariates using a balance table? *No*
- We estimate two models of propensity score, with and without interactions using a logistic regression (this choice of model is arbitrary and OLS or Probit could also be used)

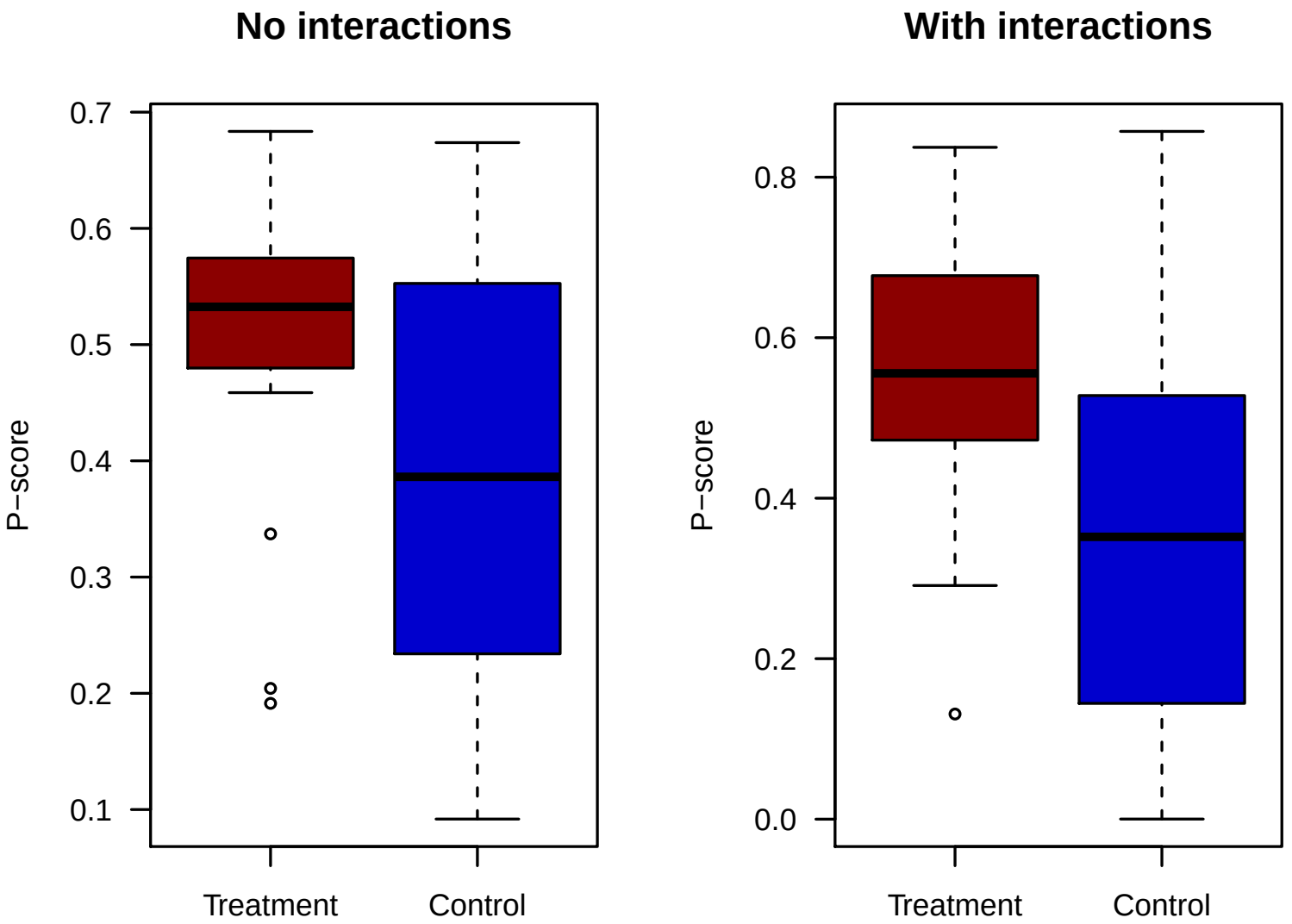
```
ps.model1 <- glm(treat~(.),data=
data.frame(treat=treat,x),family=binomial(link=logit))
ps.model2 <- glm(treat~(.)^2,data=
data.frame(treat=treat,x),family=binomial(link=logit))
```

- Where x is a data frame with all three covariates, and *treat* is a treatment indicator

Welders and DNA: P-score balance



Welders and DNA: P-score balance



matching: Some Considerations

- The basic considerations when performing matching are:
  1. What distance metric do we use?
  2. Do we match with replacement or without replacement?
  3. What do we do with ties?
  4. What do we consider a "good" match?

- We will begin with limiting ourselves to the case of matching only on the propensity score, with replacement

Distance Metrics

- In order to figure out what the "closest" match is, we have to decide what our metric for the distance between observations k and l
- Since we are only matching on one covariate, in this case the propensity score, we can use the squared distance between the two estimated propensity scores

$$d = (\hat{e}(\mathbf{x}_k) - \hat{e}(\mathbf{x}_k))^2$$

- This will punish large differences more than small distances. Alternatively, we could use the absolute value of the distance between the estimated propensity scores or the Mahalanobis distance measure
- Whatever our distance metric, "nearest-neighbor" (hence NN) matching matches the closest control unit to each treated unit (in the case of ATT) or the closest treated unit to each control unit (ATC)

With or Without Replacement

- If we match without replacement, then once we match a control unit, we take it out of the pool of potential matches for all remaining treated units.
- It is important to notice that if we do this, then depending on the order of the controls and the algorithm we use to sort through them, we may get different matches. Therefore the match should be done in away which will take that into account. A matching algorithm which is not invariant to the order of the observations is a bad algorithm
- If we match with replacement, then this means that after a control gets matched to a treated unit, it goes back into the pool of potential matches for the remaining treated units. This means that a control unit could be matched to multiple treated units
- In general, we'd like to match with replacement to make sure that we get the "best" match every time

Ties

- The case may arise that when we look for matches to a given treated unit i, there are two control units that are the same distance from i based on our distance metric d.
- What do we do?
- Flip a coin
- Allow a tie: we match both control units to treated unit i, but we give each of these controls a weight of $\frac{1}{2}$ in our matched data set (in effect, we average the control units).
  It's important to keep in mind that in this method any analysis of the data after the matching step will need to include weights

Caliper matching

- What if the closest control unit to treated unit i has a large distance, d. We may want to say that treated unit i cannot be matched because there is no control unit that is "close" to it.
- To do this, we would enforce a caliper, which says that if there is no "nearest neighbor" to treated unit i, defined as being within a certain distance of i, we say that we cannot match treated unit i.

$$|\hat{e}(\mathbf{x}_i) - \hat{e}(\mathbf{x}_k)| > w$$

  Where, if the distance is greater than the caliper w, we set the distance to infinity.
- When we drop treated observations, we are changing what we are estimating, it is no longer the ATT ...
- How should we choose the caliper? I don't know any clear and good answer

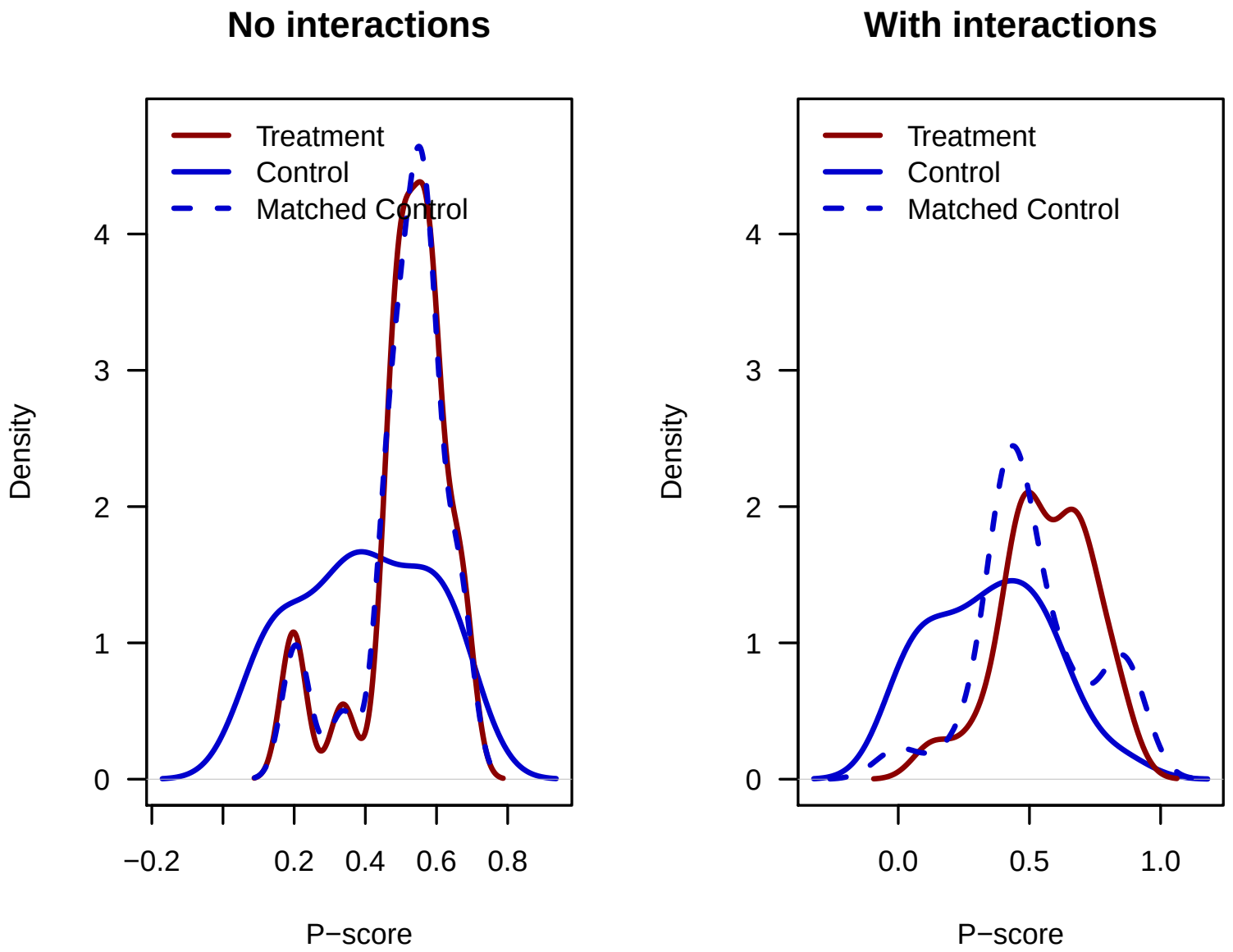
Welders and DNA: Matching

- We conduct NN matching with replacement, according to squared distance measure. The *R* code is bellow,

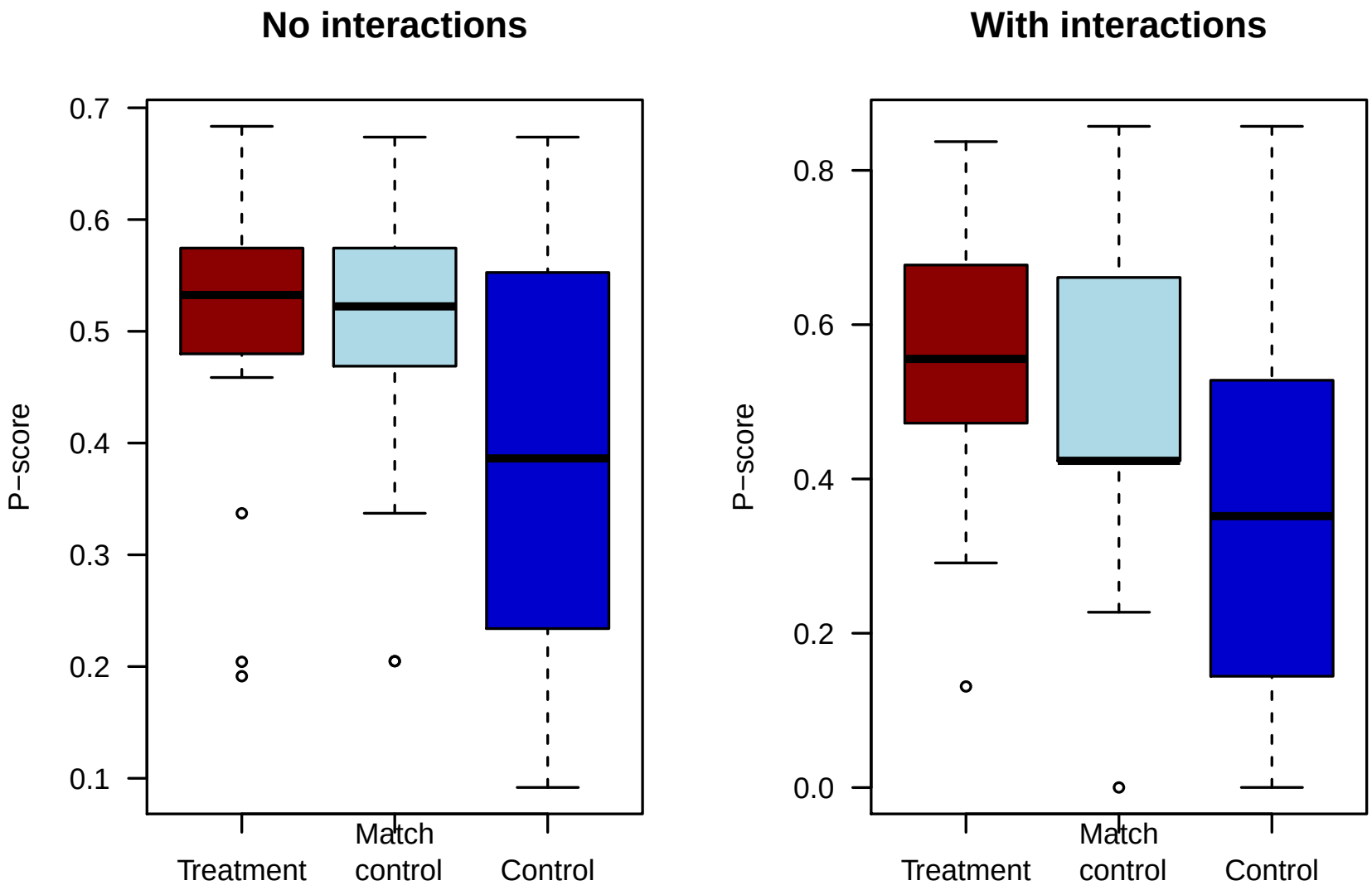
```
ps1.t = ps1[treat==1]
ps1.c = ps1[treat==0]
nt <- length(ps1.t)
d2 = function(x1,x2){return((x1-x2)^2)}
ps1.c.match <-index.control <- as.list(rep(999,nt))
for (i in c(1:nt)){
  ps1.c.match[i] = ps1.c[which.min(d2(ps1.t[i],ps1.c))]
  index.control[i] = which.min(d2(ps1.t[i],ps1.c))
}
```

- The results of the matching procedure are without ties. What is the balance in $\hat{e}(\mathbf{x})$ after matching? Did the matching generated comparable treatment and control groups?

Welders and DNA: P-score after matching



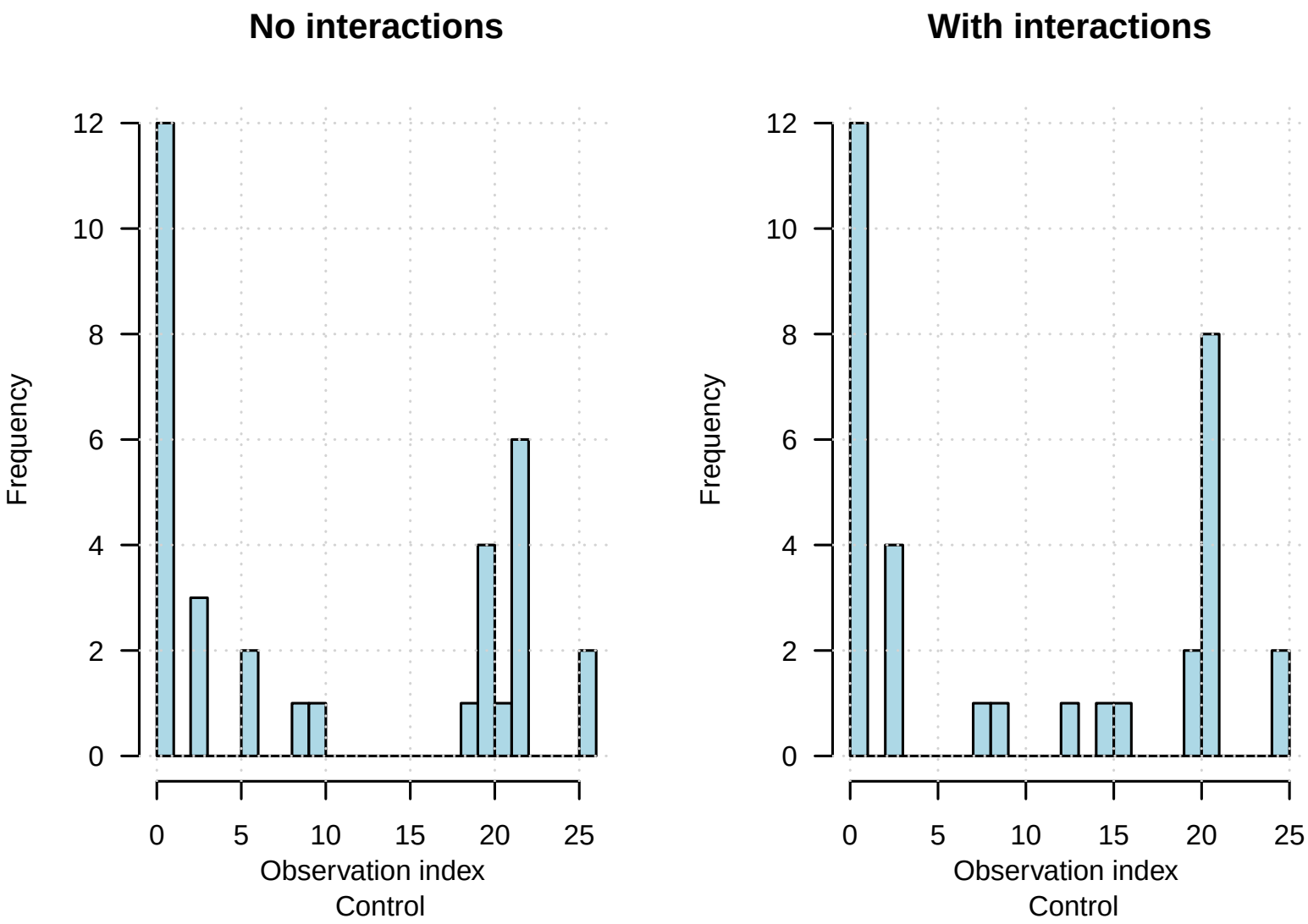
Welders and DNA: P-score after matching



Welders and DNA: P-score after matching

- If model 1 (no interactions, additive terms) is correct, we can make the treatment and control comparable in terms of $\hat{e}(\mathbf{x})$ using NN matching. KS test for differences in the P-score, $P - value = 0.9829$
- If model 2 (including interactions) is correct, the NN matching procedure dramatically improves the $\hat{e}(\mathbf{x})$ balance. However it does not completely resolves the problem. KS test for differences in the P-score, $P - value = 0.09493$
- Another important issue is: How many control observations where used multiple times? If one unit from the control group was matched to all the units in the treatment group is it a problem?

Welders and DNA: Replacement frequency



Welders and DNA: Treatment effect estimation

	No matching	No matching	Matching (P-score model 1)	Matching (P-score model 1)
(Intercept)	1.17***	1.38	1.04**	1.05
	(0.16)	(0.83)	(0.30)	(7.33)
treat	0.68**	0.63*	0.39	0.07
	(0.24)	(0.27)	(0.39)	(1.62)
age		−0.00		−0.00
		(0.02)		(0.15)
black		−0.32		0.10
		(0.36)		(0.61)
smoker		0.01		0.60
		(0.27)		(0.54)
R^2	0.15	0.17	0.05	0.12
Adj. R^2	0.13	0.09	-0.00	-0.10
Num. obs.	47	47	21	21
RMSE	0.83	0.85	0.86	0.90

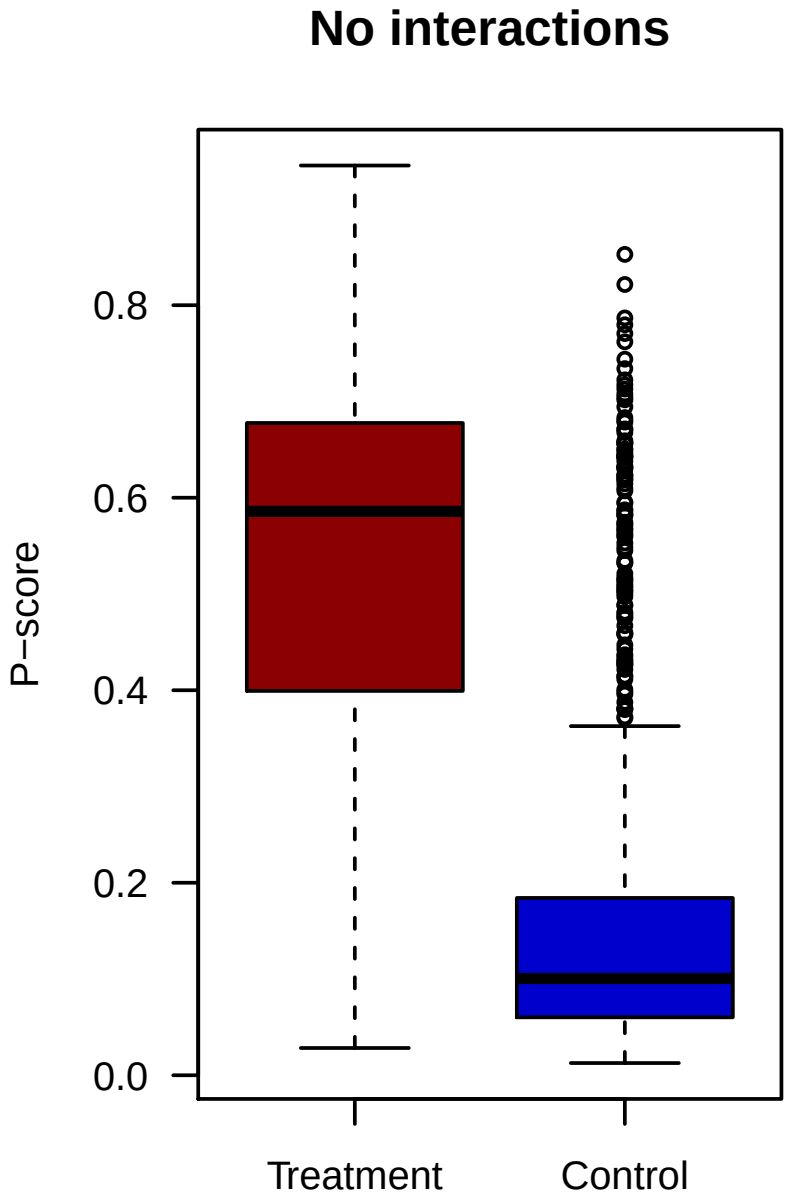
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

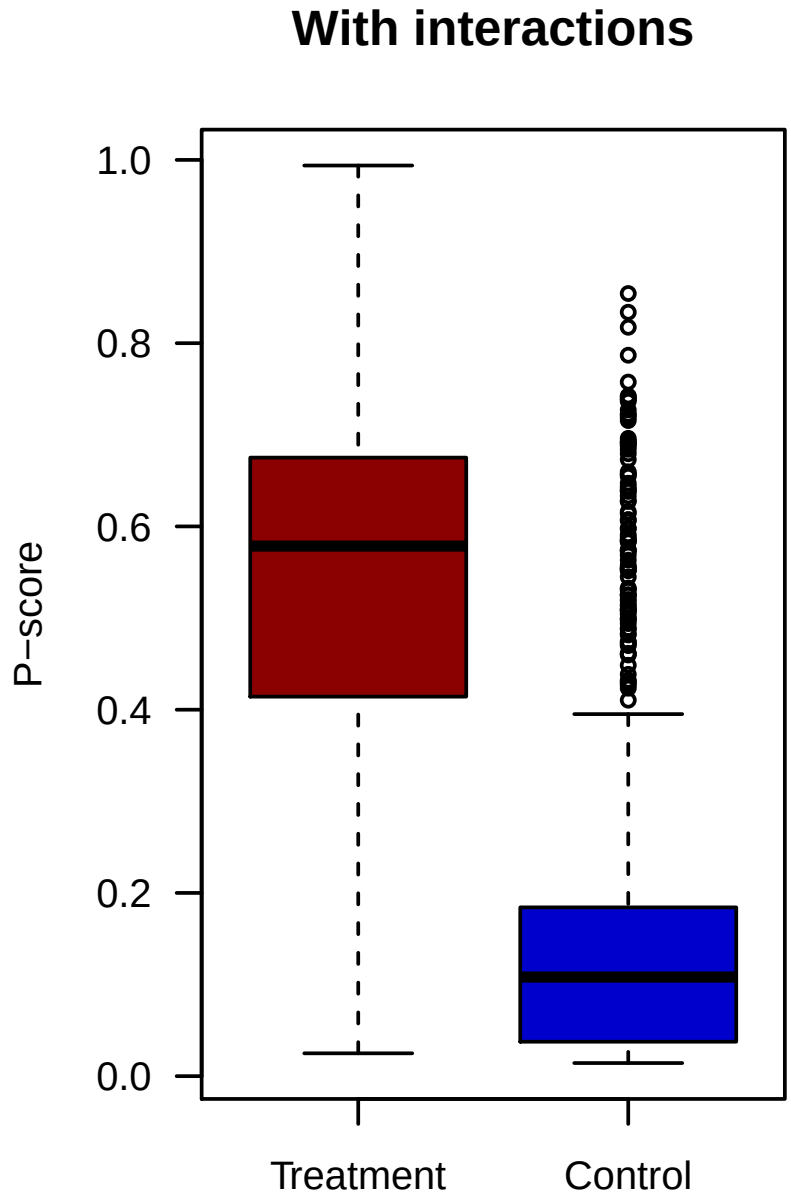
Example: Estimating the return for a college degree using PSID

- The data set is from Mroz (1987) which estimated the labor supply elasticity of working women, and evaluated the selection in to working
- Our objective is to estimate the return for college. Are college graduates similar in their observed characteristics to non-college graduates? We can test this in the data

Balance table

	Ave. College	Ave. Non-college	T-test	Wilcoxon	KS
age	41.717	42.860	0.085	0.064	0.216
k5	0.330	0.201	0.010	0.041	0.685
k618	1.245	1.396	0.138	0.297	0.571
lfp	0.679	0.525	0.000	0.000	0.001
hc	0.807	0.229	0.000	0.000	0.000
inc	25.283	18.109	0.000	0.000	0.000


No interactions
P−score
0.0
0.2
0.4
0.6
0.8
Treatment
Control


With interactions
P−score
0.0
0.2
0.4
0.6
0.8
1.0
Treatment
Control

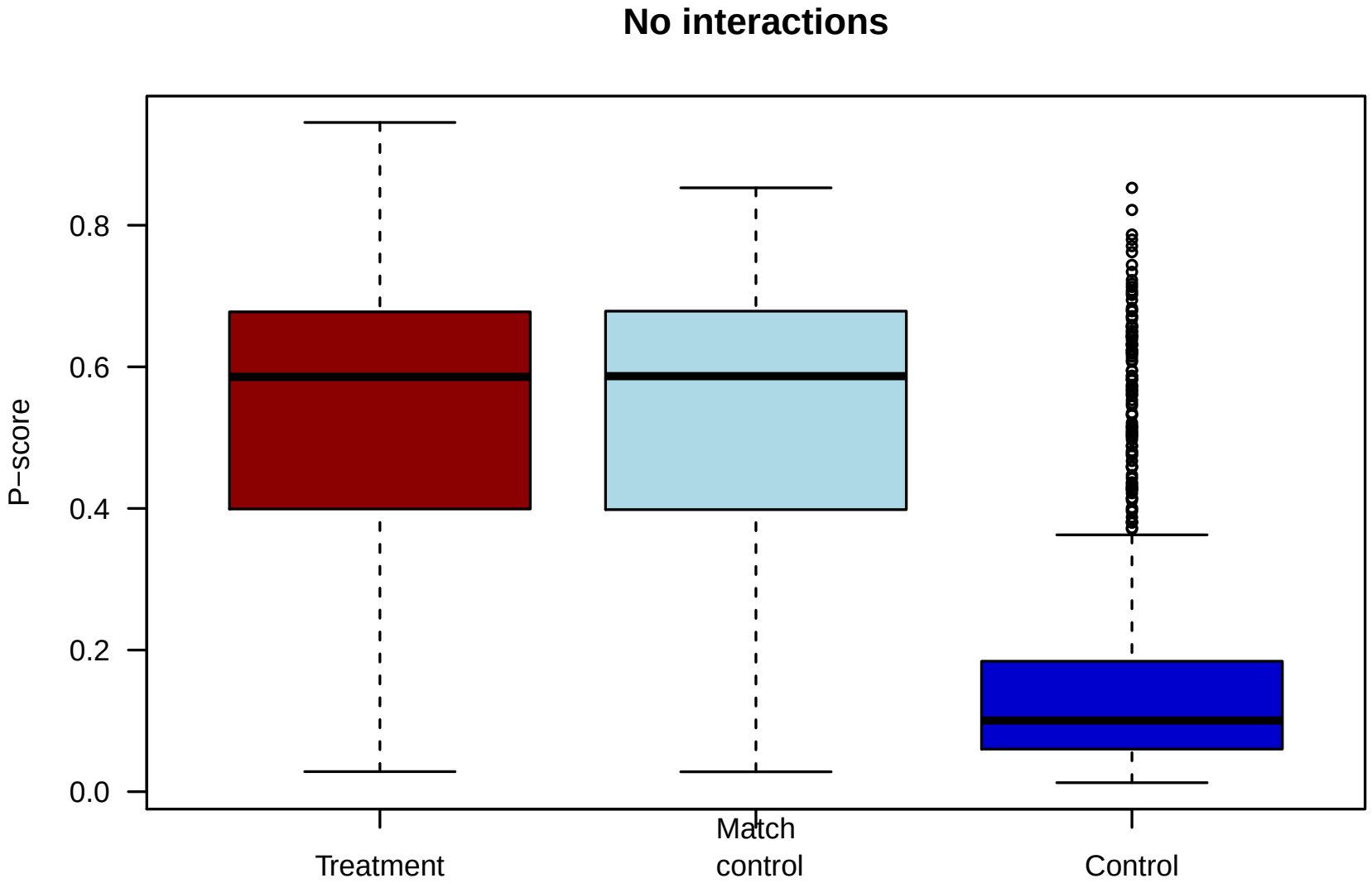
	Full sample	Only working women
(Intercept)	0.8946^{***}	1.0775^{***}
	(0.1572)	(0.2400)
lfp	0.1768^{***}	
	(0.0435)	
k618	-0.0414^{*}	-0.0576^{*}
	(0.0167)	(0.0283)
age	-0.0002	-0.0005
	(0.0031)	(0.0051)
wc	0.3419^{***}	0.3580^{***}
	(0.0540)	(0.0850)
hc	0.0147	-0.0562
	(0.0511)	(0.0847)
R^2	0.1290	0.0791
Num. obs.	753	428

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Example: PSID

- We will use the full additive no interaction specification of the p-score estimation
- Now perform nearest neighbour matching with replacement
- Check the number of replacement used from the control, how many times each observation in the control was used:

```
> summary(as.numeric(table(index.control1)))
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.767   2.000   7.000
```


No interactions
P−score
0.0
0.2
0.4
0.6
0.8
Treatment
Match control
Control

	Full sample (After matching)	Only working women (After matching)
(Intercept)	1.3975^{***}	1.6881^{***}
	(0.2417)	(0.2822)
lfp	0.1984^{**}	
	(0.0748)	
k618	-0.0772^{**}	-0.0862^{*}
	(0.0268)	(0.0353)
age	-0.0072	-0.0094
	(0.0046)	(0.0060)
wc	0.1827^{*}	0.1433
	(0.0785)	(0.0968)
hc	0.0341	0.0542
	(0.0769)	(0.0964)
R^2	0.0721	0.0506
Num. obs.	424	327

Clustering by individuals to adjust standard errors (Full sample)

```
> coeftest(lm.wc,vcov=cluster.vcov(lm.wc,dc$index ))

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.3975035  0.2674885  5.2245 2.767e-07 ***
lfp          0.1984383  0.0546547  3.6308 0.0003178 ***
k5          -0.0780152  0.0493726 -1.5801 0.1148371
k618        -0.0771820  0.0364080 -2.1199 0.0346046 *
age         -0.0071859  0.0048435 -1.4836 0.1386711
wc           0.1826633  0.0689549  2.6490 0.0083798 **
hc           0.0340553  0.0616394  0.5525 0.5809076
inc          0.0033307  0.0012389  2.6885 0.0074657 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Smith and Todd (2005) critique

- Smith and Todd (2005) critic propensity score matching, and argue it is sensitive to the specification of the propensity score.
- ST argue in favour of a *difference-in-difference matching estimator* (DDM),

$$\hat{\tau} = \frac{1}{n_t} \sum_{i \in \{T_i = 1\}} (Y_{it} - Y_{it'}) - \frac{1}{n_t} \sum_{i \in \{T_i = 0,\, M = 1\}} (Y_{it} - Y_{it'})$$

  where n_t is the number of observations in the treatment group, M is an indicator whether the observation is in the control matched data, and T is an indicator for treatment assignment.
- $\hat{\tau}$ is an estimator for the ATT

Smith and Todd (2005) critique

- The identifying assumption for this estimator to be unbiased is,

$$\mathbb{E}[Y_t(0) - Y'_t(0)|D = 1, P] = \mathbb{E}[Y_t(0) - Y'_t(0)|D = 0, P]$$

- This is slightly weaker than the CIA given the true propensity score (P),

$$Y(0), Y(1) \perp T|P$$

- This difference in the identifying assumptions is not large, however as ST argue there might be an estimation advantage for the DDM
- The DDM can also control for some unobserved factors