6. Propensity Score Matching

Econometrics II
Winter 2019
Osaka U

Shuhei Kitamura

Summary of the last lecture

Fixed effects (FE) models

- FE or RE?

Difference-in-differences (DID)

Outline

Probit and logit models

Propensity score matching (PSM)

Linear model with binary outcome

Suppose you estimate a model

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \tag{1}$$

Where Y_i is binary $\{0, 1\}$.

- E.g., Y_i is female labor participation and X_i is the number of kids. The sample constitutes only women.

Running OLS means that you are essentially computing the probability of female labor participation conditional on X_i.

$$\text{Prob}[Y_i = 1|X_i] = \alpha + \beta X_i, \tag{2}$$

And $\text{Prob}[Y_i = 0|X_i] = 1 - \text{Prob}[Y_i = 1|X_i]$.

- This $\text{Prob}[Y_i = 1|X_i]$ is sometimes called the *response probability*.
- The model assumes that the response probability is linear.
- It is thus called the *linear probability model* (LPM).

E.g., $\hat{\beta} = -0.3$ implies that one additional child decreases the probability of female labor participation by 0.3 ($=$ 30 pp).

Linear model with binary outcome (cont.)

LPM is easy to estimate and interpret, but

- You may get an estimate outside of $[0, 1]$.
- The effect of increasing child from x to $x + 1$ is assumed to be the same for any x.

  Q. Is the effect of an increase in the number of children from 1 to 2 the same as that of an increase from 2 to 3?

- There is heteroscedasticity. You should always use a robust standard error.

Nonlinear model

Consider a model

$$\text{Prob}[Y_i = 1|X_i] = F(\alpha + \beta X_i), \tag{3}$$

Where F is a function such that $0 < F(z) < 1$ for all real numbers z.

The model does not assume that the response probability is linear.

What kind of shape do we assume for F?

Nonlinear model (cont.)

Logit models assume

$$F(z) = \Gamma(z) = \frac{e^z}{1+e^z}. \tag{4}$$

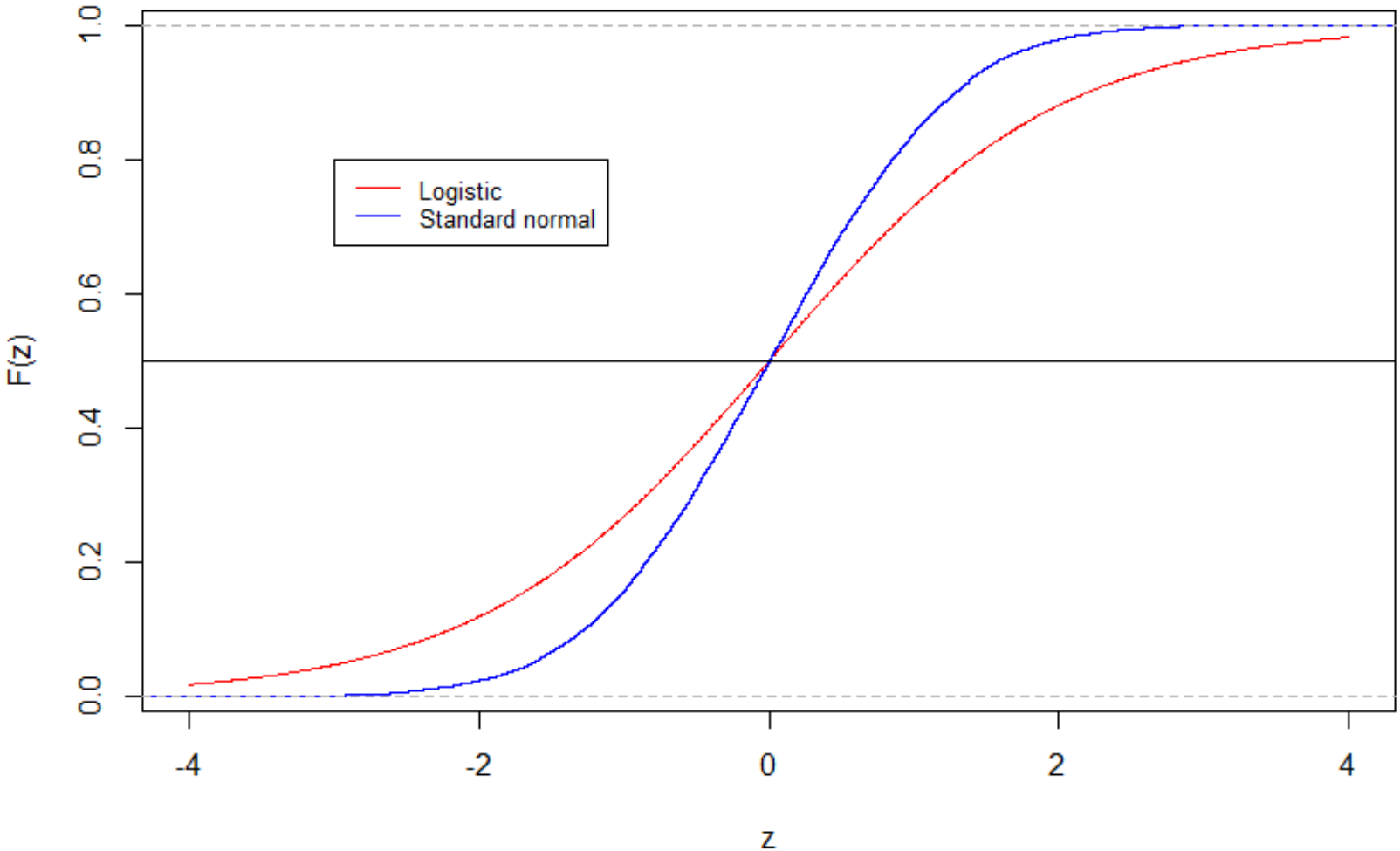
Where Γ is the cumulative distribution function (CDF) of the (standard) logistic distribution.

Probit models assume

$$F(z) = \Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt, \tag{5}$$

Where Φ is the CDF of the standard normal distribution.

Logistic vs. standard normal CDF



Nonlinear model (cont.)

You can derive these models from a *latent variable model*

$$Y_i^* = \alpha + \beta X_i + \varepsilon_i, \tag{6}$$

Where

$$Y_i = \begin{cases} 1 & \text{if } Y^* > 0 \\ 0 & \text{if } Y^* \leq 0, \end{cases} \tag{7}$$

And ε_i has either the standard normal distribution or the logistic distribution.

- For example, Y^* is the net value of working. If it is positive, we observe labor participation $Y_i = 1$.

Nonlinear model (cont.)

Then

$$\begin{aligned} \text{Prob}[Y_i = 1|X_i] &= \text{Prob}[Y^* > 0|X_i] && (8)\\ &= \text{Prob}[-(\alpha+\beta X_i) < \varepsilon_i|X_i] && (9)\\ &= 1 - F(-(\alpha+\beta X_i)) && (10)\\ &= F(\alpha+\beta X_i), && (11) \end{aligned}$$

Where I used the feature that symmetrically distributed about zero $\rightarrow F(z) = 1 - F(-z)$.

This expression is the same as (3).

Maximum likelihood estimation

We use the maximum likelihood estimation (MLE) to estimate the probit/logit model.

First, get the *likelihood function* for observation i.

$$L_i(\alpha,\beta) = f(Y_i|X_i;\alpha,\beta) = F(\alpha+\beta X_i)^{Y_i}[1-F(\alpha+\beta X_i)]^{1-Y_i}, \tag{12}$$

Where f is the probability density function.

This implies

$$\begin{cases} F(\alpha+\beta X_i) & \text{if } Y_i = 1 \\ 1-F(\alpha+\beta X_i) & \text{if } Y_i = 0. \end{cases} \tag{13}$$

The likelihood function for the entire sample is written

$$L(\alpha,\beta) = \Pi_{i=1}^{N} F(\alpha+\beta X_i)^{Y_i}[1-F(\alpha+\beta X_i)]^{1-Y_i}. \tag{14}$$

Maximum likelihood estimation (cont.)

Take the logarithm

$$l(\alpha,\beta) = \sum_{i=1}^{N} \{Y_i log(F(\alpha+\beta X_i)) + (1-Y_i) log(1-F(\alpha+\beta X_i))\}, \tag{15}$$

Which is called the *log-likelihood function*.

Probit/logit estimator is the (α,β) that maximizes this function.

- Roughly speaking, you find (α,β) that give the distribution that maximizes the probability of observing the data.
- If the distribution is symmetric, the maximum probability is found when data points are closer to the mean value. In other words, you are essentially minimizing the distance between the mean value and data points. If the distribution is normal, maximizing the likelihood and minimizing the sum of squared residuals are identical.

R exercise

Let's estimate a model of female labor participation using logit and probit estimation.

Data are taken from Mroz (1987) "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions." *Econometrica*, 55 (4), 765-799.

- The data include 753 women between ages 30-60 in 1975, with 428 working at some point during the year.

R exercise (cont.)

Launch RStudio.

Type

```
mroz <- wooldridge::mroz

reg1 <- lm(inlf ~ kidslt6 + kidsge6 + age + educ +
nwifeinc, data=mroz)
cov <- vcovHC(reg1, type = "HC0")
robust.se <- sqrt(diag(cov)

stargazer(reg1, reg1, se=list(NULL, robust.se),
column.labels=c("default","robust"), type="text")
```

Where `inlf` is the labor participation dummy, `kidslt6` and `kidsge6` are the number of kids less than 6 years old and between 6-18 years old, respectively, and `nwifeinc` husband's earnings (thousands of dollars).

R exercise (cont.)

The LPM indicates that increasing the number of kids less than 6 years old by one reduces the probability of labor participation by 0.3.

The model assumes the effect for any additional kid is the same.

- With an increase by 4, the probability is decreased by 1.19.
- You can check this by typing `reg1$coef[2]*4`.

The model does not guarantee that the predicted probabilities are contained between zero and one.

- Check the number of predicted probabilities outside the range by `length(which(fitted(reg1)<0 | fitted(reg1)>1))`.

R exercise (cont.)

How about logit and probit models?

Type

```
reg2 <- glm(inlf ~ kidslt6 + kidsge6 + age + educ +
nwifeinc, family=binomial(link="logit"), data=mroz)
reg3 <- glm(inlf ~ kidslt6 + kidsge6 + age + educ +
nwifeinc, family=binomial(link="probit"), data=mroz)

stargazer(reg1, reg2, reg3, se=list(robust.se, NULL,
NULL), type="text")
```

We get -1.45 for logit and -0.89 for probit.

How can we interpret the estimates?

Interpreting estimates

For continuous variables: take the first derivative

$$\frac{\partial P[Y=1|X]}{\partial X} = f(\alpha + \beta X) \times \beta \tag{16}$$

Where f is a probability density function.

- This captures the *partial (marginal) effect* of X on the response probability.
- Why is it called the partial effect? Consider $X = (X_1, X_2)$ and take the partial derivative with respect to X_1.

The expression (16) indicates that you need to scale up β before interpreting estimates!

- Since the sign is always the same as that of β, you can interpret the sign even without scaling it up.

Interpreting estimates (cont.)

A possible partial effect is

$$f(\hat{\alpha} + \hat{\beta}\bar{X}) \times \hat{\beta}, \tag{17}$$

Where $\bar{X}$ is the average value of X_i. That is, you compute the partial effect of the average person.

- This is called the *partial effect at the average*.

Alternatively, you can compute the *average partial effect* (*average marginal effect*)

$$\frac{1}{n}\sum_{i=1}^{N} f(\hat{\alpha} + \hat{\beta}X_i) \times \hat{\beta}. \tag{18}$$

How can we compute scale factors?

Interpreting estimates (cont.)

One can use the following

For logit

$$f(z) = \frac{e^z}{(1+e^z)^2} = F(z)(1-F(z)). \tag{19}$$

For probit

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}. \tag{20}$$

(For probit $f(0) \approx 0.4$ and for logit $f(0) = 0.25$.)

Interpreting estimates (cont.)

For discrete variables, the partial effect of X is

$$F(\alpha + \beta(X+1)) - F(\alpha + \beta X). \tag{21}$$

The average partial effect is

$$\frac{1}{n}\sum_{i=1}^{N}\{F(\hat{\alpha} + \hat{\beta}(X_i+1)) - F(\hat{\alpha} + \hat{\beta}X_i)\}. \tag{22}$$

In practice, you may get a roughly the same number by applying (18) for discrete variables as well.

R exercise (cont.)

Let's compute the average partial effects.

Type

```
reg2$coef * mean(reg2$fit*(1-reg2$fit))
reg3$coef * mean(dnorm(qnorm(reg3$fit)))
```

Compare the estimates with the OLS estimate. What do you find?

R exercise (cont.)

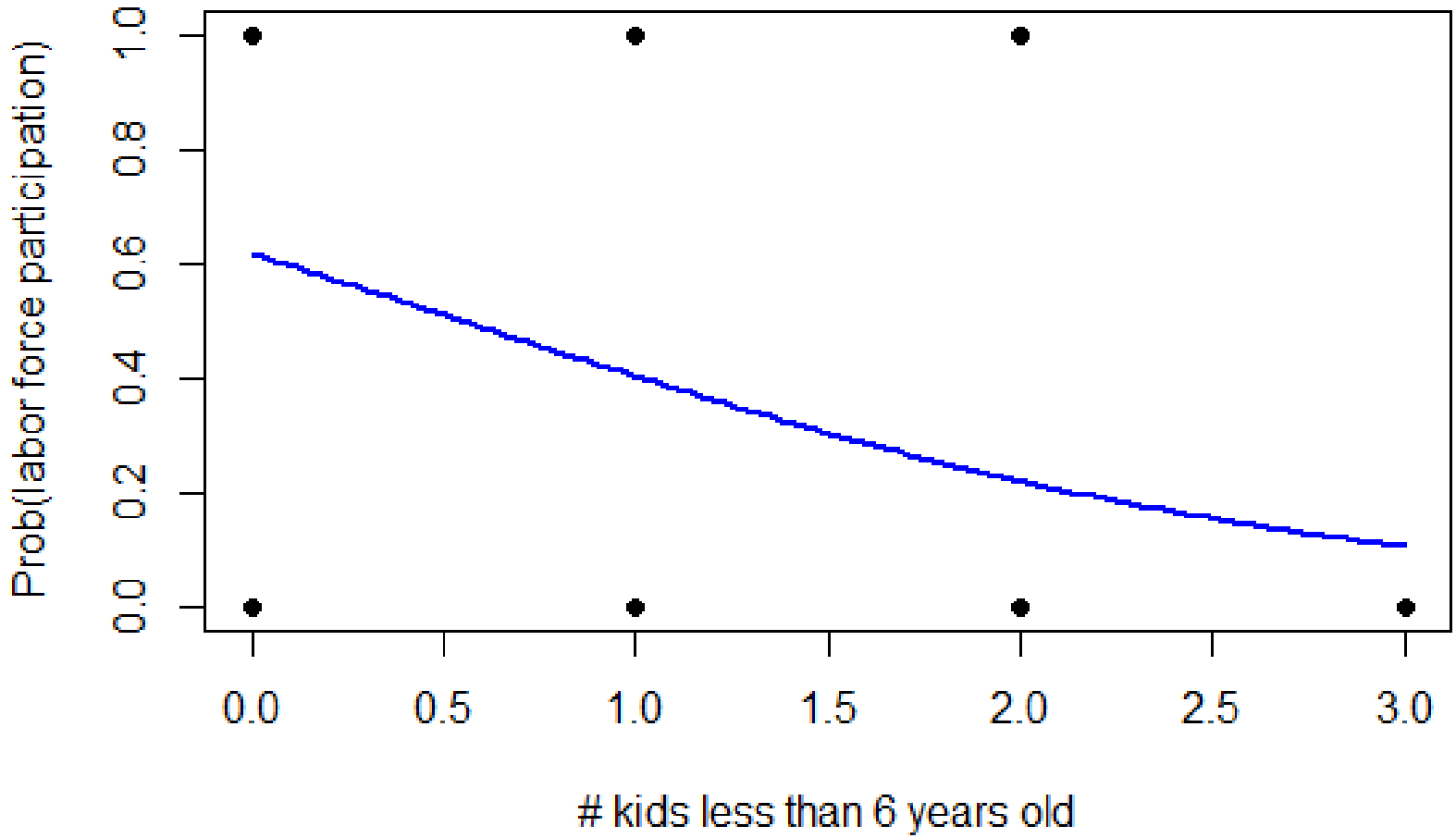
You can estimate the marginal effect at a specific value.

Type

```
margins(reg2, variables="kidslt6", at=list(kidslt6=3))
margins(reg3, variables="kidslt6", at=list(kidslt6=3))
```

Change `kidslt6` to 2, and then 1. What do you find?

Predicted response probability (logit)



Multiple hypothesis testing

There are several ways to test multiple hypotheses for MLE.

- The likelihood ratio statistic $LR = 2(L_{ur} - L_r)$, where L_{ur} and L_r are the log-likelihood value for the unrestricted model and for the restricted model, respectively. $LR \sim \chi^2_q$. q is the degree of freedom.
- The Wald statistic (F-statistic)

See Chapter 17 in Wooldridge for more details.

R exercise (cont.)

Type

```
df <- length(reg2$coef) - 1

LL <- reg2$dev/(-2)
LL0 <- reg2$null.dev/(-2)

LR <- 2*(LL-LL0)
LRp <- 1 - pchisq(LR, df=df)

LR
LRp
```

Goodness-of-fit measure

There are several goodness-of-fit measures for MLE.

A commonly used one is *pseudo R-squared* measures. One of them is the McFadden's pseudo R-squared

$$1 - \frac{L_{ur}}{L_0}. \tag{23}$$

Where L_0 is the log-likelihood value for the model with only an intercept.

R exercise (cont.)

Type

```
pR2 <- 1 - LL/LL0

pR2
```

Ordered/multinominal probit/logit

Consider an outcome which is ordered but has more than two options

$$Y_i = \begin{cases} 0 & \text{if } Y^* < 0 \\ 1 & \text{if } 0 \leq Y^* < \bar{Y} \\ 2 & \text{if } Y^* \geq \bar{Y}. \end{cases} \tag{24}$$

Then, use *ordered probit/logit*.

If the outcome is not ordered but has more than two options, then use *multinominal probit/logit*.

Regression and matching (review)

We have seen that the regression estimand can be written as the matching estimand. Let's review it.

Consider a model

$$Y_i = \alpha + \delta D_i + \gamma X_i + \varepsilon_i, \tag{25}$$

Where Y_i is earnings, D_i is job training, and X_i is sex.

The matching estimand is written as the weighted average of the treatment effect on mean earnings (δ_X) for each group (sex).

$$E[Y_i(1) - Y_i(0)|D_i = 1] = \sum_x \delta_x \text{Prob}[X_i = x|D_i = 1], \tag{26}$$

Where $\text{Prob}[X_i = x|D_i = 1]$ can be replaced by

$$\frac{I_x N_x^1}{\sum_x I_x N_x^1}. \tag{27}$$

$I_x = I(N_x^1 > 0, N_x^0 > 0)$ is an indicator variable taking value one if both $N_x^1 > 0$ and $N_x^0 > 0$, and zero otherwise.

Propensity score matching

In the above example, observations are matched based on individual characteristics, i.e., sex.

We can match observations based on the propensity score, instead of control variables, as defined by

$$p(X_i) = \text{Prob}[D_i = 1|X_i]. \tag{28}$$

Wait, didn't we see the expression before?

$\rightarrow$ We will use probit (or logit) to compute the propensity score!

Propensity score matching (cont.)

We need the common support assumption

$$0 < \text{Prob}(D_i = 1|X_i) < 1 \tag{29}$$

- For PSM, the assumption is often called the *overlap* assumption.

Recall the CIA

$$D_i|X_i \perp (Y_i(1), Y_i(0)). \tag{30}$$

One can show that if the CIA holds, then

$$D_i|P(X_i) \perp (Y_i(1), Y_i(0)) \tag{31}$$

Also holds.

- For PSM, this assumption is often called *unconfoundedness*.

Propensity score matching (cont.)

If the CIA holds, then

$$
\begin{aligned}
& E[Y_i(1) - Y_i(0) | D_i = 1] \\
= \; & E\{E[Y_i(1) - Y_i(0) | P(X_i), D_i = 1] | D_i = 1\}. \\
= \; & E\{E[Y_i(1) | P(X_i), D_i = 1] - E[Y_i(0) | P(X_i), D_i = 1] | D_i = 1\}. \\
= \; & E\{E[Y_i(1) | P(X_i), D_i = 1] - E[Y_i(0) | P(X_i), D_i = 0] | D_i = 1\}. \\
= \; & E[\delta_p | D_i = 1]. \qquad (32)
\end{aligned}
$$

Propensity score matching (cont.)

Recall the matching estimand

$$E[Y_i(1) - Y_i(0)|D_i = 1] = \sum_x \delta_x \text{Prob}[X_i = x|D_i = 1]. \tag{33}$$

For PSM, individuals are split into subpopulations based on the propensity score, and the matching estimand is the weighted average of mean difference in outcomes over common support, weighted by the propensity score distribution of individuals (e.g., (27)).

- The common support assumption ensures that observations are found for both treatment and control groups in each subpopulation.

Propensity score and regression

Let's use the job training (NSW program) experimental data to check common support.

Type

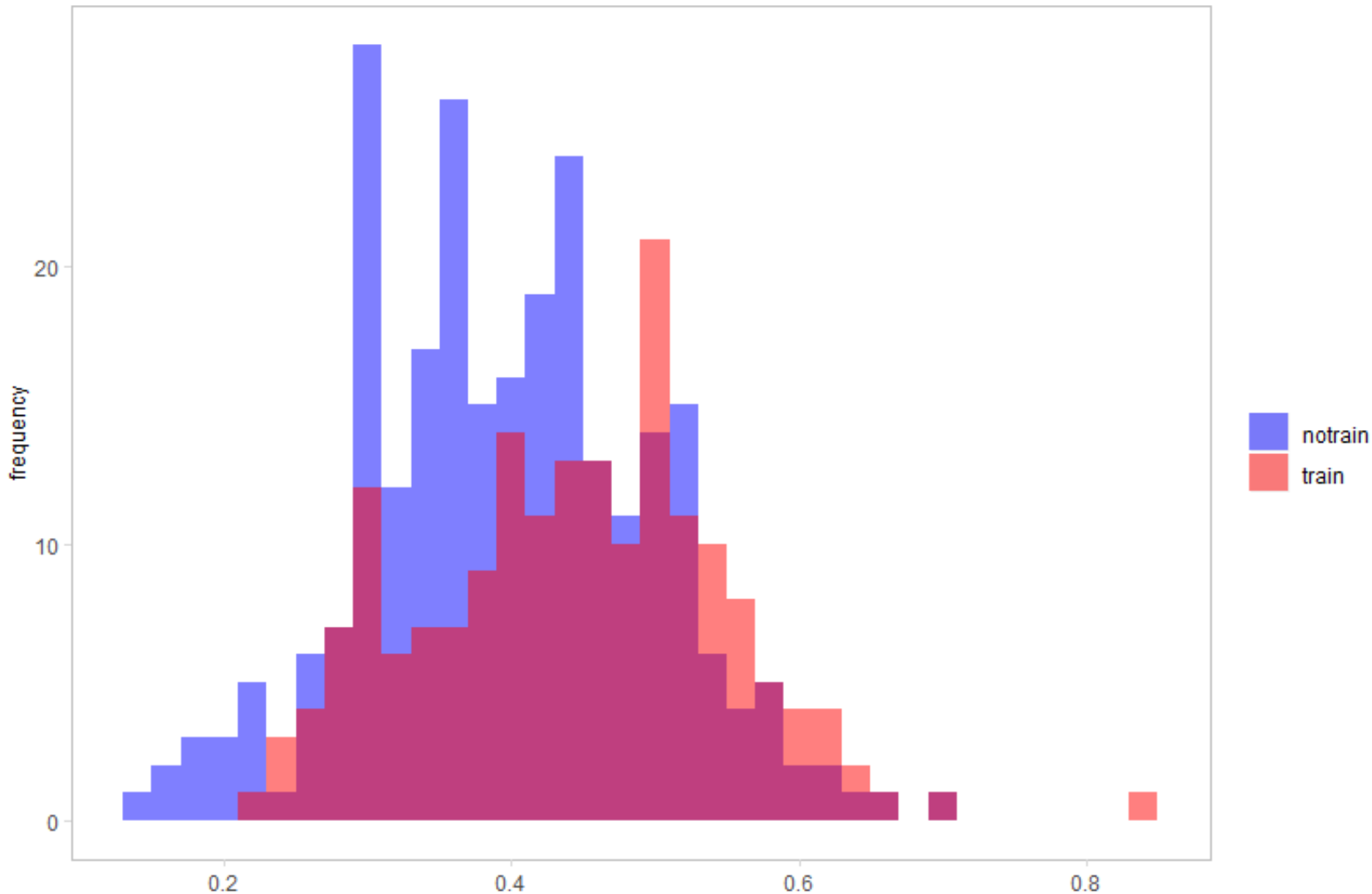
```
jt <- wooldridge::jtrain2

reg1 <- glm(train ~ age + educ + black + hisp + married
+ re74 + re75 + re78 + unem74 + unem75 + unem78,
family=binomial(link="probit"), data=jt)

jt$p1 <- reg1$fit

ggplot(data=jt, aes(x=p1, fill=factor(train))) +
    geom_histogram(binwidth=0.02, alpha=0.5,
position='identity') +
    scale_fill_manual(name="", values=c("blue", "red"),
labels=c("notrain", "train")) +
    labs(title="", x="", y="frequency")
```

Distribution of propensity score (experimental)



Propensity score and regression (cont.)

However, if you instead use the non-experimental data of the same program, you do not get a similar picture.

- As we saw earlier, estimates are also very different.

Type

```
setwd("C:/path_to_the_file")
jt2 <- read.csv("cps1re74.csv")

jt2$re78 <- jt2$re78/1000
jt2$re74 <- jt2$re74/1000
jt2$re75 <- jt2$re75/1000
```

Propensity score and regression (cont.)

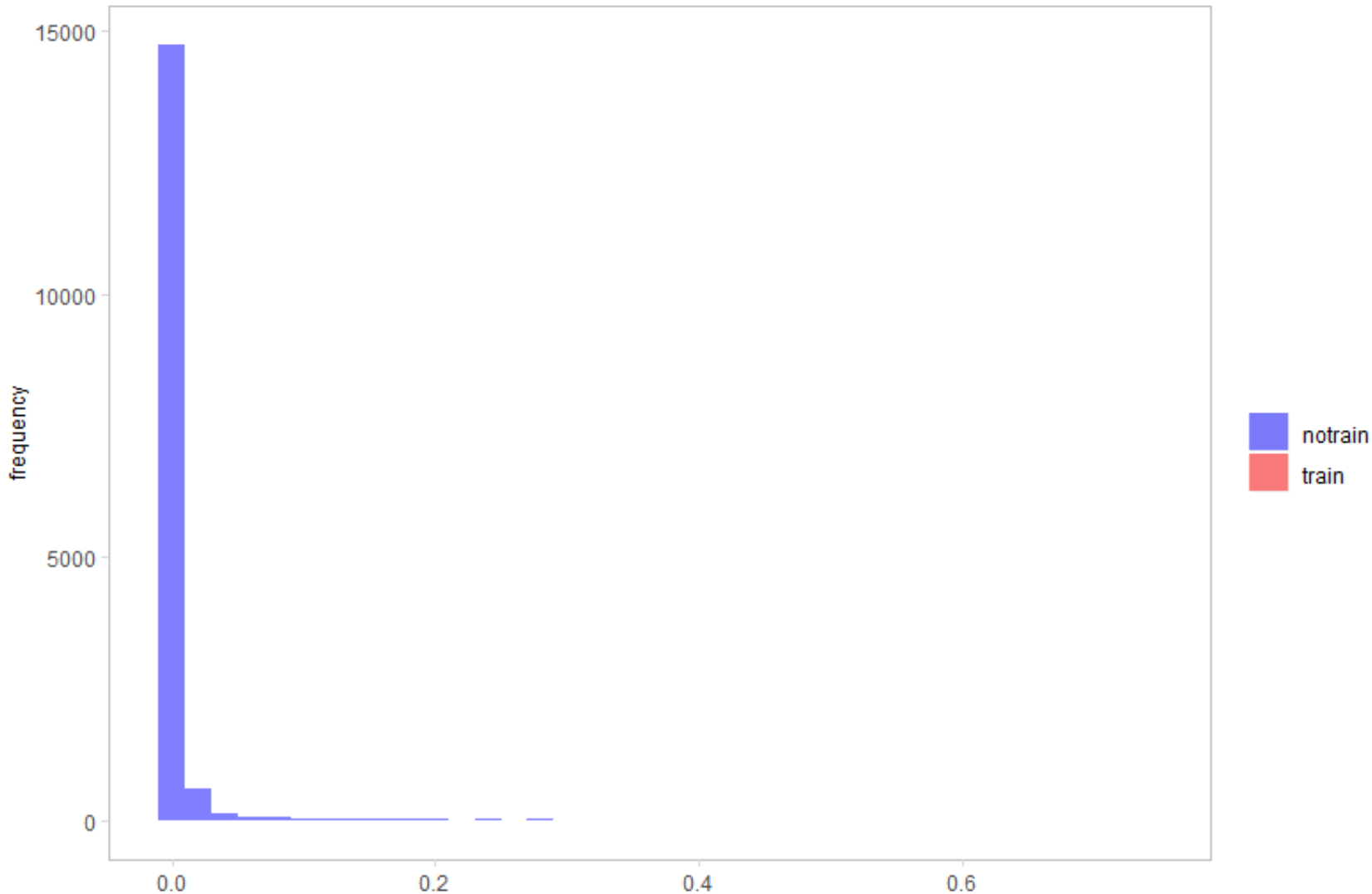
Type

```
reg2 <- glm(treat ~ age + age2 + ed + black + hisp +
nodeg + married + re74 + re75,
family=binomial(link="probit"), data=jt2)

jt2$p1 <- reg2$fit

ggplot(data=jt2, aes(x=p1, fill=factor(treat))) +
    geom_histogram(binwidth=0.02, alpha=0.5,
position='identity') +
    scale_fill_manual(name="", values=c("blue", "red"),
labels=c("notrain", "train")) +
    labs(title="", x="", y="frequency")
```

Distribution of propensity score (non-experimental)



Propensity score and regression (cont.)

Let's make groups more comparable.

Crump et al. (2009) suggest that the propensity score can be used for systematic sample selection.

We first run a probit or logit model to predict probabilities.

Then use only samples with predicted probabilities of treatment being included in $(0.1, 0.9)$.

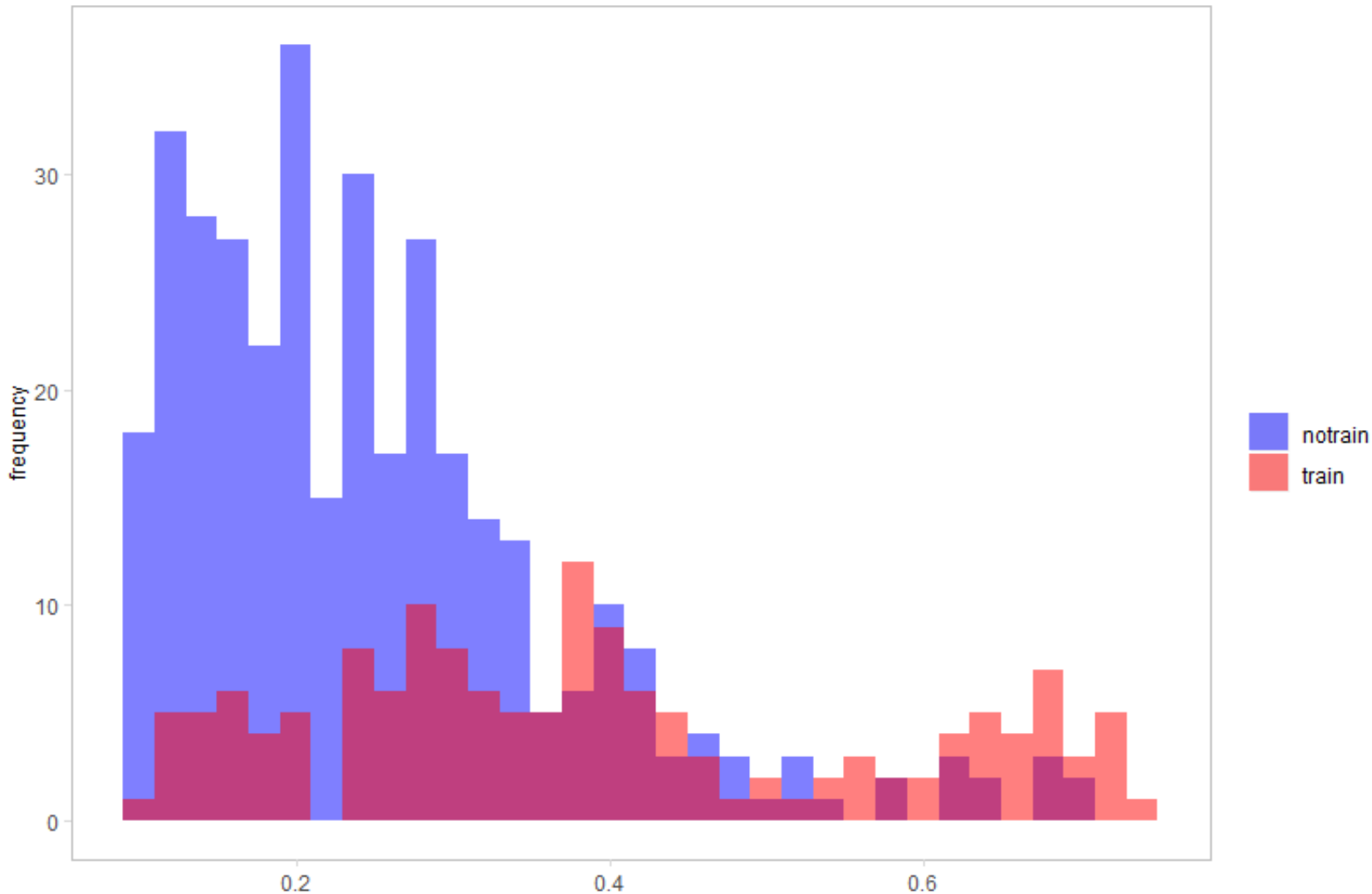
- By doing so, we are tyring to drop observations which do not have common support.
- Another option is to keep observations whose scores fall within the maximum and minimum values of scores for the treatment (or control) group.

Propensity score and regression (cont.)

Type

```
ggplot(data=jt2[(jt2$p1>0.1 & jt2$p1<0.9),], aes(x=p1,
fill=factor(treat))) +
    geom_histogram(binwidth=0.02, alpha=0.5,
position='identity') +
    scale_fill_manual(name="", values=c("blue", "red"),
labels=c("notrain", "train")) +
    labs(title="", x="", y="frequency")
```

Distribution of propensity score (cont.)



Propensity score and regression (cont.)

This looks better.

Are covariates balanced?

Type

```
res1 <- compareGroups(train ~ age + agesq + educ + black
+ hisp + nodegree + married + re74 + re75, data = jt)
res2 <- compareGroups(treat ~ age + age2 + ed + black +
hisp + nodeg + married + re74 + re75, data = jt2)
res3 <- compareGroups(treat ~ age + age2 + ed + black +
hisp + nodeg + married + re74 + re75, data = jt2,
subset=(p1>0.1 & p1<0.9))

createTable(res1)
createTable(res2)
createTable(res3)
```

This is a bit rough comparison, as we are interested in checking a balance between the treatment and control groups within each subpopulation.

Propensity score and regression (cont.)

How about regression estimates?

Type

```
reg1 <- lm(re78 ~ train + age + agesq + educ + black +
hisp + nodegree + married + re74 + re75, data=jt)
reg2 <- lm(re78 ~ treat + age + age2 + ed + black + hisp
+ nodeg + married + re74 + re75, data=jt2)
reg3 <- lm(re78 ~ treat + age + age2 + ed + black + hisp
+ nodeg + married + re74 + re75, data=jt2,
subset=(p1>0.1 & p1<0.9))

stargazer(reg1, reg2, reg3, type="text")
```

Propensity score matching (cont.)

So far we used the propensity score to select samples for OLS.

They can also be used directly for computing estimates.

There are several methods.

- Matching on the estimated score (or PSM)
- Using a weighting scheme

Common matching methods include

- One-to-one matching
- Nearest neighbor matching
- Kernel matching
- Radius matching
- Stratification matching

Matching on the estimated score

The first example uses a stratification method used by Dehejia and Wahba (1999).

- Stratify individuals in the treatment and control groups based on propensity scores.
- Compute within-stratum difference in means between the treatment and control groups.
- Sum them up where the sum is weighted by the share of treated individuals within each stratum.

R exercise

Type

```
jt2 <- read.csv("cps1re74.csv")

jt2$re78 <- jt2$re78/1000
jt2$re74 <- jt2$re74/1000
jt2$re75 <- jt2$re75/1000

jt2$ed2 <- jt2$ed^2
jt2$ed.re74 <- jt2$ed * jt2$re74
jt2$age3 <- jt2$age^3

jt2$u74 <- ifelse(jt2$re74==0,1,0)
jt2$u75 <- ifelse(jt2$re75==0,1,0)

reg2 <- glm(treat ~ age + age2 + age3 + ed + ed2 +
ed.re74 + black + hisp + nodeg + married + re74 + re75 +
u74 + u75, family=binomial(link="logit"), data=jt2)
jt2$p1 <- reg2$fit
```

R exercise (cont.)

Type

```
jt2.subset <- subset(jt2, p1>min(p1[treat==1]) &
p1<max(p1[treat==1]))

jt2.subset$pcentile <- with(jt2.subset, cut(p1,
breaks=quantile(p1, probs=seq(0,1,by=0.02))))

jt2.subset$pcentile <- as.numeric(jt2.subset$pcentile)
```

R exercise (cont.)

Type

```
q = c(); n = c()

for (i in 1:50) {
    q[i] <- with(jt2.subset, mean(re78[pcentile==i &
treat==1], na.rm = TRUE)) - with(jt2.subset,
mean(re78[pcentile==i & treat==0], na.rm = TRUE))
    n[i] <- with(jt2.subset, length(re78[pcentile==i &
treat==1]))
}

q[is.na(q)] <- 0; n[is.na(n)] <- 0

stratified <- t(q)%*%n/sum(n)

stratified
```

Using a weighting scheme

The second example uses the weighted average estimand by Hirano et al. (2003).

$$E\left\{g(X_i)\left[\frac{Y_i D_i}{p(X_i)} - \frac{Y_i(1-D_i)}{(1-p(X_i))}\right]\right\}, \tag{34}$$

Where $g(X_i)$ is a weighting function.

- $g(X_i) = 1$ for ATE
- $g(X_i) = \frac{p(X_i)}{P(D_i=1)}$ for ATT

R exercise (cont.)

Type

```
att <- with(jt2, mean(p1/mean(treat)*(re78*treat/p1 -
re78*(1-treat)/(1-p1))))

att
```

Summary

Probit/logit model

Propensity score matching (PSM)