Intermediate Econometrics

05/12/2025 - Vincenzo Gioia

IV estimator on binomial models

Endogeneity

- Some regressors may be **endogenous** and this creates:
  - Biased estimates
  - Inconsistent coefficient estimates
  - Wrong causal interpretation
- Sources (Omitted variables, Reverse causality, Measurement error, Selection bias)
- In **linear models** the two Stage Least Square works
- In **binary models**, nonlinearity requires special methods

IV estimator on binomial models

Endogeneity

- We focus exclusively on the **IV-Probit framework**, because the **probit model** is:
  - Mathematically convenient
  - Fully compatible with IV methods
  - The **standard reference in the econometrics literature**
- A IV estimator for logit:
  - Is not standard
  - Requires complex structural assumptions
  - Less commonly used in core applied research

Probit IV estimator on binomial models

The Basic Idea of IV-Probit

- There exists an **unobserved latent variable**
- We only observe:
  - 1 = decision taken
  - 0 = decision not taken
- In formulas, we observe a **binary outcome**, y_i, while the latent variable y_i^* represents the underlying economic decision

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

Probit IV estimator on binomial models

The Basic Idea of IV-Probit

- The model is

$$y_i^* = x_i^\top \beta + \varepsilon_i$$

- We split $x_i^\top$ into two parts :
  - w_i: exogenous regressors
  - z_i = endogenous regressors
- While ε_i is the structural error
- This leads the model to be written as

$$y_i^* = w_i^\top \alpha + z_i^\top \gamma + \varepsilon_i$$

Probit IV estimator on binomial models

First-Stage (Reduced-Form) Equations

- Each endogenous regressor is generated by:

$$z_{gi} = \pi_g^\top u_i + v_{gi}$$

  - $u_i = (w_i, r_i)$
  - r_i = **external instruments**
  - v_{gi} = first-stage error
- Endogeneity is captured by **correlation between the errors**:

$$\mathrm{Cov}(\varepsilon_i, v_i) \neq 0$$

- This implies:
  - Standard probit is **inconsistent**
  - Instruments are needed to **break this correlation**

Probit IV estimator on binomial models

Conditional Mean and Control Function

- Under joint normality of $(\varepsilon_i, v_i)^\top$, the conditional mean of the latent outcome is:

$$\mathrm{E}(y_i^* \mid w_i, z_i, u_i) = \gamma^\top x_i + \rho^\top v_i$$

- ρ measures the **strength of endogeneity**

Probit IV estimator on binomial models

Control Function IV-Probit

- This leads to the **Control Function Probit**:

$$P(Y_i = 1 \mid w_i, z_i, \hat{v}_i) = \Phi(\gamma^\top x_i + \rho^\top \hat{v}_i)$$

  - $\hat{v}_i$ are residuals from the first stage
  - ρ captures endogeneity
- Thus, we can test:

$$H_0 : \rho = 0$$

  - If rejected → endogenous regressor is truly endogenous
  - If not rejected → standard probit is sufficient
- This is the **Rivers–Vuong Two-Step IV-Probit Estimator**

Probit IV estimator on binomial models

IV Estimation Strategies in the Probit Model

- Within the **probit framework**, we can use the following approaches:

1. **Maximum Likelihood (ML-IV Probit)**
2. **Two-Step / Control Function (Rivers–Vuong)**
3. **Minimum Chi-Square (Newey)**

- These are the **standard IV estimators for probit models with endogeneity**

Probit IV estimator on binomial models

Method 1: Maximum Likelihood (ML)

- **Pros**
  - Highly efficient
  - Best statistical properties
- **Cons**
  - Very complex
  - Many parameters
  - Sensitive to misspecification
- Used mainly in advanced research.

Probit IV estimator on binomial models

Method 2: Two-Step / Control Function

- **Step 1**
  - Regress endogenous variables on instruments
  - Save the residuals
- **Step 2**: Run probit using:
  - Original regressors
  - Plus residuals
- **Key Intuition**
  - Residuals capture endogeneity
  - If they matter, then endogeneity exists
- **Pros**:
  - Direct economic intuition
  - Easy to implement
  - Allows explicit testing for endogeneity
  - Same logic as 2SLS in linear models
- **Cons**: Slightly less efficient than full ML

Probit IV estimator on binomial models

Method 3: Minimum Chi-Square

- Combines several intermediate estimates
- Chooses parameters that minimize statistical distance
- **Pros**
  - More efficient than Two-Step
- **Cons**:
  - Highly technical
  - Rare in empirical practice

Probit IV estimator on binomial models

Empirical Example: US Banks and Derivatives

- **Research Question**: What drives banks to use FX derivatives?
- **Dependent Variable**: Use of derivatives: Yes / No
- **Key Explanatory Variables**
  - Managerial ownership
  - Bonus
  - Stock options
  - Financial leverage

Probit IV estimator on binomial models

Empirical Example: US Banks and Derivatives

- This example is from Adkins, Carter, and Simpson (2007) and Adkins (2012)
- These authors analyzed the effect of managerial incentives on the use of foreign-exchange derivatives for hedging by U.S. bank holding companies, for the 1996-2000 period.
- The dependent variable `federiv` is 1 if the bank uses foreign-exchange derivatives.
- The first set of covariates concerns **Manager ownership**. When managers have a higher ownership position in the bank, their behavior is more in line with the preferences of shareholders, and they therefore have an incentive to take risk
  - The logarithm of the percentage of total shares outstanding that are owned by officers and directors (`linsown`) should therefore have a negative effect on the probability of using foreign-exchange derivatives. However, incentives provided by regulation may dominate the expected incentive relation and lead to a negative effect on the probability
  - On the contrary, institutional blockholders have imperfect information and, therefore, the logarithm of the percentage of total shares outstanding that are owned by all institutional investors (`linstown`) should have a negative effect on the probability of using foreign-exchange derivatives

Probit IV estimator on binomial models

Empirical Example: US Banks and Derivatives

- The second set of covariates concerns CEO compensation:
  - Value of option awards (`optval`) should induce managers to take more risk and therefore should have a negative effect on the probability
  - On the contrary, cash bonus (`bonus`) may increase the probability of hedging in order to decrease variability in the firm's cash flows
- The other covariates are:
  - the leverage (`eqrat`)
  - the logarithm of total assets (`ltass`)
  - the return on equity (`roe`)
  - the market to book ratio (`mktbt`)
  - the foreign to total interest income ratio (`perfor`)
  - a derivative dealer activity dummy (`dealdum`)
  - dividends paid (`div`) and the year from 1996 to 2000 (`year`).

Probit IV estimator on binomial models

Why Endogeneity Is Expected

- Compensation affects risk-taking
- Risk strategy also affects compensation
- Leverage is jointly determined with risk
- Three covariates are suspected to be endogenous:
  - the leverage (`eqrat`)
  - the option awards (`optval`)
  - the bonus, (`bonus`)
- The external instruments are:
  - the number of employees (`no_emp`)
  - the number of subsidiaries (`no_subs`)
  - the number of officies (`no_off`)
  - the CEO age (`ceo_age`)
  - the 12 month maturity mismatch (`gap`)
  - the ratio of cash flow to total assets (`cfa`)
- These affect incentives but not directly derivative use.

Probit IV estimator on binomial models

Main Results (Interpretation)

- To fit the models we use binomreg function of the micsr package

```
library(micsr)
form <- federiv ~ eqrat + optval + bonus + ltass +
                  linsown + linstown + roe + mktbk +
                  perfor + dealdum + div + year |
                  ltass + linsown + linstown + roe + mktbk +
                  perfor + dealdum + div + year + no_emp +
                  no_subs + no_off + ceo_age + gap + cfa

bank_2st <- binomreg(form, data = federiv, link = "probit",
                     method = "twosteps")
```

Probit IV estimator on binomial models

Main Results (Interpretation)

- The coefficients of linstown, bonus and optval have the expected sign. linsown has a positive sign, which must be driven by the strength of the regulatory constraints
  - **Bonuses** → Increase hedging
  - **Stock options** → Increase risk-taking → Less hedging
  - **Institutional ownership** → Less hedging
  - **Managerial ownership** → Effect reversed due to regulation

```
1 summary(bank_2st)
```

```
Two-steps estimation
               Estimate  Std. Error   z-value  Pr(>|z|)
(Intercept) -9.7201e+00  1.2183e+01   -0.7978    0.4250
ltass        3.6651e-01  1.9406e+02    0.0019    0.9985
linsown      2.5689e-01  2.7409e+01    0.0094    0.9925
linstown     3.7219e-01  4.0792e+01    0.0091    0.9927
roe         -3.3155e-02  1.9525e+02   -0.0002    0.9999
mktbk       -1.8501e-03  3.2246e+03    0.0000    1.0000
perfor      -3.4735e+00  7.6485e-01   -4.5414 5.589e-06 ***
dealdum     -2.7954e-01  7.3928e+00   -0.0378    0.9698
div         -8.3666e-01  3.7868e+00   -0.2209    0.8251
```

```
year1997    -2.4438e-02  5.7481e+00   -0.0043    0.9966
year1998    -2.4397e-01  5.6832e+00   -0.0429    0.9658
year1999    -2.3807e-01  5.8173e+00   -0.0409    0.9674
year2000    -1.2869e-01  5.5020e+00   -0.0234    0.9813
eqrat        2.1825e+01  1.0361e+00   21.0644 < 2.2e-16 ***
```

Probit IV estimator on binomial models

Endogeneity

- A Wald test that $\rho = 0$ can be performed more simply using the `miscsr::endogtest` function:
- Endogeneity is weak, but statistically significant at 10%

```
endogtest(form, federiv)
```

```
    Smith-Blundell / Rivers-Vuong test

data:  form ...
chisq = 7.5472, df = 3, p-value = 0.05636
alternative hypothesis: endogeneity
```

Matching in Observational Studies

Endogeneity

- The fundamental problem of causal inference with observational data is that:
  - Treated and control units **not** come from the same population
  - As a result:
    - Observable characteristics differ
    - Unobservable characteristics may also differ
- This leads to **selection bias** in treatment effect estimation.

Matching in Observational Studies

Basic Idea of Matching

- The idea of matching is:
  - For each **treated** observation, find a **control** observation that is as similar as possible, based on **observable characteristics**
- If matching is successful, the resulting sample:
  - Mimics a randomized experiment
  - Allows causal interpretation under selection on observables

Matching in Observational Studies

The Propensity Score

- However, with many continuous covariates, exact matching is impossible.
- Solution: **Propensity Score Matching**
- Propensity score is:

$$p(x) = P(T = 1 \mid x)$$

  - T is treatment
  - x are the covariates
- It is estimated using Logit or Probit, where the treatment indicator as dependent variable
- Each treated unit is matched with the control unit having the **closest propensity score**.

Matching in Observational Studies

Practical Matching Algorithm

1. Estimate the **propensity score** using a rich model
2. Divide observations into **strata** (e.g. 5 blocks)
3. Test equality of mean scores within each stratum
4. If the test fails → split the stratum
5. Test balance for **all covariates**
6. If balance fails → re-estimate a more flexible model

- Achieve **covariate balance** between treated and control groups

Matching in Observational Studies

Estimation of the Treatment Effect

- Once strata are defined, the treatment effect is estimated as:

$$\sum_k (\bar{y}_k^T - \bar{y}_k^C) f_k$$

```
- $\bar{y}_k^T$ = mean outcome of treated units in stratum $k$
- $\bar{y}_k^C$ = mean outcome of control units in stratum $k$
- $f_k = n_k / n_T$: frequency of each group
```

- This gives the **Average Treatment Effect on the Treated (ATT)**

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- Study of Ichino, Mealli, and Nannicini (2008)
- **Goal**: Examine the effect of temporary work agency (TWA) jobs on the probability of finding a stable job
- The data set, called `twa`, contains 2030 observations (511 treated and 1519 untreated) for two regions, Tuscany and Sicily
- Let's focus to the Tuscany obtaining a sample where `group` is the treatment variable

```
1 tuscany <- twa[twa$region == "Tuscany", ]
2 table(tuscany$group)
```

```
control treated
    628     281
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The outcome is also a factor indicating the employment status one year after the program. Its levels are `none` (no job), `other`, `fterm` for fixed-term contract and `perm` for a permanent contract. Following the authors, we define the outcome of interest as a dummy for a permanent contract
- **Outcome: Permanent contract after 1 year** (`perm`)
- To get a first idea of the treatment effect, we compute the mean of the outcome for the two groups
- The proportion of individuals who have a permanent job is 31.3% for the treated group and 16.6% for the control group and the apparent treatment effect is therefore 14.7%

```
1 tuscany$perm <- ifelse(tuscany$outcome == "perm", 1, 0)
2 cbind(control = mean(tuscany$perm[tuscany$group == "control"]),
3       treated = mean(tuscany$perm[tuscany$group == "treated"]))
```

```
      control   treated
[1,] 0.1656051 0.3131673
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The outcome is also a factor indicating the employment status one year after the program. Its levels are `none` (no job), `other`, `fterm` for fixed-term contract and `perm` for a permanent contract. Following the authors, we define the outcome of interest as a dummy for a permanent contract
- **Outcome: Permanent contract after 1 year** (`perm`)
- To get a first idea of the treatment effect, we compute the mean of the outcome for the two groups
- The proportion of individuals who have a permanent job is 31.3% for the treated group and 16.6% for the control group and the apparent treatment effect is therefore 14.7%

```
1 tuscany$perm <- ifelse(tuscany$outcome == "perm", 1, 0)
2 cbind(control = mean(tuscany$perm[tuscany$group == "control"]),
3       treated = mean(tuscany$perm[tuscany$group == "treated"]))
```

```
      control   treated
[1,] 0.1656051 0.3131673
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The `micsr::pscore` function implements the algorithm previously described. The first two arguments are:
  - `formula`: it should have two variables on the left-hand side, the first indicating the outcome and the second the group (the group variable can be either a dummy or a factor with two levels, the second indicating the treated individuals)
  - `data`
- To use the same formula of Ichino, Mealli, and Nannicini (2008) we need to consider
  - square term for the distance to the next agency
  - an interaction between self-employed status and the city of Livorno

```
tuscany$dist2 <- tuscany$dist^2
tuscany$livselfemp <- (tuscany$city == "livorno") * (tuscany$occup == "selfemp")

ftusc <- perm + group ~ city + sex + marital + age +
  loc + children + educ + pvoto + training +
  empstat + occup + sector + wage + hour + feduc + femp + fbluecol +
  dist + dist2 + livselfemp
ps <- pscore(ftusc, tuscany)
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- Three supplementary arguments of `pscore` can be used
  - the maximum number of iterations (default 4)
  - the tolerance level for the t-tests (default 0.005)
  - the link for the binomial model (default to logit).
- `pscore` returns a `pscore` object which contains three tibbles:
  1. `strata` contains information about the stratas (the frequencies, the average propensity scores and the probability value of the hypothesis of no difference of the propensity scores in the two groups)
  2. `cov_balance` has a line for every covariate and contains the strata for which the probability value is the smallest
  3. `model` contains the original data sets with some supplementary columns:
  - `pscore` contains the propensity score for every observation
  - `.gp` is a factor with levels `"control"` and `"treated"`
  - `.cs` is a boolean indicating whether the propensity score for an observation lies in the interval of scores for the treated
  - `.resp` contains the response
  - `.cls` indicates the strata for the observation

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- A `summary` method is provided and the `print` method for `summary.pscore` objects has a `step` argument that allows to print the result of each step of the estimator.
- Two of the initial stratas were cut in halves (0-0.2 and 0.6-0.8). The control subsample is restricted to the range of the values of the propensity score for the treated; therefore, only 592 observations of the control group out of 628 are used.

```
1 ps$strata[, c(1,5,6,2,3)]
```

```
        cls n_control n_treated ps_control ps_treated
1   [0,0.1)       217        11 0.05269215 0.06465599
2 [0.1,0.2)       138        24 0.15149207 0.15285985
3 [0.2,0.4)       118        60 0.29378350 0.29129358
4 [0.4,0.6)        81        60 0.50136633 0.49695526
5 [0.6,0.7)        21        35 0.64611288 0.66792193
6 [0.7,0.8)        14        56 0.74478424 0.75384585
7   [0.8,1)         3        35 0.85032925 0.83665869
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- With `step = "covariates"`, we get a (long) table indicating, for each covariate, the results of the balance test between the treatment and the control group

```
1 ps$cov_balance
```

	name	classe	pvalue
1	citygrosseto	[0.4,0.6)	0.3913460128
2	citylivorno	[0.8,1)	0.0005301713
3	citypisa	[0.1,0.2)	0.0309213550
4	citylucca	[0.2,0.4)	0.1312218542
5	sexmale	[0.6,0.7)	0.2827215803
6	age	[0,0.1)	0.0951303241
7	loccentro	[0,0.1)	0.3608490981
8	locsud	[0,0.1)	0.4251824782
9	locestero	[0.2,0.4)	0.3132300977
10	children	[0,0.1)	0.0467478226
11	educ	[0,0.1)	0.3126301521
12	pvoto	[0.7,0.8)	0.2780706319
13	training	[0.8,1)	0.1100596133
14	empstatunemp	[0.6,0.7)	0.0446371340
15	empstatolf	[0.6,0.7)	0.0648372489
16	occupselfemp	[0,0.1)	0.1102833396

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The values of the estimated ATET (average treatment effect of the treated)
  - The estimated treatment effect is 0.177, which is slightly higher than the treatment effect computed with the whole sample which was 14.7%
  - It is highly significant, being the standard deviation 0.035

```
1 ps
```

```
Number of strata 7
pscore range  : 0.001605198 - 0.9068141
common support: 0.01314591 - 0.9068141
untreated obs used: 592 out of 628
ATET: 0.1769476 (0.03543038)
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- An alternative to using strata and computing the ATET as a weighted average of the difference of the mean of the outcome between treated and control observations in each stratum is
  - **to match each treated observation to one or several observations in the control group**
- The simplest algorithm consists of selecting, for every treated observation, the control observation which has the closest value of propensity score
- Using the `model` tibble returned by `pscore`, we begin with constructing two tibbles, one for the treatment group and one for the control group, containing only the index of the observation and the value of the propensity score

```
1 tusc_tr <- ps$model[ps$model$group == "treated", c("id", "pscore")]
2 colnames(tusc_tr) <- c("id_tr", "ps_tr")
3
4 tusc_ctl <- ps$model[ps$model$group == "control", c("id", "pscore")]
5 colnames(tusc_ctl) <- c("id_ctl", "ps_ctl")
6
7 print(head(tusc_tr, 2))
```

```
    id_tr     ps_tr
313   214 0.1484308
330   310 0.1343662
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- We then need to join the two tables
  - inequality: one can for example match a treated observation with all the observations in the control group with higher propensity scores
  - rolling: select only one observation, the closest one
- **Inequality**: We do it below only for the first two treated observations and they are matched to respectively 309 and 331 observations in the control group

```
 1  tusc_tr_2 <- tusc_tr[1:2, ]
 2
 3  result <- data.frame(
 4    id_tr = tusc_tr_2$id_tr,
 5    n = sapply(1:2, function(i) {
 6      sum(tusc_ctl$ps_ctl >= tusc_tr_2$ps_tr[i])
 7    })
 8  )
 9
10  print(result)
```

```
  id_tr   n
1   214 309
2   310 331
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- We then need to join the two tables
  - inequality: one can for example match a treated observation with all the observations in the control group with higher propensity scores
  - rolling: select only one observation, the closest one
- **Rolling**: this time one row is returned for every treated observation, and the same control observation can be matched with several treated observations

```
id_sup <- ps_sup <- numeric(nrow(tusc_tr))

for (i in seq_len(nrow(tusc_tr))) {
  ps_treated <- tusc_tr$ps_tr[i]
  candidates <- tusc_ctl[tusc_ctl$ps_ctl >= ps_treated, ]

  if (nrow(candidates) == 0) {
    id_sup[i] <- NA
    ps_sup[i] <- NA
  } else {
    j <- which.min(candidates$ps_ctl)
    id_sup[i] <- candidates$id_ctl[j]
```

```
    ps_sup[i] <- candidates$ps_ctl[j]
  }
}

match_sup <- data.frame(
  id_tr  = tusc_tr$id_tr,
  ps_tr  = tusc_tr$ps_tr,
  id_sup = id_sup
```

```
  id_tr      ps_tr id_sup     ps_sup
1   214 0.14843075   4098 0.14946177
2   310 0.13436619   4594 0.13549186
3   332 0.07885062   4765 0.07889962
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- Next, we match with the closest lower propensity score control:

```r
id_inf <- ps_inf <- numeric(nrow(tusc_tr))

for (i in seq_len(nrow(tusc_tr))) {
  ps_treated <- tusc_tr$ps_tr[i]
  candidates <- tusc_ctl[tusc_ctl$ps_ctl <= ps_treated, ]

  if (nrow(candidates) == 0) {
    id_inf[i] <- NA
    ps_inf[i] <- NA
  } else {
    j <- which.max(candidates$ps_ctl)
    id_inf[i] <- candidates$id_ctl[j]
    ps_inf[i] <- candidates$ps_ctl[j]
  }
}

match_inf <- data.frame(
  id_tr  = tusc_tr$id_tr,
  ps_tr  = tusc_tr$ps_tr,
```

```
20    id_inf = id_inf,
```

```
  id_tr      ps_tr id_inf     ps_inf
1   214 0.14843075   4283 0.14719677
2   310 0.13436619   4857 0.13243997
3   332 0.07885062   3369 0.07806057
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- We then join the two tables (by the index for treated observations id_tr) and select the control observation which is the closest:

```r
m <- merge(match_sup, match_inf, by = "id_tr", all.x = TRUE)
m$id_ctl <- ifelse(
  is.na(m$ps_inf) | (!is.na(m$ps_sup) & (m$ps_tr.y - m$ps_inf) >= (m$ps_sup - m$ps_t
  m$id_sup,
  m$id_inf
)

m$ps_ctl <- ifelse(
  m$id_ctl == m$id_inf,
  m$ps_inf,
  m$ps_sup
)

match_nearest <- m[, c("id_tr", "id_ctl", "ps_tr.x",  "ps_tr.y", "ps_ctl")]
head(match_nearest, 3)
```

```
  id_tr id_ctl   ps_tr.x   ps_tr.y    ps_ctl
1     9   4211 0.8199733 0.8199733 0.8323604
2    11   4038 0.8600481 0.8600481 0.8809654
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- For a couple of treated observations, the propensity score is greater than the highest propensity score in the control group
- Therefore, `ps_sup` is `NA`, and we set it to an arbitrary high value so that the `id_inf` observation is selected
- We can then compute the number of observations in the control group that are used and select control observations which are the most often used to match treated observations (for example, observation 4935 in the control group matches 16 observations in the treatment group)

```
1 length(unique(match_nearest$id_ctl))
```

```
[1] 146
```

```
1 tab <- table(match_nearest$id_ctl)
2 tab_sorted <- sort(tab, decreasing = TRUE)
3 head(tab_sorted, 2)
```

```
4935 4211
  16   13
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- For some observations, the algorithm may result in a poor match for some treated observations if the difference between the probability scores of this observation and the matched control observation is high.
- In our sample, the highest difference is about 2.6%.

```
match_nearest$diff <- match_nearest$ps_tr.x - match_nearest$ps_ctl
match_nearest[order(-match_nearest$diff), ][1:3, ]
```

```
    id_tr id_ctl   ps_tr.x   ps_tr.y    ps_ctl       diff
34     89   4038 0.9068141 0.9068141 0.8809654 0.02584868
166   412   4935 0.8088115 0.8088115 0.7862707 0.02254076
154   375   4935 0.8087364 0.8087364 0.7862707 0.02246570
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The sample can be reduced to observations for which the difference is lower than a given value: **caliper matching**.
- For example, to restrict the sample to treated observations for which the propensity score difference with its matched control observation is lower than 1%
- We then lose 25 treated observations

```
1 match_caliper <- match_nearest[abs(match_nearest$ps_tr - match_nearest$ps_ctl) < 0.0
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- To compute the treatment effect, we pivot the tibble in “long” format (one line for each observation) and we join it to `tuscany` to get the response (`perm`) for each observation

```r
tmp <- match_nearest[, !(names(match_nearest) %in% c("ps_tr", "ps_ctl"))]
id_tr <- tmp$id_tr
id_ctl <- tmp$id_ctl
group_tr <- rep("tr", length(id_tr))
group_ctl <- rep("ctl", length(id_ctl))

match_long <- data.frame(
  id = c(id_tr, id_ctl),
  gp = c(group_tr, group_ctl)
)

match_smpl <- merge(match_long, tuscany[, c("id", "perm")], by = "id", all.x = TRUE)
head(match_smpl, 2)
```

```
  id gp perm
1  9 tr    0
2 11 tr    0
```

Matching in Observational Studies

Example: Temporary Work Agencies (Italy)

- The ATET (0.1886) is close to the one obtained previously using the algorithm based on stratas (0.1769)

```r
means <- tapply(match_smpl$perm, match_smpl$gp, mean)
# ATET
atet <- means["tr"] - means["ctl"]
c(treatment = means["tr"], control = means["ctl"], atet = atet)
```

```
treatment.tr  control.ctl      atet.tr
   0.3131673    0.1245552    0.1886121
```

Matching in Observational Studies

MatchIt: Introduction

- MatchIt implements advanced matching techniques
- Main function: matchit()
- Interface: formula + data
- Key argument: replace = TRUE → allows a control unit to match multiple treated units
- Default method: “nearest”

```
1 library(MatchIt)
2 ftusc2 <- update(ftusc, group ~ .)
3 mtch <- matchit(ftusc2, tuscany, replace = TRUE)
```

Matching in Observational Studies

MatchIt: Introduction

- The number of matched observations is 427 (as previously), the 281 treated observations and 146 observations of the control group

```
1 summary(mtch)$nn
```

```
               Control Treated
All (ESS)     628.0000     281
All           628.0000     281
Matched (ESS)  68.7215     281
Matched       146.0000     281
Unmatched     482.0000       0
Discarded       0.0000       0
```

Matching in Observational Studies

MatchIt: Introduction

- A `match.data` function is provided in order to extract the data frame restricted to the treated observations and the subset of observations of the control group that match
- weights → for subsequent analyses
- Treated: weight = 1
- Controls: proportional to the number of matches (For an observation of the control group that matches only one treated observation, the weight is $146/281 = 0.52$ and, for example, for an observation of the control group that matches four treated observations, the weight is :$146/281 \times 4 = 2.08$.)

```
1 matched_data <- match.data(mtch)
2 head(matched_data, 2)
```

```
      id age    sex marital children feduc fbluecol femp educ pvoto training
308 4862  32 female married        0     5         0    1   18    98        1
311 3756  33 female  single        0     5         0    0   13    80        1
        dist        nyu hour     wage    hwage contact  region     city   group
308 43.03258 0.06666667   36 981.2681 6.814362       1 Tuscany grosseto control
311 62.15774 0.02142857    0   0.0000       NA       0 Tuscany grosseto control
    sector    occup empstat contract    loc outcome perm    dist2 livselfemp
308   serv whitecol    empl     atyp centro   other    0 1851.803          0
311  nojob    nojob   unemp    nojob centro    perm    1 3863.585          0
```

```
      distance  weights
308 0.02923802 0.519573
311 0.18637402 0.519573
```

Matching in Observational Studies

MatchIt: Introduction

- Caliper matching is performed using the `caliper` argument.

```
mtch_cap <- matchit(ftusc2, tuscany, replace = TRUE,
                    caliper = 0.01)
matched_data <- match.data(mtch_cap)

att <- with(matched_data, mean(perm[group == "treated"]) - mean(perm[group == "contr
att
```

```
[1] 0.1892879
```

Matching in Observational Studies

MatchIt: Introduction

- Once the matching has been performed, the quality of the balancing process can be assessed
- The `plot` method draws a Love plot; for every covariate, two standardized mean differences are plotted, one for the raw data set and one for the balanced one

```
1  plot(summary(mtch_cap))
```