Lab session practice

Slawski and Ben-David

8/21/2019

Ground Truth: all variables with correct correspondence

- ▶ Import the original data and perform a LS regression
- ► This regression is the oracle model

R summary

```
coef(lm0)
```

```
##
               (Intercept)
                                               SEX
##
             1.0094294300
                                    -0.2135490993
          I(EXPERIENCE^2)
                                         EDUCATION as.factor
##
##
            -0.0004441696
                                     0.0730130844
   as.factor(OCCUPATION)3 as.factor(OCCUPATION)4 as.factor
##
            -0.2262765855
                                    -0.3658979172
## as.factor(OCCUPATION)6
                                             UNION
            -0.1964030978
##
                                     0.1998655883
```

Importing two data files to be linked

First import the data files.

Check both data files

What variables can be used as matching variables?

▶ What variables are approriate for linkage?

```
[1]
       "ID"
                                                    "AGE"
                  "FNAME"
                              "I.NAME."
                                         "SEX"
##
   [8] "WAGE"
##
    [1]
        "ID"
                       "OCCUPATION" "SECTOR"
                                                    "UNION"
    [6]
##
        "EXPERIENCE"
                       "AGE"
                                     "SEX"
                                                    "MARR"
   [11] "SOUTH"
                       "FNAME"
                                     "LNAME"
                                                    "ADDRESS"
```

Linking the data files

- ▶ Upload the R package **fastLink** for linkg files
- ▶ In this experiment we link two files based on **ZIPCODE**.
- ► Check fastlink function first.

```
help(fastLink)
```

##

```
## starting httpd help server ... done
```

To specify matching variables we choose varnames = "ZIPCODE".

The linked data

► We can get the linked data file form *getMatch" function.

Merging these two files

First ensure unique column names in the merged file

Merge the data linked by fastLink

```
merged_wages <-cbind.data.frame(dA, dB)</pre>
```

► Compute the fraction of mismatches (about 13%)

```
mean(merged_wages[,"A.ID"] != merged_wages[,"B.ID"])
```

```
## [1] 0.1292134831460674
```

Check the linear regression with linked file

Ceofficients of Naive LS regression

```
coef(lm_merged)
```

```
## (Intercept) B.SEX

## 1.1968475057691034813 -0.1887882270381399941 0.028082:

## I(EXPERIENCE^2) EDUCATION as.factor

## -0.0004280056942423294 0.0591585615568550119 -0.2575914

## as.factor(OCCUPATION)3 as.factor(OCCUPATION)4 as.factor

## -0.2255277397914074722 -0.3514718608619472051 -0.0752009

## as.factor(OCCUPATION)6 UNION

## -0.1809642839349065724 0.1802265346727588424
```

compare the result with the original regression with original data

Robust regression

► Try robust regression with Huber loss

Comparing estimation error of naive and robust estimation

```
sqrt(sum((coef(lm0) - coef(lm_merged))^2))
## [1] 0.2027454431130401
sqrt(sum((coef(lm0) - coef(rlm_merged))^2))
```

[1] 0.1701889571280111

Mixture model

Now we try the mixure modeling approach with composite likelihood:

```
source("../code/mixture_model.R")
X <- model.matrix(lm merged)</pre>
X <- X[,!(colnames(X) %in% "(Intercept)")]</pre>
y <- model.extract(lm merged$model, "response")
Xc \leftarrow apply(X, 2, function(z) z - mean(z))
yc \leftarrow y - mean(y)
tausq <- mean(yc^2)
f0 <- function(z) dnorm(z, mean = 0, sd = sqrt(tausq))
res <- fit_mixture(Xc, yc, f0,
                     control = list(init = "robust"))
interc <- mean(y - X %*% res$betahat)</pre>
```

Comparing the results

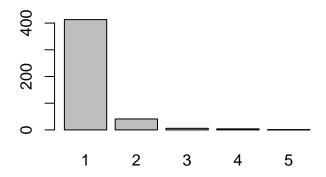
▶ Compute $\|\widehat{\beta} - \widehat{\beta}_{LS}\|$, where $\widehat{\beta}_{LS}$ denotes the estimates from the original LS regression model with original data and $\widehat{\beta}$ is the estimate from, Näive, Huber, or mixture models.

```
coef_mixture <- c(interc, res$betahat)</pre>
```

Comparing the results (continued)

```
sqrt(sum((coef(lm0) - coef mixture)^2))
## [1] 0.0272892695833842
sqrt(sum((coef(lm0) - coef(rlm_merged))^2))
## [1] 0.1701889571280111
sqrt(sum((coef(lm0) - coef(lm_merged))^2))
## [1] 0.2027454431130401
```

Historam of duplicates



Compute the estimates

```
beta_Q <- coef(lahiri_larsen_block(Xc, yc, blockix))
Q <- generate_Q_block(blockix)
beta_Q_check <- lahiri_larsen(Xc, yc, Q)

interc <- mean(y - X %*% beta_Q)
coef_Q <- c(interc, beta_Q)</pre>
```

► Compute $\|\widehat{\beta} - \widehat{\beta}_{LS}\|$:

```
sqrt(sum((coef(lm0) - coef_Q)^2))
```

```
## [1] 0.07430098233549876
```

Re-matching based on sorting

Plot of mismatches

