Evaluating the effect of smoking on blood cadmium and lead levels using the developed package `sensitivityq`

Anonymous

This file illustrates the R package "sensitivityq" for sensitivity analysis based on quantiles of hidden biases.

Install the `sensitivityq` package

We can install the package from Github. The github page of the package (https://github.com/Anonymous/sensitivityq) contains main functions with detailed explanation in R documentation. # Install the package via devtools, and then load:

```
library(devtools)
install_github("Anonymous/sensitivityq")
```

We can also install the package directly by running `devtools::install("sensitivityq")` in the console once we download the package. Then we can load the package by:

```
library('sensitivityq')
```

Sensitivity analysis for the effect of smoking on the blood cadmium level

Below we provide the code to replicate data analysis in the paper "Sensitivity Analysis for Quantiles of Hidden Biases in Matched Observational Studies".

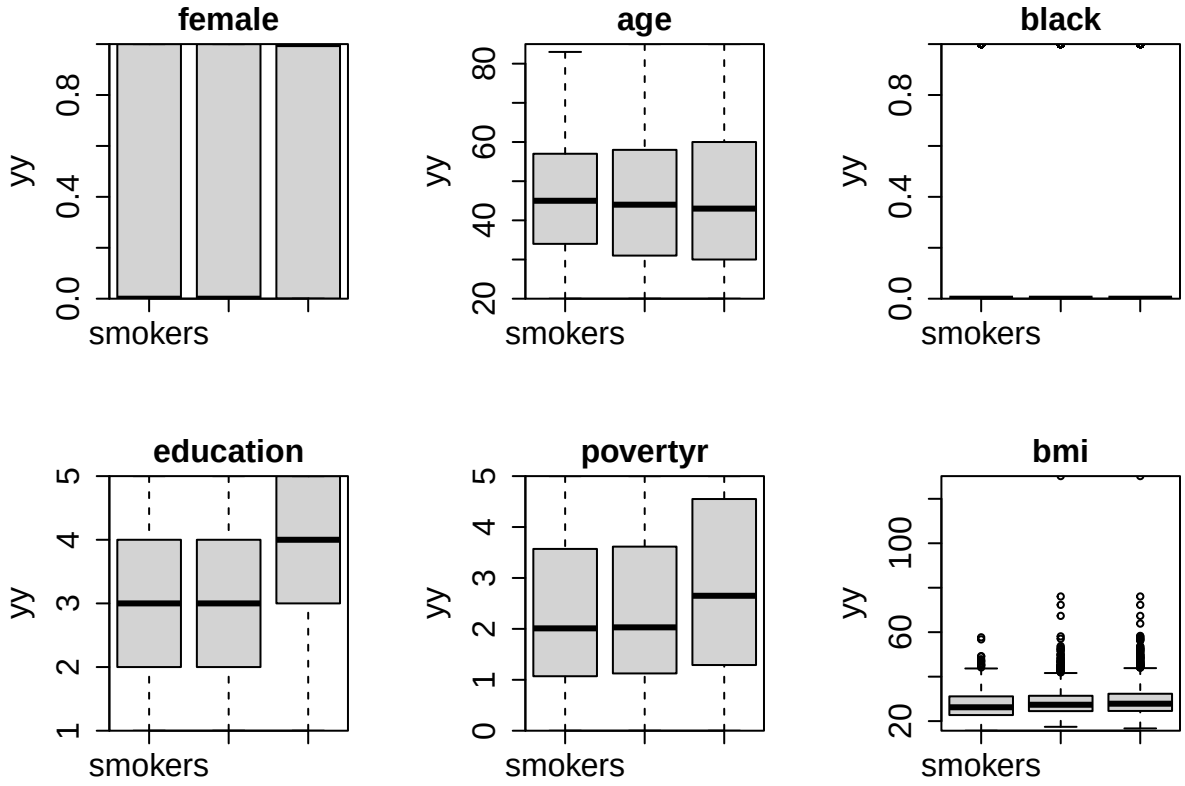
Matched data set

We use the data from the 2005–2006 National Health and Nutrition Examination Survey, which are also available in Yu (2020). The data include 2475 observations, including 521 daily smokers and 1963 nonsmokers. We use the recent optimal matching algorithm proposed by in Yu and Rosenbaum (2019), which takes into account six covariates including gender, age, race, education level, household income level and body mass index; for a more detailed description of these covariates, see, e.g., Yu (2020). The matched data consists of 512 matched sets, each of which contains one daily smoker and two nonsmokers. The original data `nh0506.rds` and the matched data `md.rds` are both available on the github page of the package (https://github.com/Anonymous/sensitivityq).

We first show the box plots of the six covariates for smokers, matched nonsmokers and original nonsmokers. From them, we can see that the balance of these covariates has been substantially improved by matching.

```
nh0506 = readRDS("nh0506.rds")
md = readRDS("md.rds")
par(
  mfrow    = c(2, 3),
  mar      = c(5, 5, 2, 2),
  xaxs     = "i",
```

```
  yaxs     = "i",
  cex.axis = 1.5,
  cex.lab  = 1.5
)
cov.vec = c("female","age","black","education","povertyr","bmi")
for(jj in 1:6){
  covariate = cov.vec[jj]
  yy = c(md[md$z==1, covariate], md[md$z==0, covariate], nh0506[nh0506$z==0, covariate])
  xx = c(rep("1", length(md[md$z==1, covariate])),
         rep("2", length( md[md$z==0, covariate])),
         rep("3", length(nh0506[nh0506$z==0, covariate])))
  boxplot(yy ~ xx, names = c("smokers","matched","nonsmoker"), xlab = "",
          main = covariate, cex.main = 1.5)
}
```



Sensitivity analysis for all quantiles of hidden biases

We then apply our method to conduct sensitivity analysis for all quantiles of hidden biases. In particular, we will use the function `Gamma_seq()` in the `sensitivityq` package. We use the difference-in-means statistic as the test statistic, which can be achieved by setting `inner=0,trim=Inf`. We are interested in sensitivity for all quantiles of hidden hidden biases, so we set `K=512`. Here `K` specifies the number of quantiles of hidden biases that we are interested in. That is, we are going to calculate the lower confidence limits for hidden biases at rank $I - K + 1 \leq k \leq I$, assuming that smoking has no effect on cadmium. As we can see, it takes about 4.5 minutes to get the lower confidence limits for all quantiles of hidden biases.

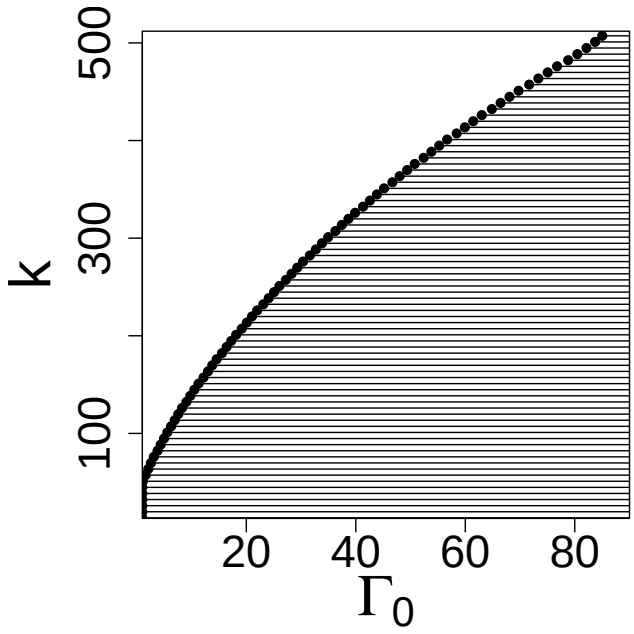
```
I = max(md$mset)
t = Sys.time()
Gamma.cadmium = Gamma_seq(md$cadmium, md$z, mset = md$mset, inner = 0, trim = Inf,
                          thres = 0.05, Ks = 512:1)
```

```r
Sys.time() - t
```

```
## Time difference of 4.419689 mins
```

We can conveniently visualize the lower confidence limits for all hidden biases.

```r
par(
  mar      = c(5, 5, 2, 2),
  xaxs     = "i",
  yaxs     = "i",
  cex.axis = 2,
  cex.lab  = 2
)
plot(1, type = "n", lty = 2,  xlim = c(1, 90), ylim = c(512 - 500 + 1, 512),
     cex.lab = 2.6, mgp = c(3.2, 1, 0), ylab = "k", xlab = expression(Gamma[0]))
for(k in seq(1, 512, by = 500 / 80)){
  lines(c(Gamma.cadmium[512 - k + 1], 90), rep(k, 2), lty = 1, lwd = 1)
  points(Gamma.cadmium[512 - k + 1], k, pch = 16, cex = 1, lwd = 1)
}
```



The following code shows the lower confidence limits for the largest, 52th largest, 154th largest, 256th largest and 359th largest hidden biases. These are equivalently the 100%, 90%, 70%, 50% and 30% quantiles of the hidden biases.

```r
beta = c(1, 0.9, 0.7, 0.5, 0.3)
rev(Gamma.cadmium)[ceiling(I * beta)]
```

```
## [1] 82.44162 72.52080 46.90880 26.88896 11.66503
```

Sensitivity analysis for a specific quantile of hidden biases with a given upper bound

If we are interested in sensitivity analysis for a specific quantile of hidden biases with a given upper bound, we can use the function `senmk()` in the `sensitivityq` package, which provides the p-value for the sensitivity analysis with given constraint. The arguments `k` and `gamma` of the function `senmk()` specifies the sensitivity

analysis constraint that the hidden bias at rank `k` is less than or equal to `gamma`. Below is an example with `k = 512` and `gamma = 82.44`.

```
senmk(md$cadmium, md$z, md$mset, k = I, gamma = 82.44, inner = 0, trim = Inf)$pval
```

```
## [1] 0.04999661
```

Analogously, the p-values for testing the null of no effect when the 90%, 70%, 50% and 30% quantiles of the hidden biases are bounded, respectively, by their upper bounds can be calculated as follows.

```
senmk(md$cadmium, md$z, md$mset, k = ceiling(I*0.9), gamma = 72.52,
      inner = 0, trim = Inf)$pval
```

```
## [1] 0.04999786
```

```
senmk(md$cadmium, md$z, md$mset, k = ceiling(I*0.7), gamma = 46.90,
      inner = 0, trim = Inf)$pval
```

```
## [1] 0.04995089
```

```
senmk(md$cadmium, md$z, md$mset, k = ceiling(I*0.5), gamma = 26.88,
      inner = 0, trim = Inf)$pval
```

```
## [1] 0.04989957
```

```
senmk(md$cadmium, md$z, md$mset, k = ceiling(I*0.3), gamma = 11.66,
      inner = 0, trim = Inf)$pval
```

```
## [1] 0.04985813
```

We can then calculate the 95% lower confidence limits for the average hidden biases $\bar{\Gamma}_g^\star$ with $g(x)$ equal to $x, \log(x), x/(1+x)$, assuming smoking has no effect on **cadmium**:

```
mean(Gamma.cadmium)
```

```
## [1] 32.18186
```

```
exp(mean(log(Gamma.cadmium)))
```

```
## [1] 17.7367
```

```
mean(Gamma.cadmium / (1 + Gamma.cadmium)) / (1 - mean(Gamma.cadmium / (1 + Gamma.cadmium)))
```

```
## [1] 8.334141
```

Sensitivity analysis for the effect of smoking on the blood lead level

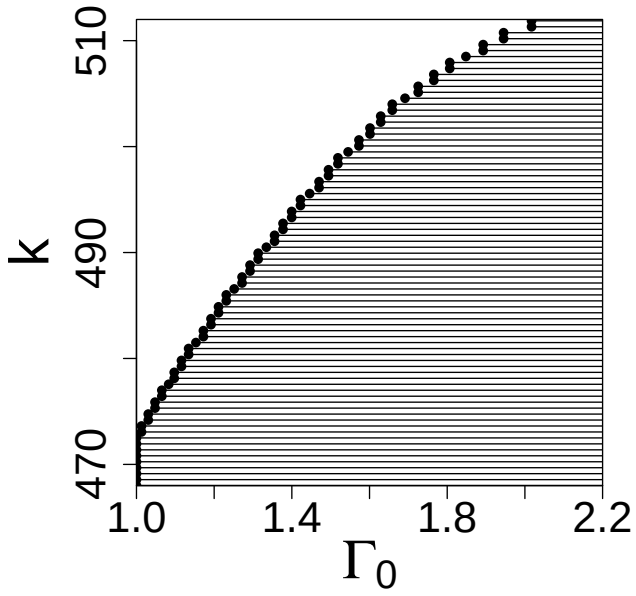
We can perform similar sensitivity analysis for the effect of smoking on the blood lead level. Below we calculate and visualize the lower confidence limits for all quantiles of hidden biases assuming smoking has no effect on lead. It takes about half a minute to get the lower confidence limits for all quantiles of hidden biases.

```
t = Sys.time()
Gamma.lead = Gamma_seq(md$lead, md$z, mset = md$mset, inner = 0, trim = Inf, thres = 0.05, Ks = 512:1)
Sys.time()-t
```

```
## Time difference of 26.15404 secs
```

```
par(
  mar      = c(5, 5, 2, 2),
  xaxs     = "i",
  yaxs     = "i",
  cex.axis = 2,
```

```
  cex.lab  = 2
)
plot( 1, type = "n", lty = 2,  xlim = c(1, 2.2), ylim = c(512 - 45 + 1, 512),
      cex.lab = 2.6, mgp=c(3.2, 1, 0), ylab="k", xlab = expression(Gamma[0]))
for(k in seq(512 - 45 + 1, 512, by = 45 / 80)){
  lines(c(Gamma.lead[512 - k + 1], 2.2), rep(k, 2), lty = 1, lwd = 1)
  points(Gamma.lead[512 - k + 1], k, pch = 16, cex = 1, lwd = 1)
}
```



The code below gives the lower confidence limits for the 100%, 90%, 70%, 50% and 30% quantiles of the hidden biases.

```
beta = c(1, 0.95, 0.9)
rev(Gamma.lead)[ceiling(I * beta)]
```

```
## [1] 2.017109 1.251559 1.000000
```

Reference

Yu, R. 2020. *Bigmatch: Making Optimal Matching Size-Scalable Using Optimal Calipers*. https://CRAN.R-project.org/package=bigmatch.

Yu, R., and P. R. Rosenbaum. 2019. "Directional Penalties for Optimal Matching in Observational Studies." *Biometrics* 75: 1380–90.