Work in Progress

Erik Bulow

9 februari 2016

Contents

1 Intro

Detta är ett arbetsdokument för att dokumentera mitt arbete då det pågår!

2 Förberedelser

```
# Try it out!
memory.limit(50000)
```

```
## [1] 50000
```

```
options(samplemetric.log = TRUE)
set.seed(123)
```

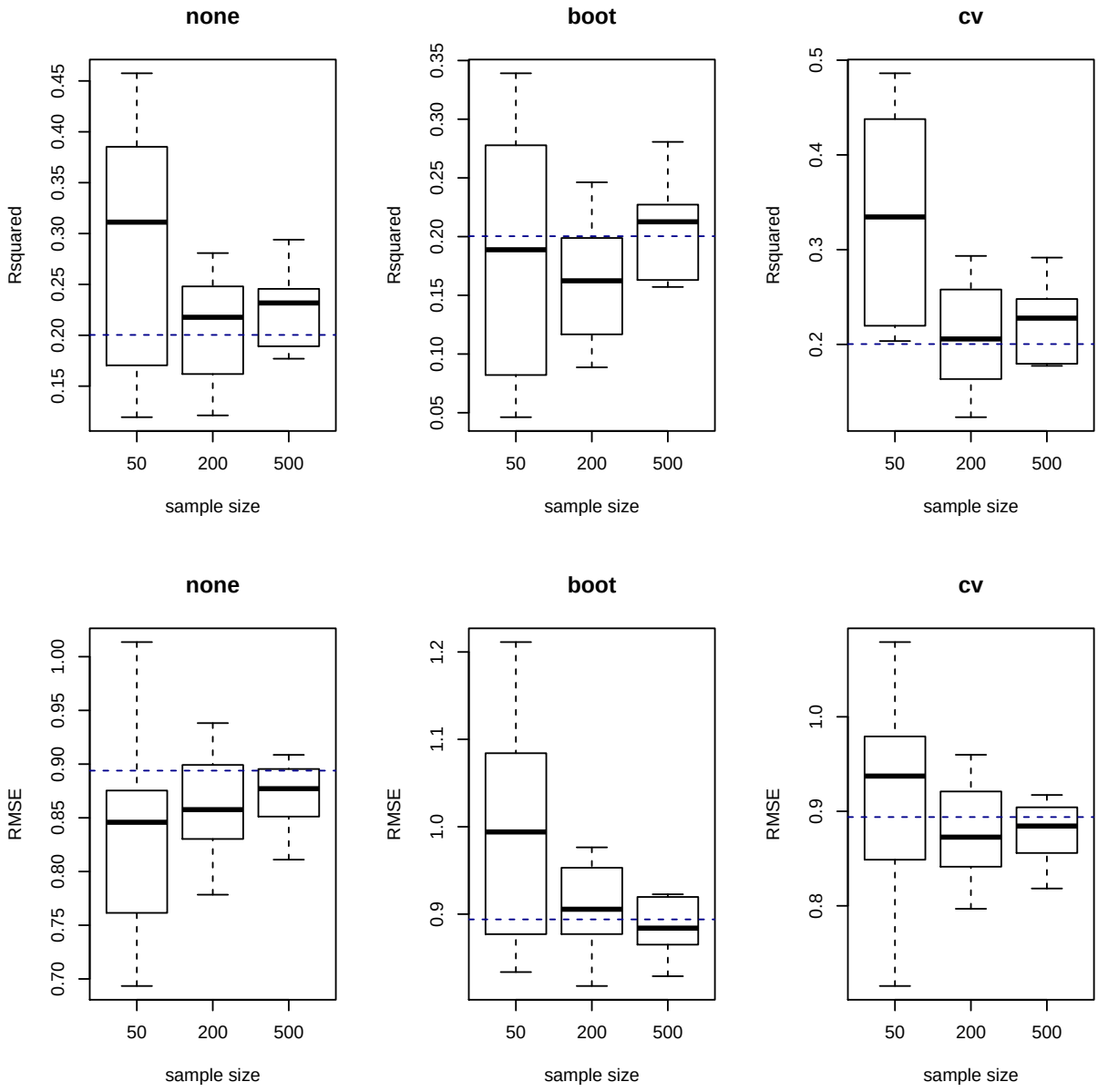
3 2016-02-09

```
# Samma för alla
d  <- sim_data()
ss <- subsamples(d, n.sample = c(50, 200, 500), N = 10)

# Beräkna för olika methods
```

```r
mthds <- c("none", "boot", "cv")
# mthds <- c("none", "boot", "boot632", "cv", "repeatedcv", "LOOCV", "LGOCV")
ms <- lapply(mthds, function(m) metrics(ss, m))
```

```
## Loading required package: lattice
## Loading required package: ggplot2
```

```r
names(ms) <- mthds

# Plotta för alla
par(mfcol = c(2, length(mthds)))
for (i in seq_along(ms)) plot(ms[[i]], main = mthds[i])
```



3.1 Slutsatser

1. Vi ska egentligen inte jämföra resultaten mot beräknade värden för hela datasettet utan använda beräkningar med "none" som facir (dvs på de mindre datasetten).
2. Vi identifierar mönstret att högre RMSE betyder mer brus => mindre R2
3. Framförallt noteras att cv överskattar resultatet och orsak till detta måste undersökar! Jag finner liknande resultat i (Steyerberg et al. 2001, 5).

3.2 Testar referenser

Enl: http://rmarkdown.rstudio.com/authoring_bibliographies_and_citations.html

Testar här: (Steyerberg et al. 2001) (vilket sätt man presenterar referenserna på kan också ställas in). Verkar inte funka med Endnote-filer (framgår också av länk ovan att detta är erkänt problem). Men funkar med många andra format, t ex .bib. Jag testar därför att istället använda Mendeley, vilket jag är riktigt nöjd med!

4 2016-20-10

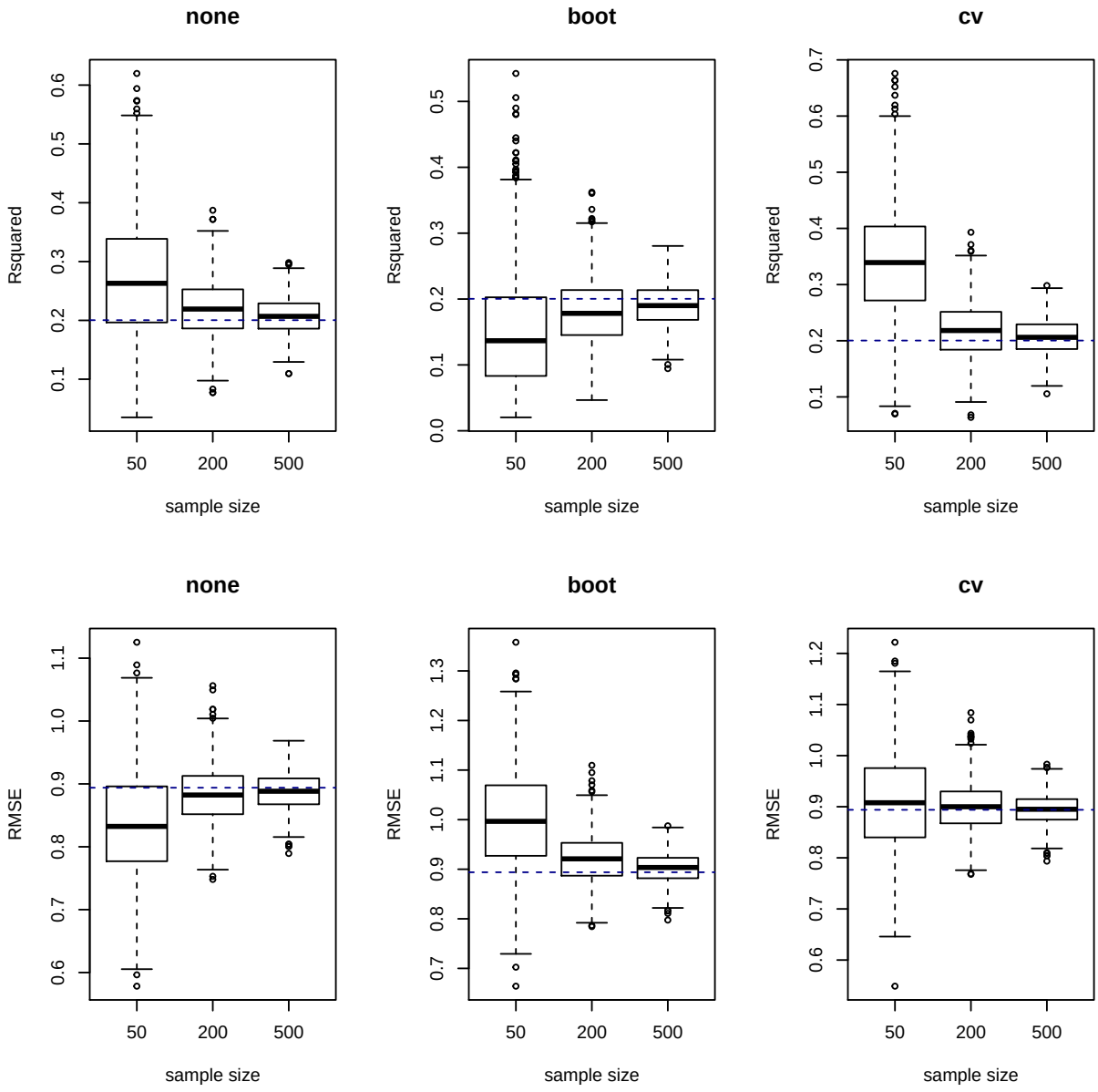
4.1 Diskussion med SN

Vi drar en del lärdomar av (Steyerberg et al. 2001):

1. Vi bör använda liknande men förenklade jämförelsemått för R2, dvs enligt tabel 1 (dock inte för Bootstrap 0.632 som har en onödigt krångig metod och även för de andra metoderna bör vi försöka använda samma mått för alla).
2. Vi bör utöka modelleringen till att även använda splines och med en mer komplicerad modell, dvs Y ~ g(.) för ngt g. (Jmfr artikeln osm anväder logistisk regression t ex).

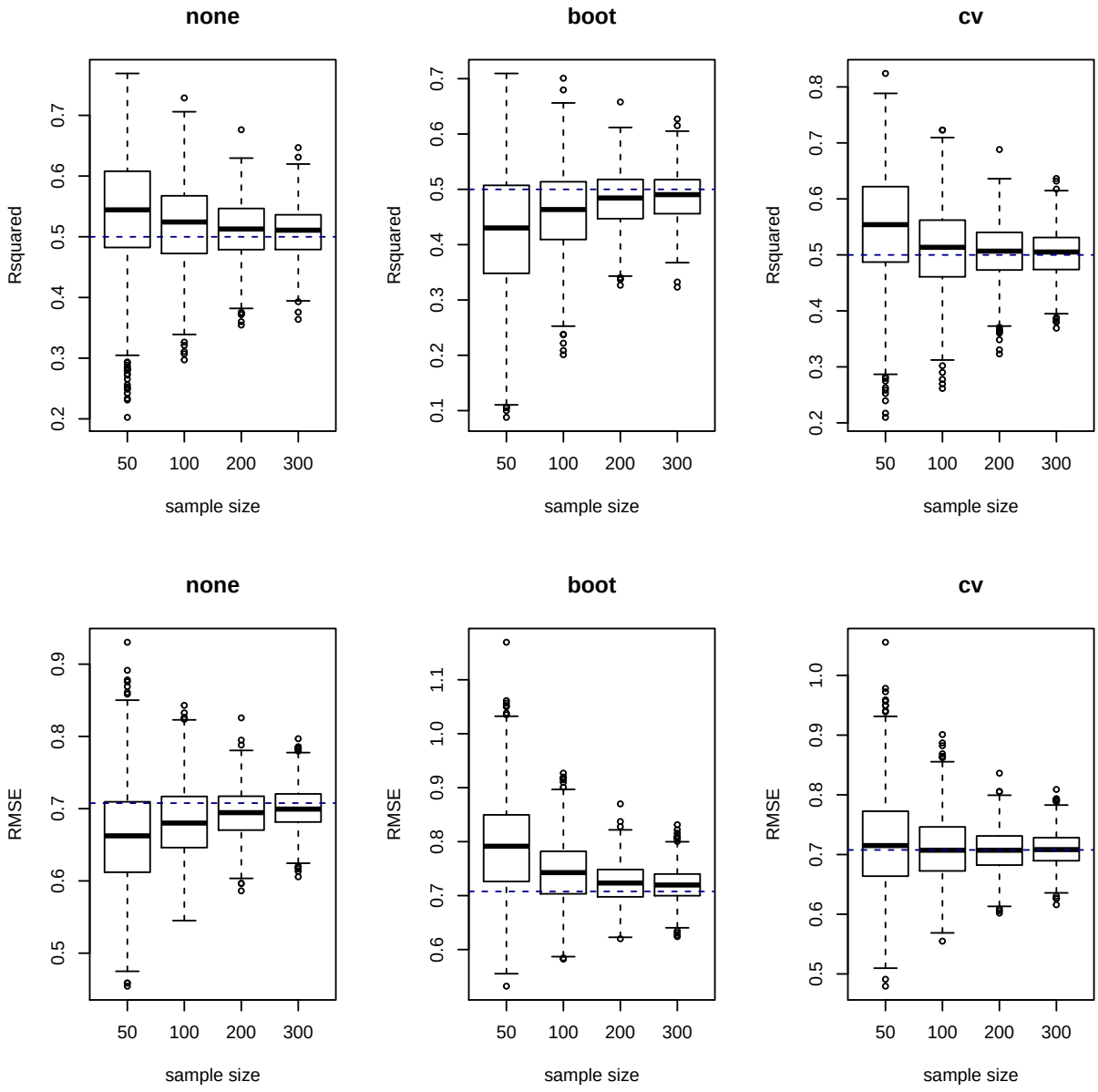
Vi ökar N till 1000:

```
# Samma för alla
ss <- subsamples(d, n.sample = c(50, 200, 500), N = 1000)

# Beräkna för olika methods
mthds <- c("none", "boot", "cv")
ms <- lapply(mthds, function(m) metrics(ss, m))
names(ms) <- mthds

# Plotta för alla
par(mfcol = c(2, length(mthds)))
for (i in seq_along(ms)) plot(ms[[i]], main = mthds[i])
```



Vi ändrar r2 och n.sample:

```
# Samma för alla
d  <- sim_data(r2 = .5)
ss <- subsamples(d, n.sample = c(50, 100, 200, 300), N = 1000)

# Beräkna för olika methods
mthds <- c("none", "boot", "cv")
ms <- lapply(mthds, function(m) metrics(ss, m))
names(ms) <- mthds

# Plotta för alla
par(mfcol = c(2, length(mthds)))
for (i in seq_along(ms)) plot(ms[[i]], main = mthds[i])
```



Referenser

Steyerberg, Ewout W, Frank E Harrell, Gerard J.J.M Borsboom, M.J.C Eijkemans, Yvonne Vergouwe, and J.Dik F Habbema. 2001. "Internal validation of predictive models." *Journal of Clinical Epidemiology* 54 (8): 774–81. doi:10.1016/S0895-4356(01)00341-9.