Arbetslogg 2016 vecka 6

Erik Bulow

9 februari 2016

Contents

1 Intro

Detta är ett arbetsdokument för att dokumentera mitt arbete då det pågår! Min plan är att skapa en sådan logg för varje påbörjad vecka!

2 Förberedelser

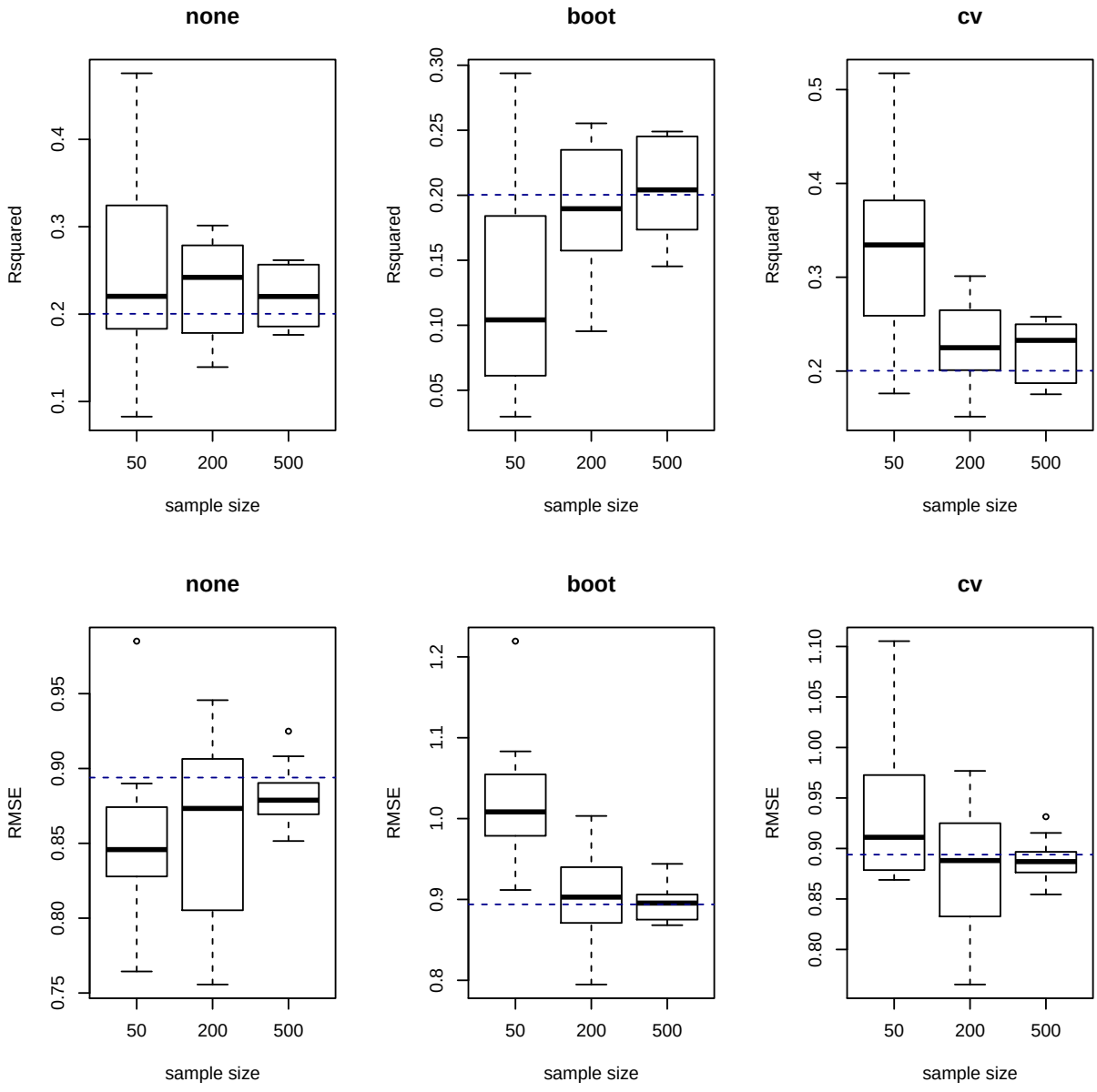
```
# Try it out!
memory.limit(50000)
```

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

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

3 2016-02-09

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

# Beräkna för olika methods
mthds <- c("none", "boot", "cv")
# mthds <- c("none", "boot", "boot632", "cv", "repeatedcv", "LOOCV", "LGOCV")
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])
```



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

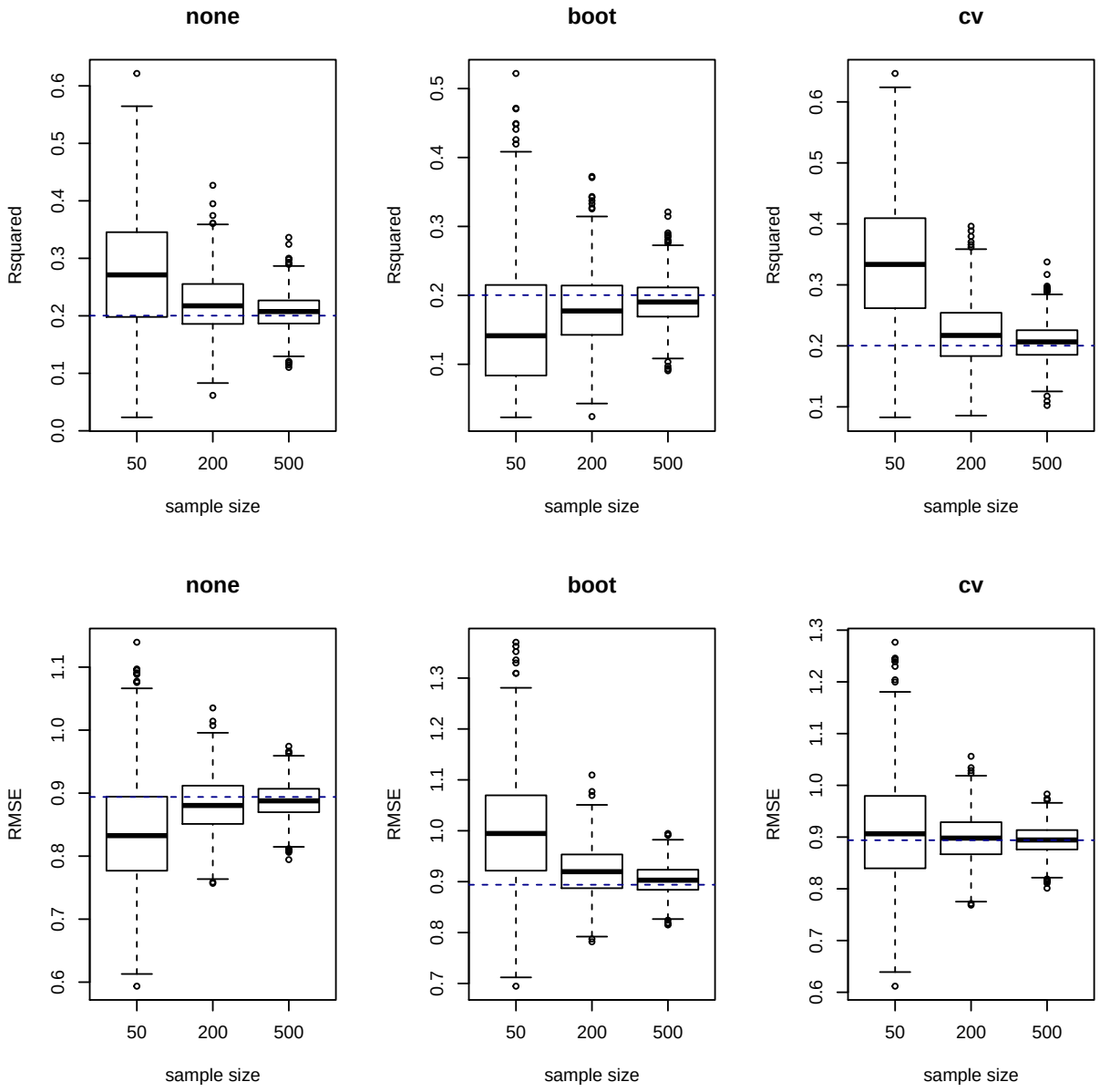
4.1 Diskussion med SN

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])
```



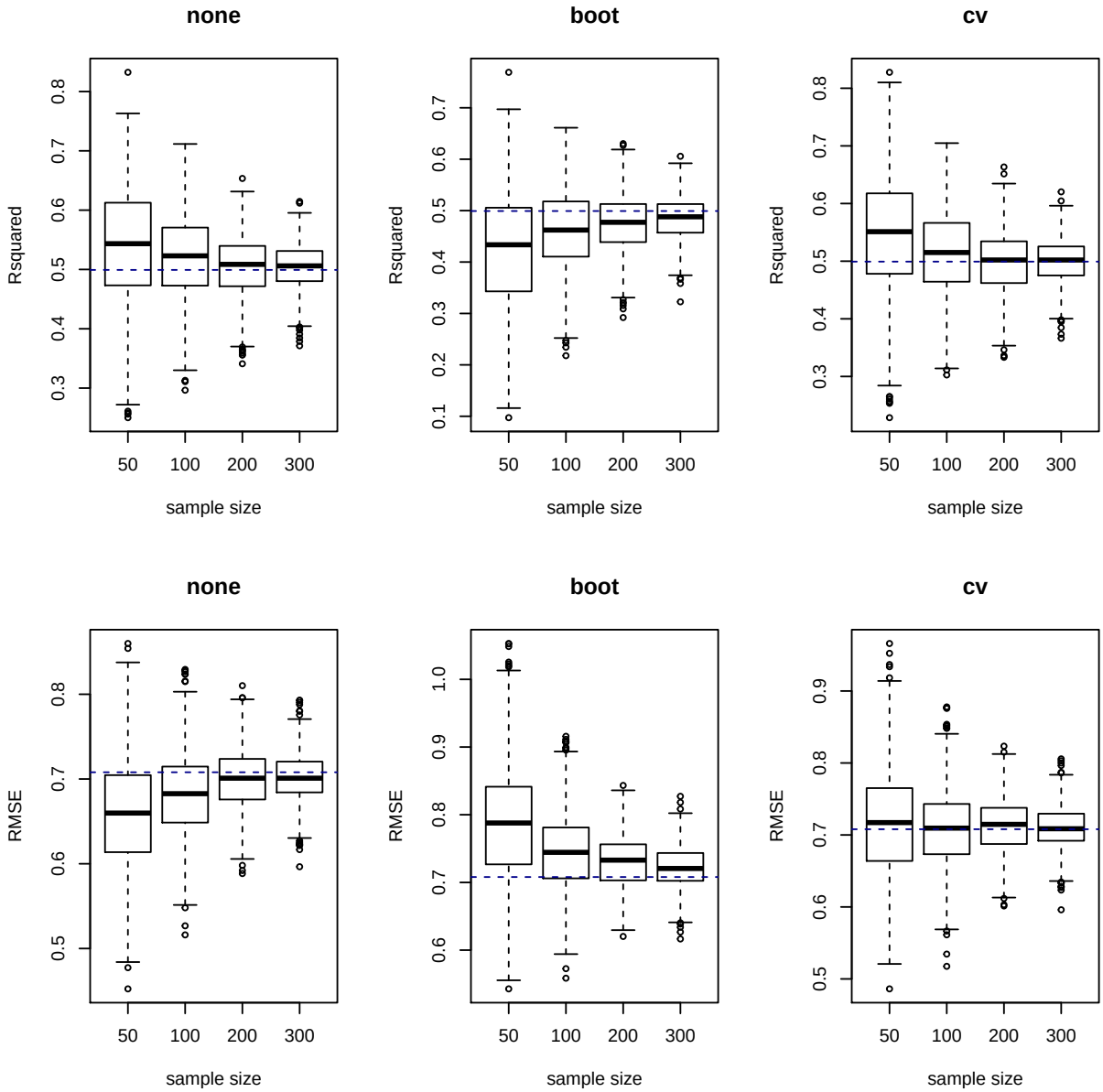
Vi kan nu med större säkerhet dra slutsatsen att boot tycks underskatta och cv överskatta resultatet. Att vi överskattar för none kan dock rimligen bero på vårt låga r2 = 0.2. Vi förväntar oss ett mer symmetriskt resultat för r2 = .5 och testar med det. Vi vill undersöka för lite fler n.samples.

När vi kör cv delar vi datamängden i 10 delar, dvs 5 observationer i varje och validering sker på dessa fem datapunkter, dvs samma antal som antalet kovariater.

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))
```

```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])
```



Bra förklaring till Bootstrap och .632 version: http://stats.stackexchange.com/questions/96739/what-is-the-632-rule-in-bootstrapping Harrell skriver också i ett svar att 0.632+-versionen inte nödvändig om man använder deviance-based värde etc.

CV är en mer traditionell metod än boot. Boot ger mindre varians för små stickprovsstorlekar (Kim 2009, p 1).

Har satt upp en gratisinstans av R + RStudio på AWS. Gratisversion i ett år. Dock ganska begr resurser om man inte betalar. Men funkar bra. Nås via: http://ec2-54-93-99-36.eu-central-1.compute.amazonaws.com/

4.1.1 Fundering

1. Hur går det att undersöka R2 om vår modell inte är linjär. Vårt R2 bygger ju på korrelation i det linjära fallet men måste bli ett annat mått med annan modell (**???**).

5 2016-02-11

Jag separerer R-paketet från loggen då R-paketet annars riskerar bli alldeles för stort. Jag versionshanterar f.ö. inte cache-filer från loggarna, vilka blir ganska stora. Jag tänker också att jag påbörjar en ny loggfil varje vecka för att inte riskera att kah måste uppdatera ett alldeles orimligt långt dokument etc.

Senaste bilden från igår visar f.ö. inte den symmetri som vi hoppades på för r2 simulerat som .5. Vi ser fortfarande samma mönster som tidigare för none. Boot ger fortfarande en underskattning. Dock tycks biasen från cv jmfrt med none ha minskat ganska avsevärt . Vi får fortfarande en högre skattning men inte mkt i förhållande till none (vilket vi bör se som facit).

5.1 Diskussion med SN

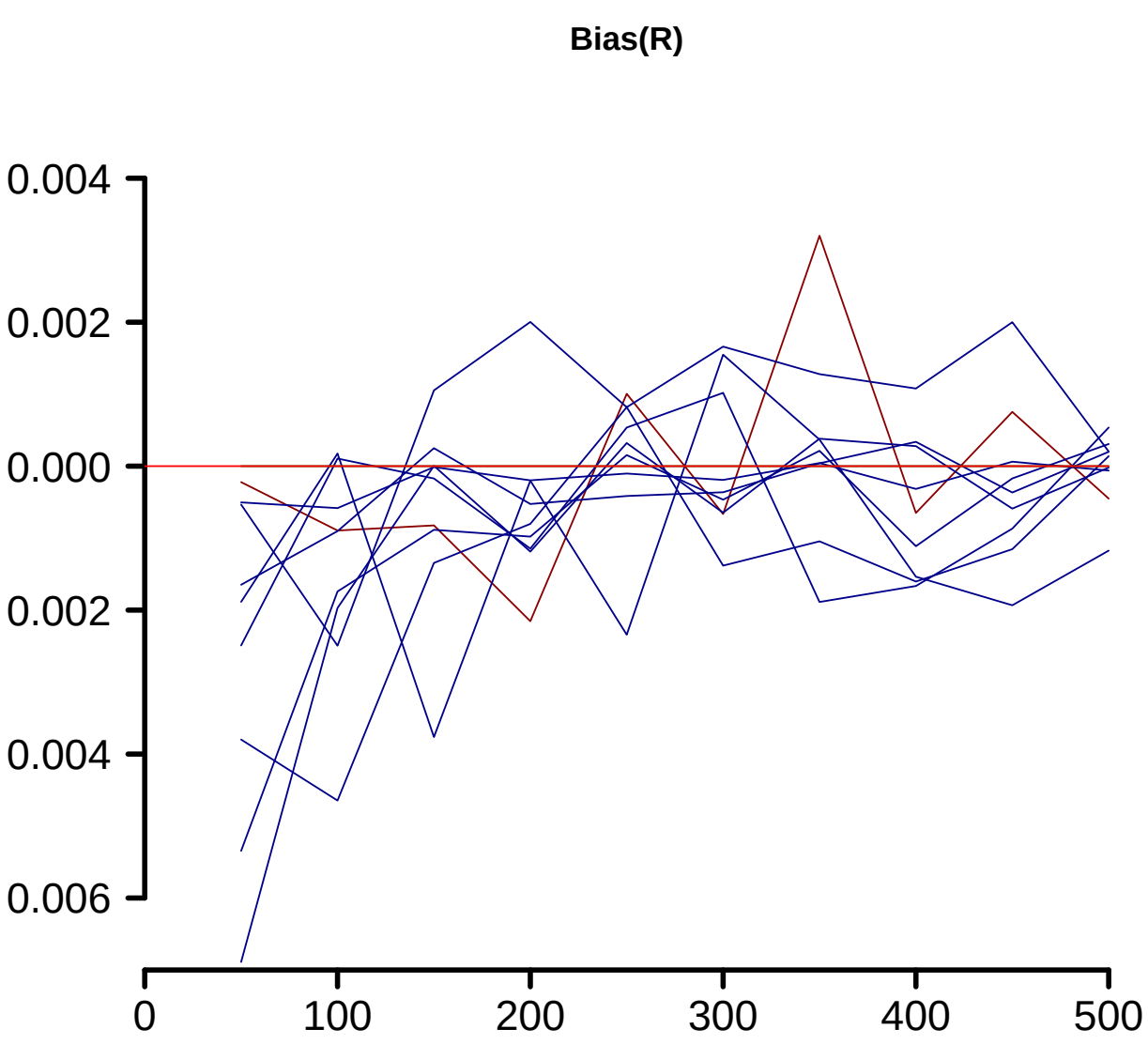
Resultatet ovan var förvånande men då vi kollar med (**???**) så visar hans artikel samma sak, dock utan att gå in på detaljer om bias (fokuserar istället på varians).

Vi vill nu dock göra en mkt enkel simulering med bara två variabler och ingen validering. Vi gör det direkt utan paketet för att säkerställa att vi får motsvarande resultat och minskar därmed risken för programeringsfel och dylikt. Vi gör detta både för R2 (dvs vanlig korrelation mellan de båda variablerna då vi bara har just två variabler) men också för korrelationen (dvs R).

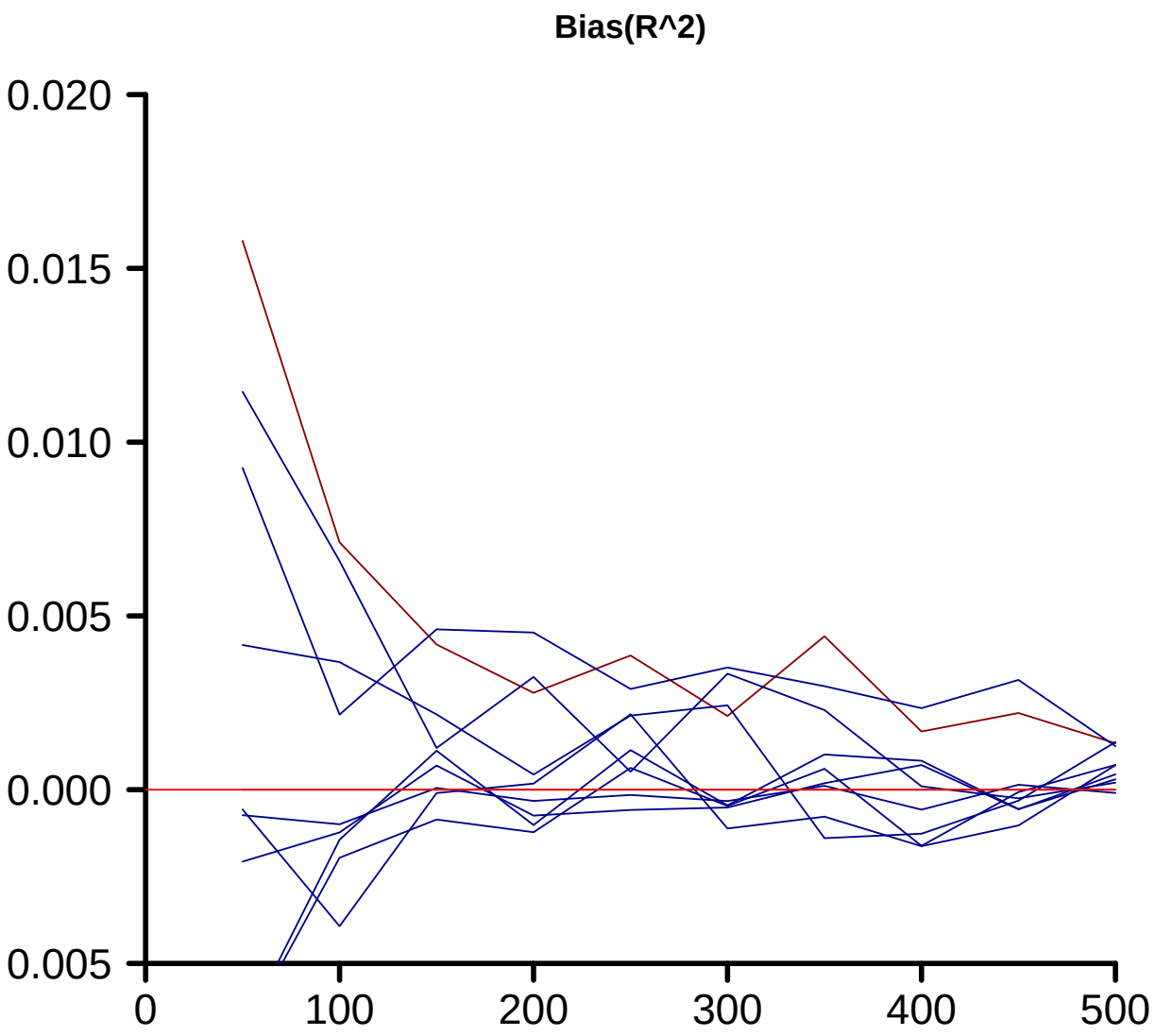
```
# Settnigs
true_r2            <- seq(.1, 1, .1)
subsample_sizes    <- seq(50, 500, 50)
repeat_subsampling <- 1000

# initiera vektorer för att kunna spara simulerade r2-värden
r_bias <- r2_bias <- r_mse <- r2_mse <- matrix(nrow = length(true_r2), ncol = length(subsample_sizes))

# Nested loop to make two matrix, one correlation matrix and one r2

# Olika sanna r2-värden
for (i in seq_along(true_r2)) {
  r          <- sqrt(true_r2[i])
  message("true_r2 = ", r ^ 2)
  big_sample <- mvtnorm::rmvnorm(1e6, sigma = matrix(c(1, r, r, 1), ncol = 2))
  big_sample <- as.data.frame(big_sample)

  # Med olika stickprovsstorlekar
  for (j in seq_along(subsample_sizes)) {
    message("subsamplesize = ", j)

    # Upprepa 1000 ggr
    tmp_rk <- numeric(repeat_subsampling)
    for (k in seq_len(repeat_subsampling)) {
      subsample   <- dplyr::sample_n(big_sample, subsample_sizes[j])
      tmp_rk[k]   <- cor(subsample$V1, subsample$V2)
```

```
    }
    r_bias[i, j]  <- mean(tmp_rk     - r)
    r2_bias[i, j] <- mean(tmp_rk ^ 2 - r ^ 2)
    r_mse[i, j]   <- mean((tmp_rk     - r) ^ 2)
    r2_mse[i, j]  <- mean((tmp_rk ^ 2 - r ^ 2) ^ 2)
  }
}
```

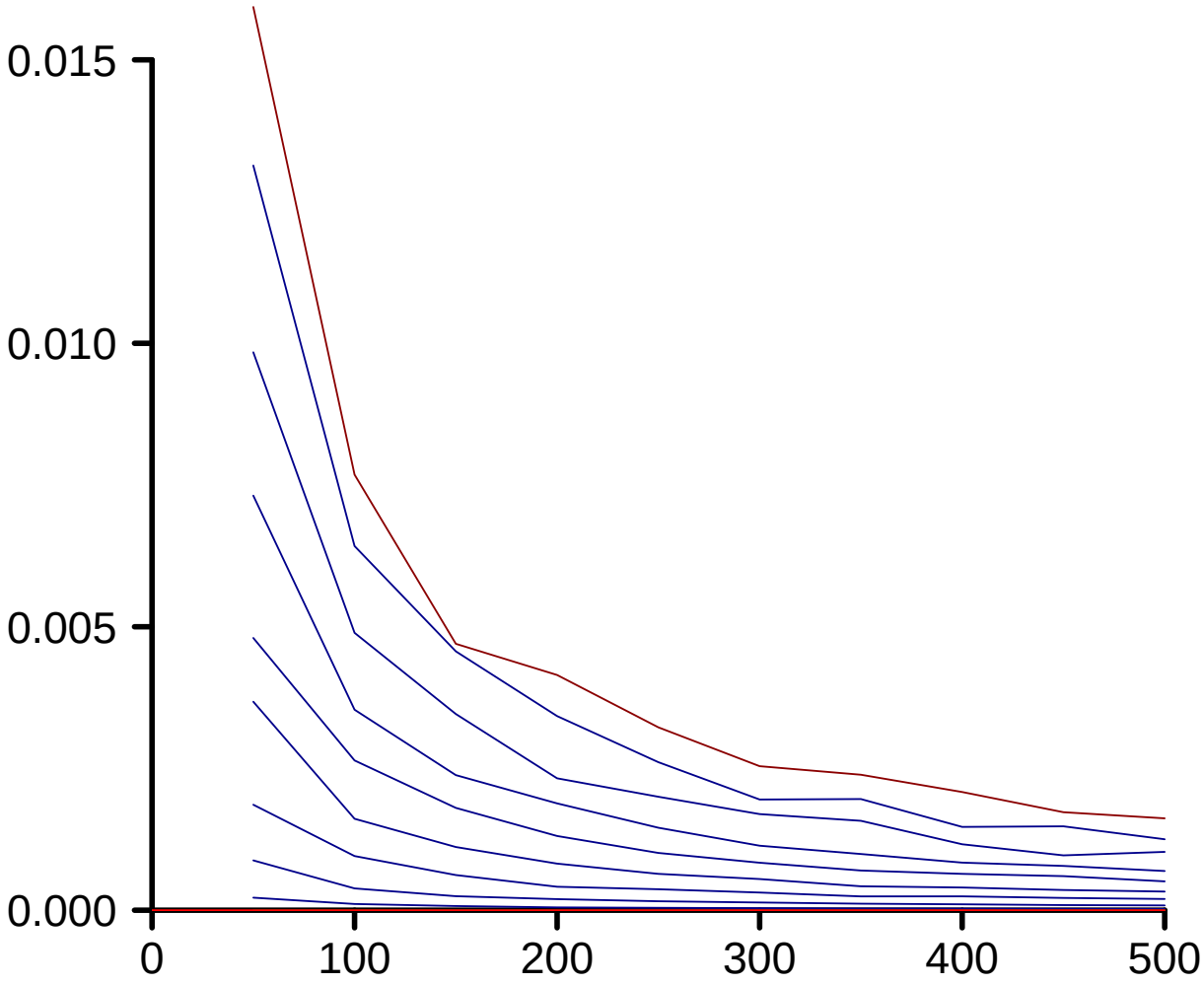
```
# Plot för R2
plot_r2_bias_sme <- function(x, ylim, main) {
  plot.new()
  plot.window(ylim = ylim, xlim = c(0, 500), xaxs = 'i', yaxs = 'i')
  axis(1, lwd = 3, cex.axis = 1.5)
  axis(2, lwd = 3, cex.axis = 1.5, las = 2)
  lines(subsample_sizes, x[1,], col = "darkred")
  for (i in 2:9) lines(subsample_sizes, x[i,], col = "darkblue")
  lines(subsample_sizes, x[10,], col = "darkgreen")
  abline(h = 0, col = "red")
  title(main = main)
}

plot_r2_bias_sme(r_bias,  c(-0.007, 0.005), "Bias(R)")
```



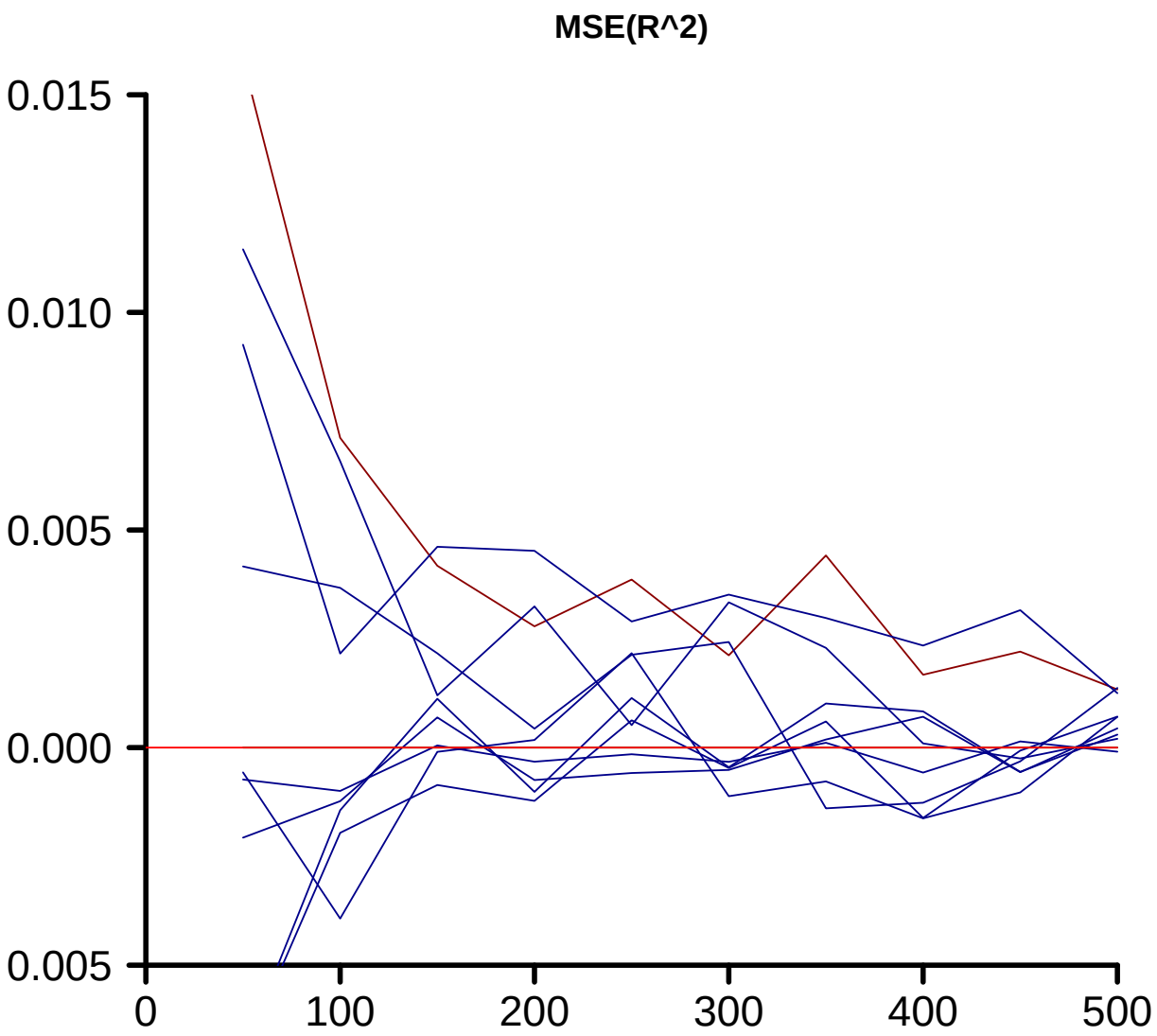

```
plot_r2_bias_sme(r2_bias, c(-0.005, 0.02), "Bias(R^2)")
```




```
plot_r2_bias_sme(r_mse,   c(0, 0.016), "MSE(R)")
```




```
plot_r2_bias_sme(r2_bias, c(-0.005, 0.015), "MSE(R^2)")
```



5.1.1 Slutsatser från graferna

1. Vad vi ser här är att sambandet mellan R och R2 förstås är deterministiskt. Vi kan framöver släppa R och fortsatt koncentrera oss på R2.
2. Vi ser att mindre R2 ger upphov till större bias (röd linje) och större R2 (grön linje) ger mindre bias. Varför? Kan undersökas närmare!
3. Vi ser också (vilket bekräftar tidigare observerat mönster enl ovan) att mindre stickprovsstorlek ger högre bias och att biasen minskar med ökad stickprovsstorlek.
4. Vi ser dock också att den bias vi får nu är bet mindre än den vi observerat ovan. Det som skiljer sig är att vi hade p = 5 ovan men här bara p = 1. SN har teori om additiv bias men detta behöver undersökas!

5.2 Additiv bias

Vi vill nu undersöka om det kan vara så att vanlig sampling "none" ger additiv bias för fler parametrar, dvs högre p. Vi gör detta mha paketet. Vi har dock ännu niget mått för bias men skapar åtm grafer där differensen framgår av avståndet mellan ref-linjen och boxploten. Se också: https://en.wikipedia.org/wiki/Mean_squared_error#Estimator Här nämns f.ö. att MSE tar lite för mkt hänsyn till outliers, vilket alltså bör få störst påverkan vid små stickprov. Där nämns därför att man ibland istället använder absolutvärde eller liknande istf kvadrat.

Referenser

Kim, Ji Hyun. 2009. "Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap." *Computational Statistics and Data Analysis* 53 (11): 3735–45. doi:10.1016/j.csda.2009.04.009.

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.