Arbetslogg 2016 vecka 7

Erik Bulow

15 februari 2016

Contents

1 Förberedelser

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

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

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

2 2016-02-15

SN läst på ordentligt föreg fredag och rekommenderar artiklar: (Warren 1971, Olkin and Pratt (1958), Hogben (1968)).

(Hogben 1968) beskriver att R2 som korrelsation mellan två bivariat normalfördelade variabler är fördelad som non central beta och att detta är välkänt sedan tidigare.

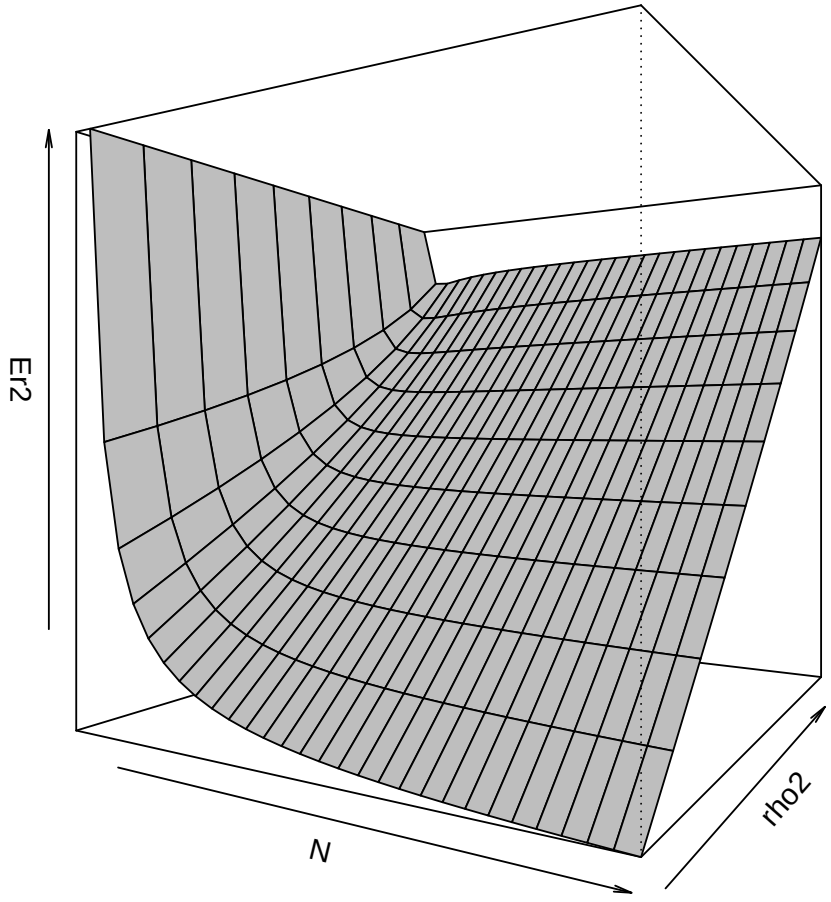
Beskrivs i (Warren 1971) att även E[r2] har känd fördelning som bero rpå N och ursprungligt rho och en hypergeometrisk fördelning (anv paketet `hypergeo` som finns på CRAN). Idé att skapa 3d-graf som beskriver denna relation. Nämns uttryckligen (s 154) att man i artikeln inte har mölighet göra större numerisk studie av hur E[r2] varierar med N, p och rho^2. Det som ändå inkluderas är dok ganska omfattande avseende selective sampling men inte random sampling. Man har heler inga 3D-grafer. Här kanske vi därmed kan tillföra ngt!? Också välkänt (men ofta förbiset) faktum att r och r2 biased för små stickprovsstorlekar. Den del av artikeln som behandlar selective sampling kan vi förstås bortse ifrån då detta inte längre är en nödv eller önskad metod. En slutsats att man inte ska skatta rho från icke slumpmässigt sample. De stickprovsstorlekar som används är rätt mkt mindre än de vi tänkt oss.

Konstaterar att (Olkin and Pratt 1958), i enlighet med SN:s initiala kommentar beskriver en unbiased version av R2 som dock ter sig ganska komplex och som inte tycks ha fått så stort genomslag efter att artikeln skrevs. Kan använda artikeln som referens men utan att vidare fördjupa mig i densamma.

2.1 E[R2]

Vi undersöker förhållandet mellan N (stickprovsstorlek), rho2 och väntevärdet av R2 gm 3d-plot

```
N <- 1:30
rho2 <- seq(0, .9, .1)
Er2_fun  <- function(N, rho2) {
  hg <- hypergeo::hypergeo(1, 1, .5 * (N + 1), rho2)
  y <- 1 - ((N - 2) / (N - 1)) * (1 - rho2) * hg
  y <- Re(y)
  y[is.infinite(y)] <- NA
  y
}

z <- outer(N, rho2, Er2_fun)
Er2 <- matrix(z, length(N), length(rho2))

persp(N, rho2, Er2, phi = 0, theta = 30, col = "gray")
```

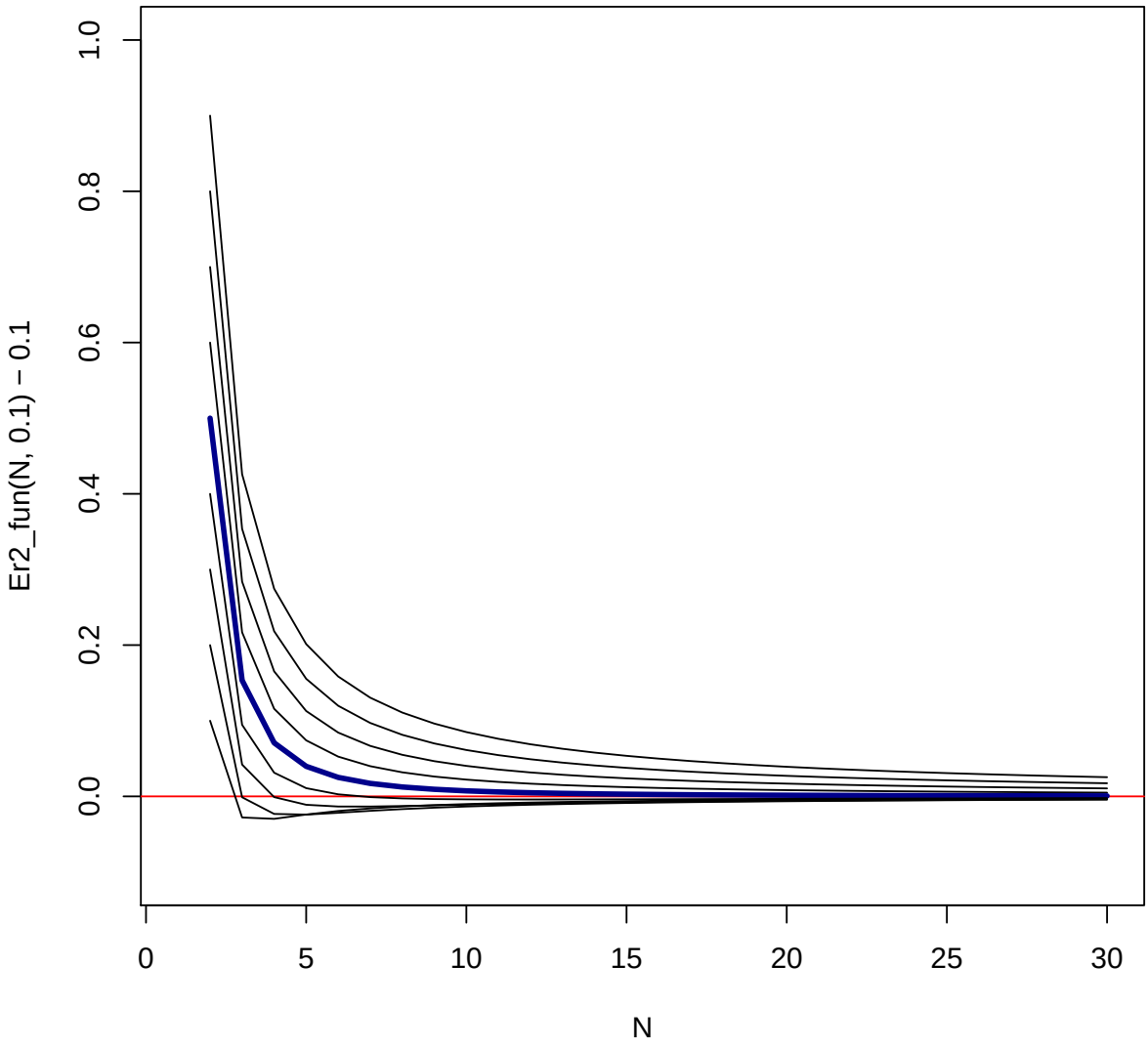


Slutsatser:

1. Vi har pos bias, dvs E[R2] > rho2
2. Biasen avtar med större N
3. Störst bias för mindre rho2
4. Biasen blir nog negligerbar för typ N = 30

Vi kan också plotta bias mot N. Vi ser här att biasen blir väldiogt liten redan då N = 30 eller t o m mindre än så. Denna bias är alltså inte förklaringen till varför vi får överskattade R2-värden. Denna graf kan jämföras med (Warren 1971) och kan då tolkas som att vi får bet bättre resultat när vi inte använder den selective samlping som beskrivs där (motsv då deras p -> Inf).

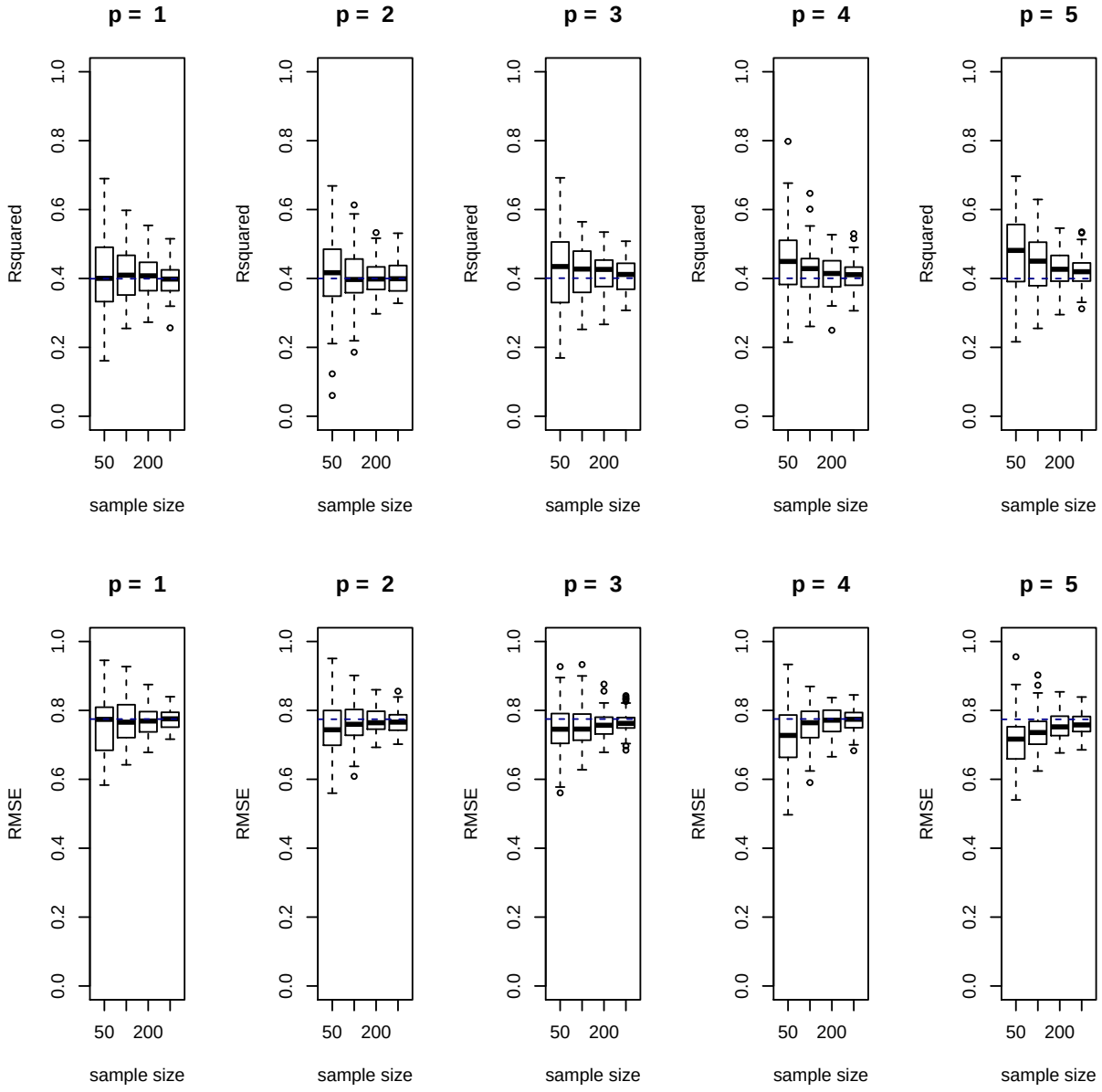
```
plot(N, Er2_fun(N, .1) - .1, type = "l", ylim = c(-.1, 1))
abline(h = 0, col = "red")
lines(N, Er2_fun(N, .2) - .2, type = "l")
lines(N, Er2_fun(N, .3) - .3, type = "l")
lines(N, Er2_fun(N, .4) - .4, type = "l")
lines(N, Er2_fun(N, .5) - .5, type = "l", col = "darkblue", lwd = 3)
lines(N, Er2_fun(N, .6) - .6, type = "l")
lines(N, Er2_fun(N, .7) - .7, type = "l")
lines(N, Er2_fun(N, .8) - .8, type = "l")
lines(N, Er2_fun(N, .9) - .9, type = "l")
```



2.2 Undersökning av bias för olika p

Vi undersöker hur mkt vi avviker från verkliga värdet med bara vanlig samplnig. Kan vi här se ett mönster som tyder på additiv bias med ökat antal oberoende variabler?

```
par(mfcol = c(2, 5))

sim_p <- function(p = 1) {
sim_data(.4, p = p) %>%
  subsamples(N = 100) %>%
  metrics(method = "none", n.sample = c(50, 100, 200, 300)) %>%
  plot(main = paste("p = ", p), ylim = 0:1)
}

lapply(1:5, sim_p)
```



Vi ser att vår bias ökar med p. Med linjär regression och med angiven kovariansmatris i `sim_data` beräknar vi vårt R2 som korrelation mellan två variabler X och Y där Y ~ N(0, 1) och X = sum(Xi) ~ N(0, p). Defenitionen R2 = cor(X, Y)^2 ger då att för fixt rho ökar Cov(X, Y) med p enl Cov(X, Y) = sqrt(p) * R.

Alltså om vi säger att vi har en bivariat normalfördelning Y, X där Y ~ N(0, 1) och X ~ N(0, p) så tycks det som att bias(cor(X, Y)) ökar med p för små stickprov även om rho är fix. Denna bias -> 0 då N -> Inf.

(Hogben 1968) beskriver fallet då X och Y är oberoende, vilket vi inte har då vi på förhand simulerar dessa värden med en fix correlation.

Vi gör en simulering så att X och Y verkligen är oberoende:

```
d <- sim_data(0)
Y <- d$Y
X <- rowSums(d[, -1])
cor(Y, X)
```

```
## [1] 0.0004178273
```

```
cov(Y, X)
```

```
## [1] 0.0009340186
```

```
var(X)
```

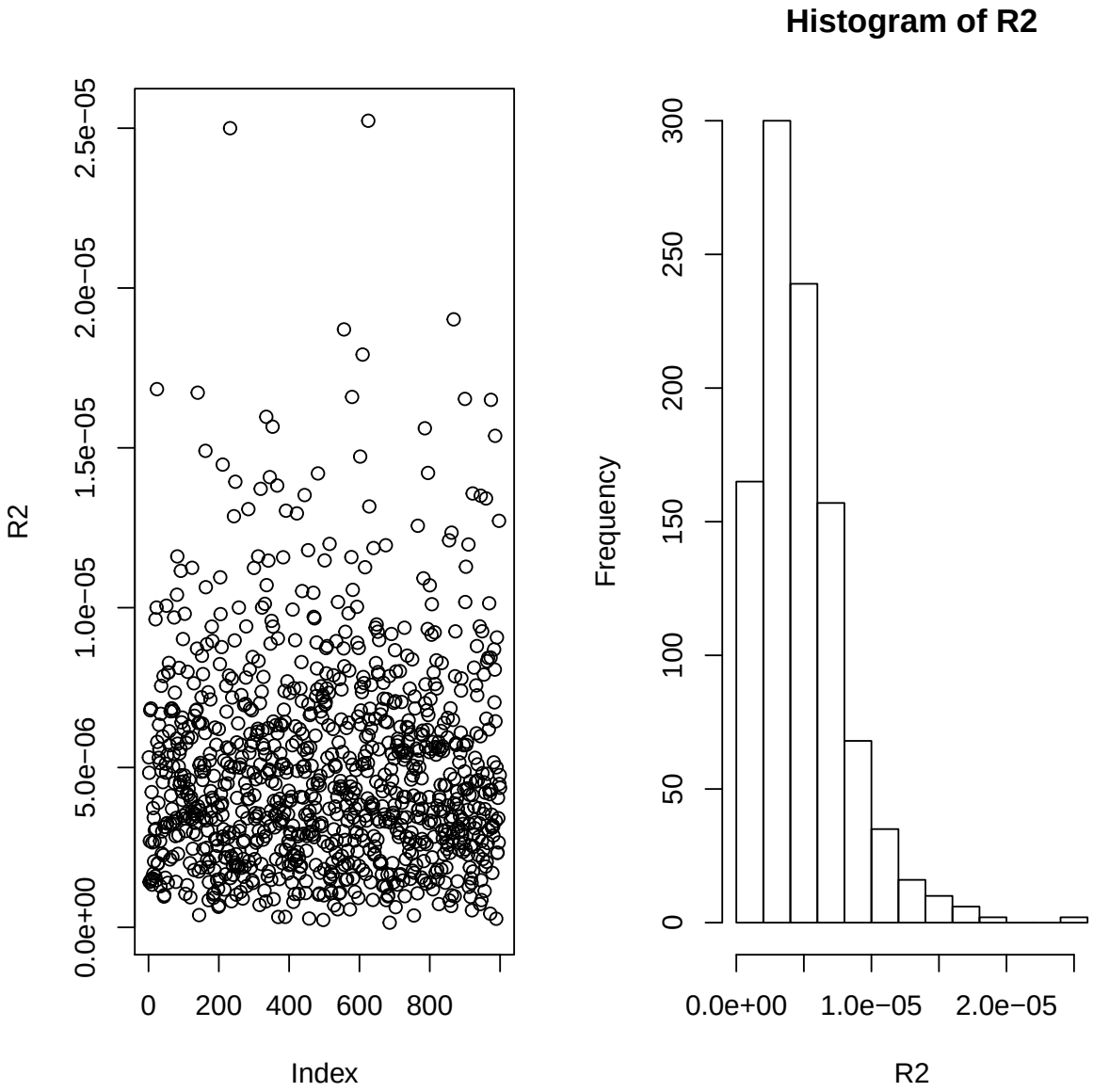
```
## [1] 5.002332
```

```
var(Y)
```

```
## [1] 0.9989529
```

```
# Om vi genererar flera sådana R2 så ska de enligt teorin förlja en given fördelning
R2 <- replicate(1000, summary(lm(Y ~., data = sim_data(0)))$r.squared)
```

```
par(mfrow = c(1, 2))
plot(R2)
hist(R2)
```



3 2016-02-16

Vi ser av bilden ovan att correlationen tycks följa ngt som liknar en exopnentialfördelning (förmodar att detta är en betafördelning enligt teorin). Det tycks också som att vi har en viss (men väldigt liten) bias. Vi har 4.9252332×10^{-6}. Vi kan jämföra detta med väntevärdet enligt (Warren 1971) 1.000001×10^{-6}. Vi har där en differens på 3.9252322×10^{-6}. Är denna diff stor eller liten?

(Hotelling 1953) förklarar ganska tydligt fördelningen av r i termer av hypergeometrisk fördelning. Han poängterar också (början av avs 3) att vårt fall med högre varians för X är likbördigt med fallet för unit variance då sådan transformation inte ska påverka "sample value of r". Funktioner beror dock på n (i artikeln = frihetsgrad = N - 1 där N = sample size). Även om vi bör få samma resultat oberoende av var(X) så kan vi alltså få en bias baserat på stickprovsstorleken. Vi får en formel för $f_n(r, \rho)$ enl (25). Denna sägs konvergera

ganska snabbt för små n och ännu snabbare för stora n. Användning av denna formell rekommenderas förutmo för extremt små n.

Kan och bör undersöka fördelning (25) pss som $E[R^2]$ ovan. Undrar om det finns en implementation av denna funktion i R?

Går också att få approximativt väntevärde och varians för arctanh(r) då $\rho = 0$ enl http://stats.stackexchange.com/questions/7218/distribution-of-sample-correlation

Från: https://en.wikipedia.org/wiki/Fisher_transformation

> The Fisher transformation is an approximate variance-stabilizing transformation for r when X and Y follow a bivariate normal distribution. This means that the variance of z is approximately constant for all values of the population correlation coefficient ρ. Without the Fisher transformation, the variance of r grows smaller as $|\rho|$ gets closer to 1. Since the Fisher transformation is approximately the identity function when $|r| < 1/2$, it is sometimes useful to remember that the variance of r is well approximated by 1/N as long as $|\rho|$ is not too large and N is not too small. This is related to the fact that the asymptotic variance of r is 1 for bivariate normal data.

```
r_dist <- function(n, r, rho) {
  stopifnot(n > 2, r >= -1, r <= 1, rho >= -1, rho <= 1)
  t1 <- (n - 1) * sqrt(2 * pi)
  g  <- gamma(n) / (gamma(n + 1 / 2))
  t2 <- (1 - rho ^ 2) ^ (0.5 * n)
  t3 <- (1 - r ^ 2) ^ (0.5 * (n - 3))
  t4 <- (1 - rho * r) ^ (0.5 - n)
  hg <- hypergeo::hypergeo(1 / 2, 1 / 2, n + 1 / 2, (1 + rho * r) / 2)
  prod(t1, g, t2, t3, t4, hg)
}

r_cumdist <- function(n, r, rho, .s_lim = 10) {
  stopifnot(n > 2, r >= -1, r <= 1, rho >= -1, rho <= 1,
            2 * (r - rho ) ^ 2 >= (1 - rho * r) ^ 2)

  # a as in [@Hotelling1953, (32)]
  a <- (1 - rho ^ 2) * (1 - r ^ 2) / (1 - rho * r) ^ 2

  # The incomplete beta function with x = b = 1 - a as in [@Hotelling1953, (50)]
  inc_beta <- function(s) pbeta(1 - a, 0.5 * (s + 1), 0.5 * (n - 1))

  # Individual terms in the infinite sum in [@Hotelling1953, (44)]
  term <- function(k, s) {
    gamma(3 / 2 - k) / (gamma(3 / 2 - k - s) * factorial(s)) *
      rho ^ s * inc_beta(s)
  }

  # N as in [@Hotelling1953, (52)]
  N <- function(k) {
    .5 * (1 - rho ^ 2) ^ k *
      sum(vapply(0:.s_lim, function(s) term(k, s), numeric(1)))
  }

  # Q as in [@Hotelling1953, (47)]
  Q <- function(n = n, r = r, rho = rho) {
    N0 <- N(0); N1 <- N(1)
```

```
    (n - 1) * gamma(n) / (sqrt(2 * pi) * gamma(n + 1 / 2)) *
      (N0 + (2 * N0 - N1) / (4 * (2 * n + 1)))
  }

  Q()
}
```

Vi kan testa att plotta `r_dist` som surface-plot enl ovan för fixt *rho* resp variabla n och r.

4 2016-02-17

Disskusion med SN: Vi behöver nästa vecka lägga upp en lite mer konkret plan för det fortsatta arbetet då det annars riskerar att svälla och bli lite för stort.

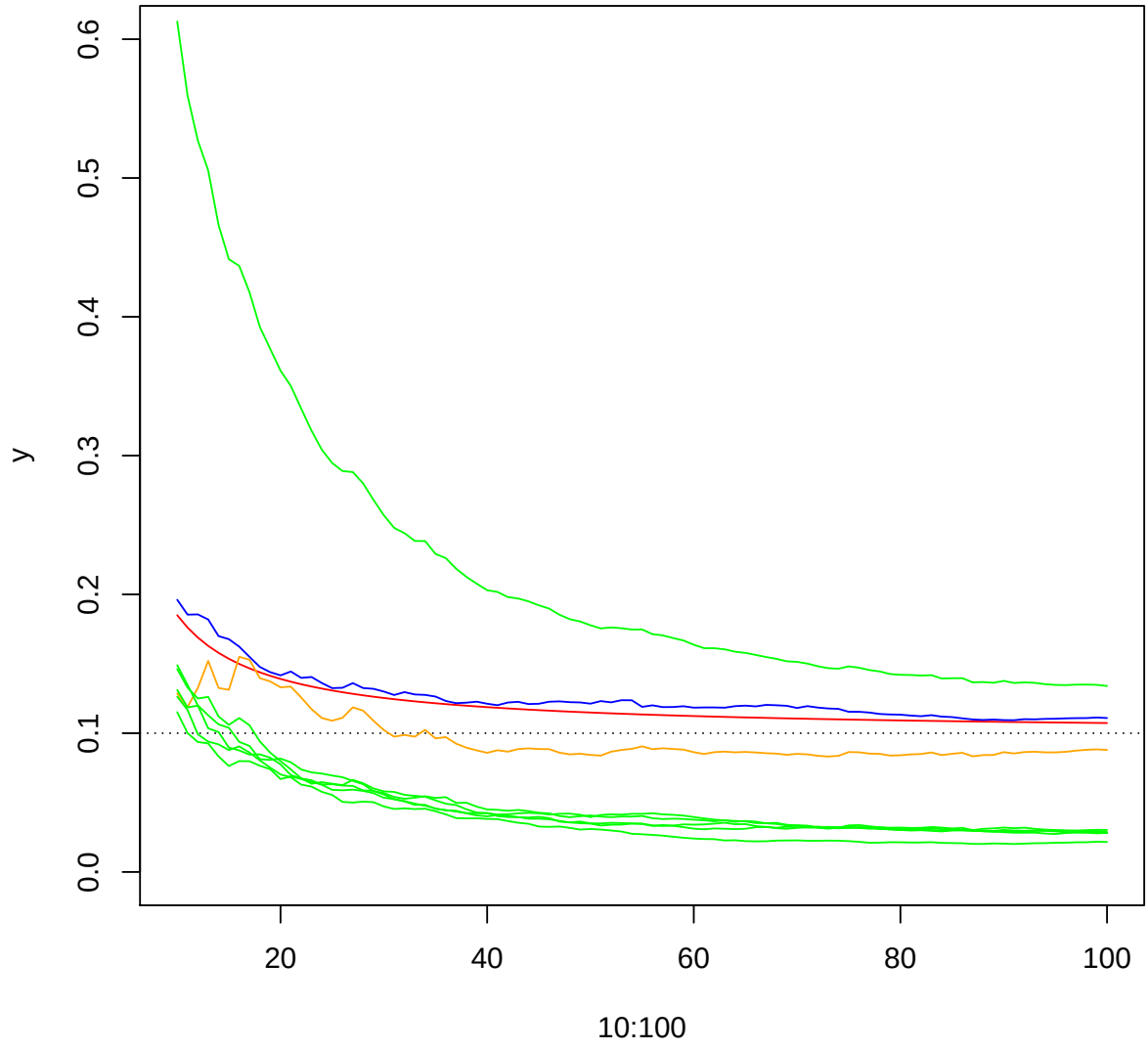
Teori att även om vi simulerar data så att $cov(X_i, X_j) = 0$ för alla $i, j : i \neq j$ så finns ett beroende mellan dem eftersom de skapas pss att deras samanvägda korrelation mot Y blir ρ. Detta låter sig inte mätas med den vanliga kovariansmatrisen men ett sådant mått finns (se: https://en.wikipedia.org/wiki/Multiple_correlation).

Plan att just nu skapa en graf med dels den teoretiska $E[r^2]$, mot n, dels boxplottar från simulering och att göra detta för olika p för att undersöka hur valet av p påverkar vår observerade bias. Detta kan vi nog göra enklast genom att utveckla `plot.mertics` för att också lägga till en linje med $E[r^2]$ som referens.

Vi kan börja med $\rho = 0$ för att åtmonstone litegrann efterlikna teorin so mbygger på oberoende observationer etc.

```
 m <- sim_data(.1, n = 10000, p = 1) %>%
        subsamples(n.max = 100, N = 100) %>%
        metrics(n.sample = seq(10, 100, 1))

  y     <- Vectorize(Er2_fun, "N")(10:100, 0.1)
  plot(10:100, y, type = "l", ylim = c(0, .6), col = "red")
  lines(x = seq(10, 100, 1), colMeans(m$Rsquared), type = "l", col = "blue")
  abline(h = .1, lty = 3)


  # Vi lägger också på en grön linje för p = 5
 ss <- sim_data(.1, n = 10000, p = 5) %>%
        subsamples(n.max = 100, N = 100)
  m2 <- metrics(ss, n.sample = seq(10, 100, 1))

  R2 <- colMeans(m2$Rsquared)
  lines(x = seq(10, 100, 1), R2, type = "l", col = "green")

  # Lägg på en linje med adjusted R2
  adj.R2 <- mapply(adj.r2, r2 = R2, n = as.numeric(names(R2)), MoreArgs = list(p = 5))
  lines(x = seq(10, 100, 1), adj.R2, type = "l", col = "orange")

  # Vi vill dessutom ha enskilda linjer för samma data men anpassad till modeller med enbart Y ~ Xi för
individual_R2 <- function(X) {
  ss1 <- lapply(ss, function(ssi) ssi[, c("Y", X)])
  class(ss1) <- "subsamples"
    metrics(ss1, n.sample = seq(10, 100, 1))$Rsquared %>%
    colMeans()
```

```
}
R2_1 <- individual_R2("X1")
R2_2 <- individual_R2("X2")
R2_3 <- individual_R2("X3")
R2_4 <- individual_R2("X4")
R2_5 <- individual_R2("X5")
lines(x = seq(10, 100, 1), R2_1, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_2, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_3, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_4, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_5, type = "l", col = "green")
```



Yipey! Här får vi en närmast perfekt överrensstämmelse! Observera att linjen anger mean och att den sammanfaller med $E[r^2]$ medan boxplottarna fokuserar på median, vilket ger ett lägre värde då r har en

assymetrisk fördelning.

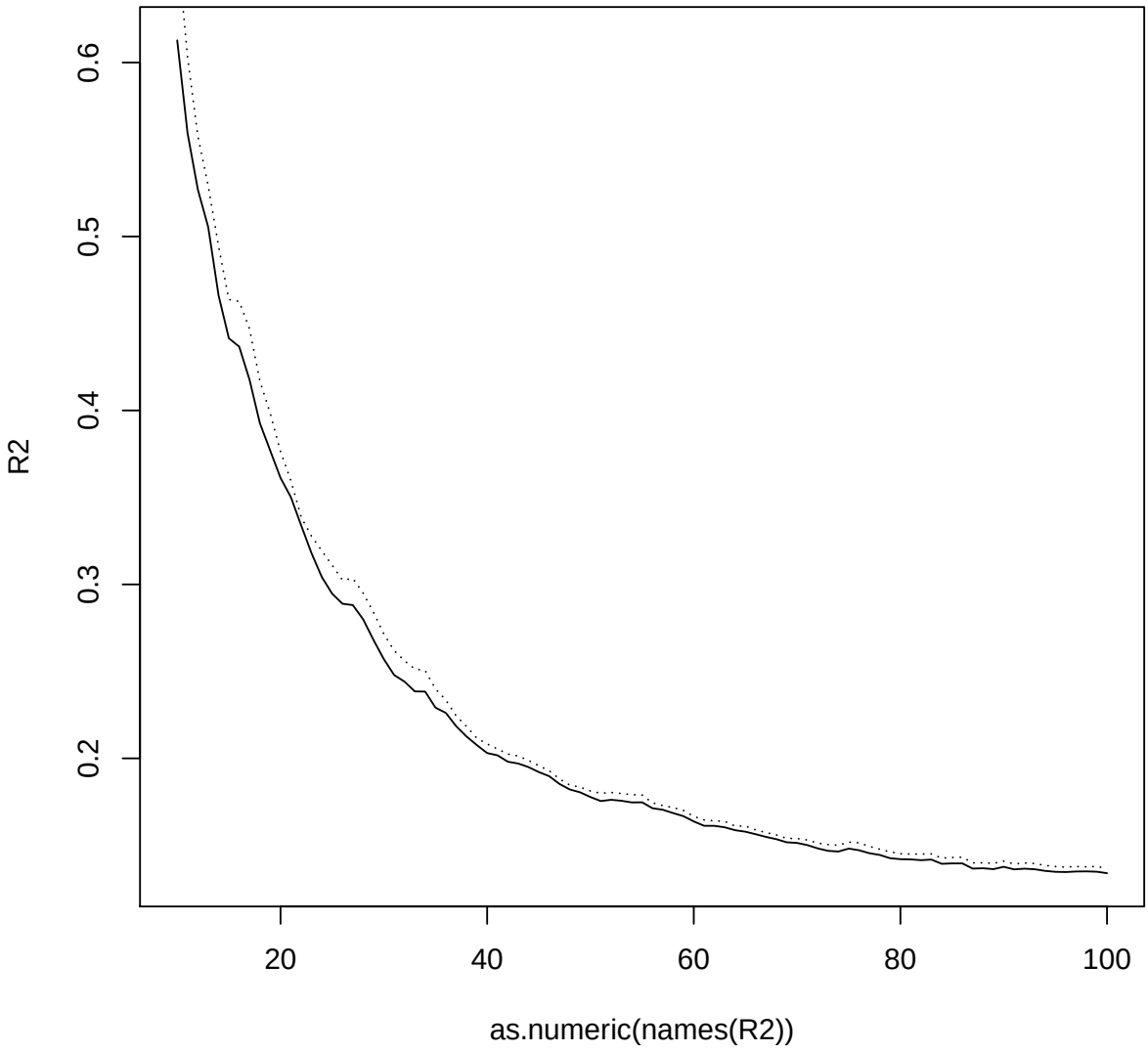
F.ö. unersökte jag och SN varför det tar så lång tid att köra koden ovan. Det är anropet till `caret::train` som tar tid. Detta är delvis diskuterat på nätet men då främst för andra mer avancerade modeller. Vi undersökte detta bla genom profiling (ett paket från RStudio). När vi kör method = "none" behöver vi egentligen inte gå via caret utan kan hårdkoda att använda lm direkt. Vi gör det för att snabba upp tiden litegrann (har nu också gjort så).

Vi ser av den oranga linjen ovan att adjusted R2 kompenserar ganska bra för den bias vi observerar. Vi bör kolla källa för hur/när/varför adjusted R2 infördes. Sannolikt var det väl just denna situation man var medvetna om?

Vi ser ovan att det tycks rimligt anta att linjärkombination av X_i för $i = 1, ..., 5$ ger fem ggr så mkt bias som för individuella modeller med Y Xi för enskilt i.

För en tydligare bild plottar vi dels R2, dels summan av alla R_i^2:n.

```
plot(as.numeric(names(R2)), R2, type = "l")
lines(as.numeric(names(R2)), R2_1 + R2_2 + R2_3 + R2_4 + R2_5, type = "l", lty = 3)
```

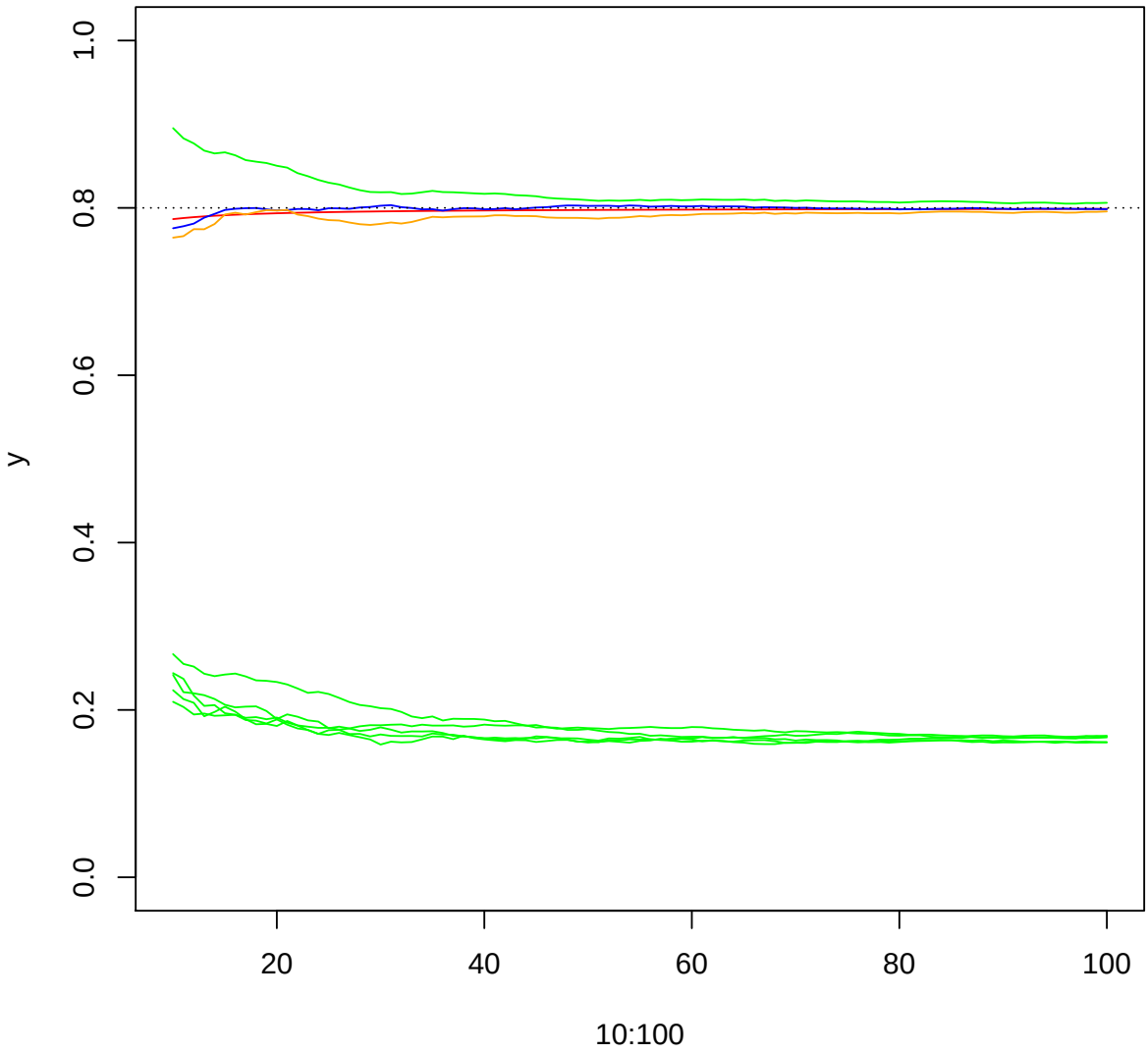


Vi ser här att dessa linjer följer varandra och drar därmed slutsatsen att biasen är additiv för p!

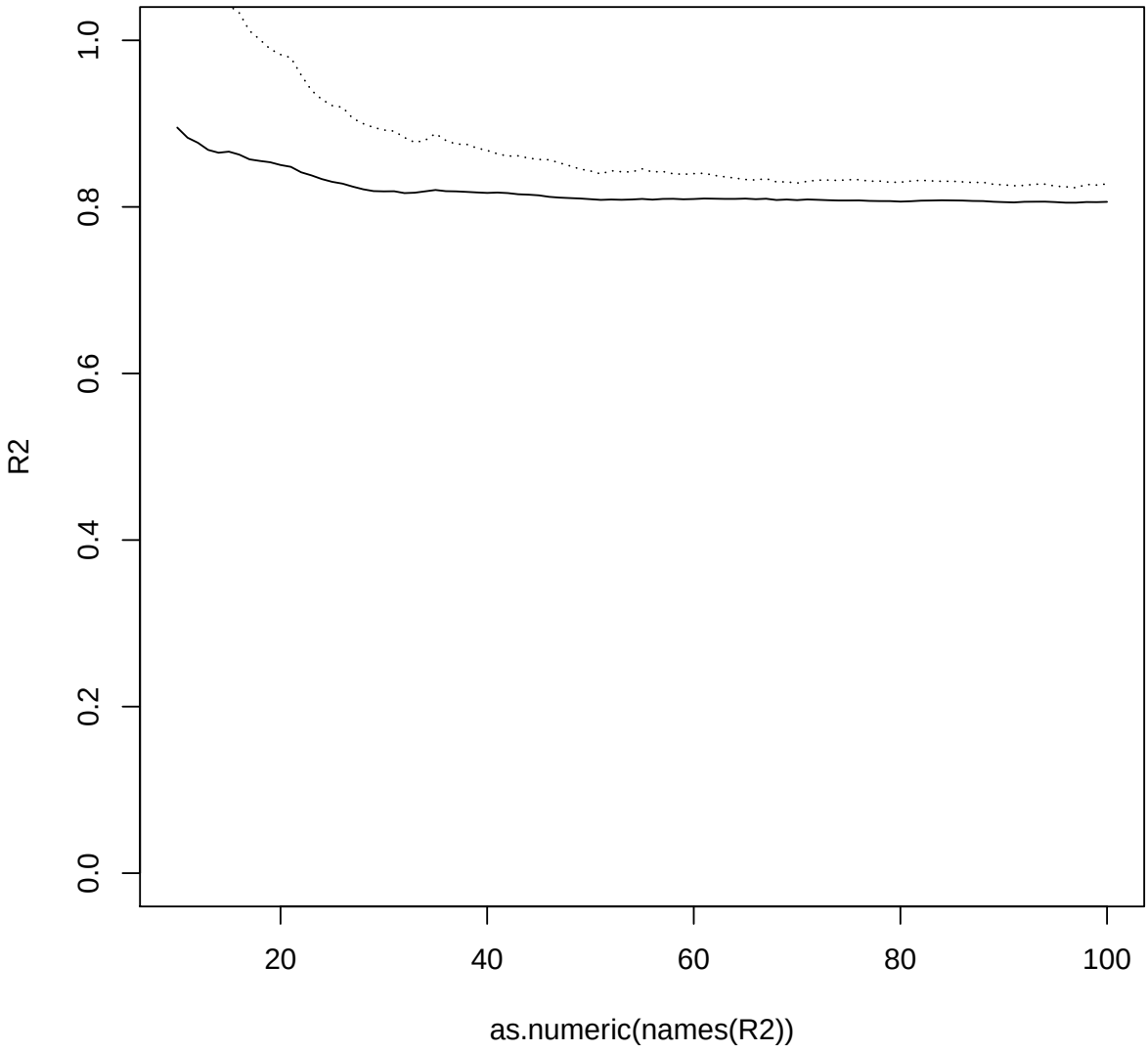
4.1 Upprepas med $\rho = 0.8$

```
m <- sim_data(.8, n = 10000, p = 1) %>%
       subsamples(n.max = 100, N = 100) %>%
       metrics(n.sample = seq(10, 100, 1))

 y     <- Vectorize(Er2_fun, "N")(10:100, 0.8)
 plot(10:100, y, type = "l", col = "red", ylim = 0:1)
 lines(x = seq(10, 100, 1), colMeans(m$Rsquared), type = "l", col = "blue")
 abline(h = .8, lty = 3)
```

```r
  # Vi lägger också på en grön linje för p = 5
 ss <- sim_data(.8, n = 10000, p = 5) %>%
       subsamples(n.max = 100, N = 100)
  m2 <- metrics(ss, n.sample = seq(10, 100, 1))

  R2 <- colMeans(m2$Rsquared)
  lines(x = seq(10, 100, 1), R2, type = "l", col = "green")

  # Lägg på en linje med adjusted R2
  adj.R2 <- mapply(adj.r2, r2 = R2, n = as.numeric(names(R2)), MoreArgs = list(p = 5))
  lines(x = seq(10, 100, 1), adj.R2, type = "l", col = "orange")

  # Vi vill dessutom ha enskilda linjer för samma data men anpassad till modeller med enbart Y ~ Xi för
individual_R2 <- function(X) {
  ss1 <- lapply(ss, function(ssi) ssi[, c("Y", X)])
  class(ss1) <- "subsamples"
    metrics(ss1, n.sample = seq(10, 100, 1))$Rsquared %>%
    colMeans()
}
R2_1 <- individual_R2("X1")
R2_2 <- individual_R2("X2")
R2_3 <- individual_R2("X3")
R2_4 <- individual_R2("X4")
R2_5 <- individual_R2("X5")
lines(x = seq(10, 100, 1), R2_1, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_2, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_3, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_4, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_5, type = "l", col = "green")
```



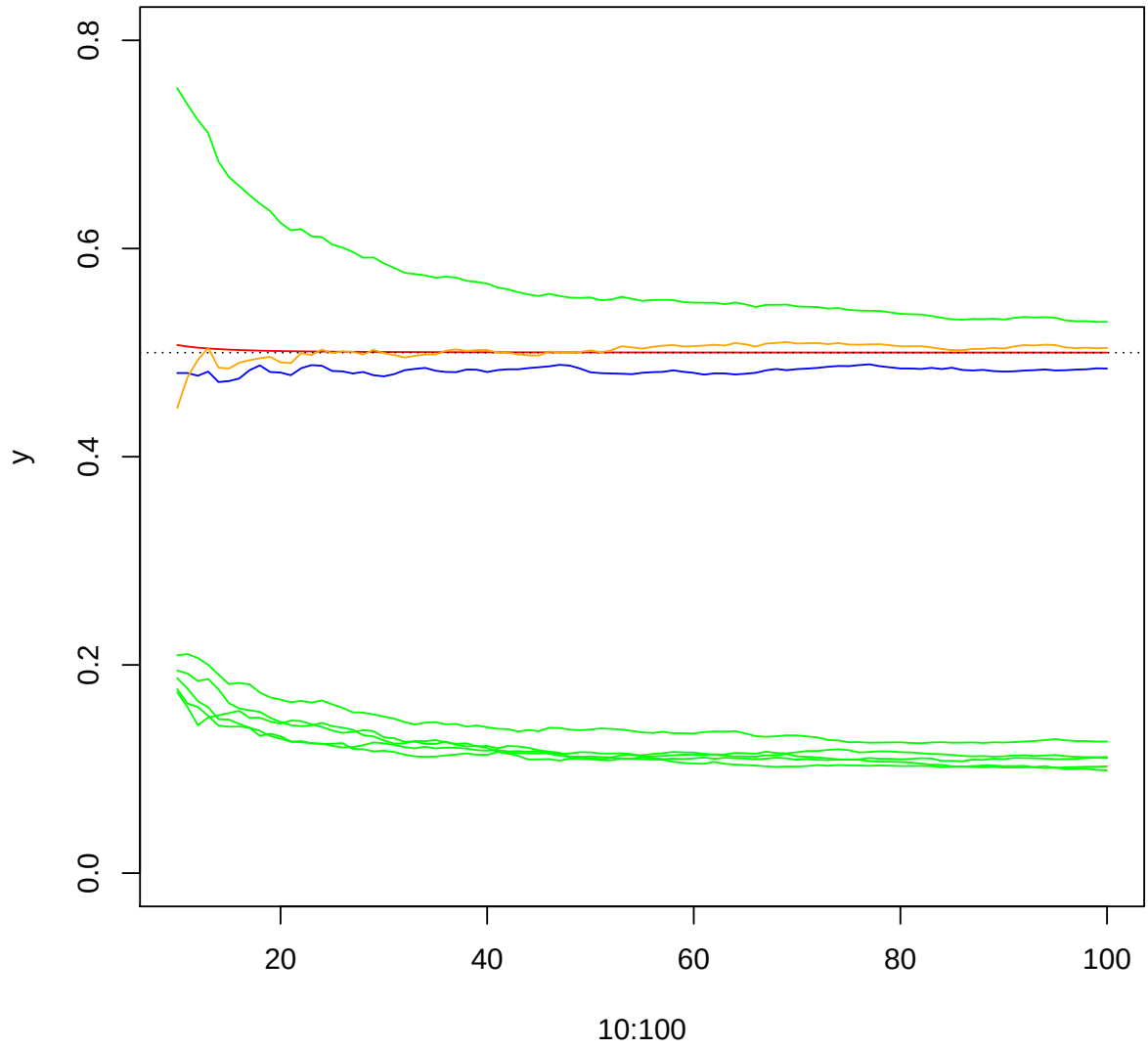

```
plot(as.numeric(names(R2)), R2, type = "l", ylim = 0:1)
lines(as.numeric(names(R2)), R2_1 + R2_2 + R2_3 + R2_4 + R2_5, type = "l", lty = 3)
```



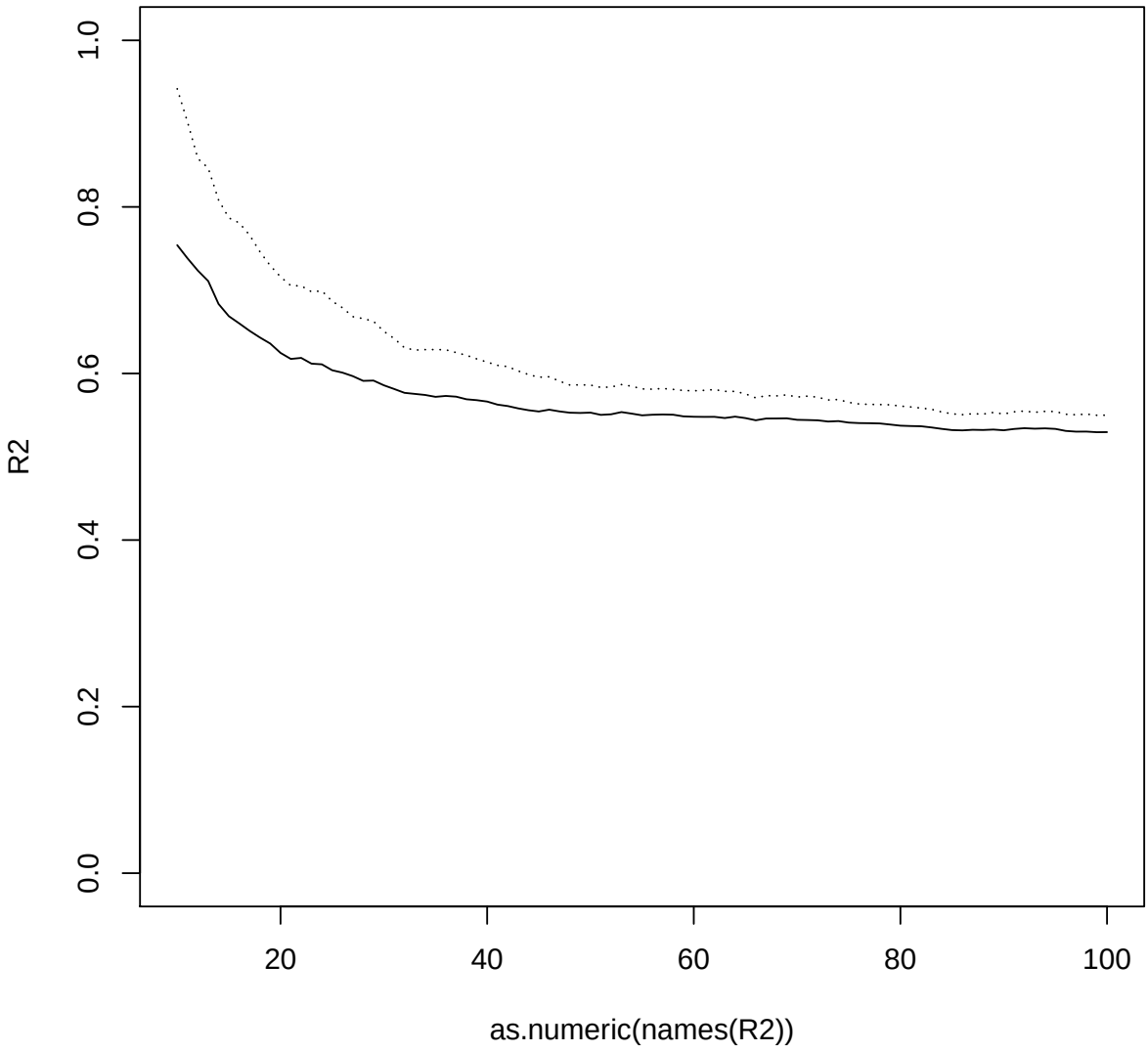
4.2 Upprepas med $\rho = 0.5$

```
m <- sim_data(.5, n = 10000, p = 1) %>%
       subsamples(n.max = 100, N = 100) %>%
       metrics(n.sample = seq(10, 100, 1))

 y     <- Vectorize(Er2_fun, "N")(10:100, 0.5)
 plot(10:100, y, type = "l", col = "red", ylim = c(0, .8))
 lines(x = seq(10, 100, 1), colMeans(m$Rsquared), type = "l", col = "blue")
 abline(h = .5, lty = 3)
```

```r
  # Vi lägger också på en grön linje för p = 5
 ss <- sim_data(.5, n = 10000, p = 5) %>%
        subsamples(n.max = 100, N = 100)
  m2 <- metrics(ss, n.sample = seq(10, 100, 1))

  R2 <- colMeans(m2$Rsquared)
  lines(x = seq(10, 100, 1), R2, type = "l", col = "green")

  # Lägg på en linje med adjusted R2
  adj.R2 <- mapply(adj.r2, r2 = R2, n = as.numeric(names(R2)), MoreArgs = list(p = 5))
  lines(x = seq(10, 100, 1), adj.R2, type = "l", col = "orange")

  # Vi vill dessutom ha enskilda linjer för samma data men anpassad till modeller med enbart Y ~ Xi för
individual_R2 <- function(X) {
  ss1 <- lapply(ss, function(ssi) ssi[, c("Y", X)])
  class(ss1) <- "subsamples"
    metrics(ss1, n.sample = seq(10, 100, 1))$Rsquared %>%
    colMeans()
}
R2_1 <- individual_R2("X1")
R2_2 <- individual_R2("X2")
R2_3 <- individual_R2("X3")
R2_4 <- individual_R2("X4")
R2_5 <- individual_R2("X5")
lines(x = seq(10, 100, 1), R2_1, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_2, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_3, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_4, type = "l", col = "green")
lines(x = seq(10, 100, 1), R2_5, type = "l", col = "green")
```




```
plot(as.numeric(names(R2)), R2, type = "l", ylim = 0:1)
lines(as.numeric(names(R2)), R2_1 + R2_2 + R2_3 + R2_4 + R2_5, type = "l", lty = 3)
```

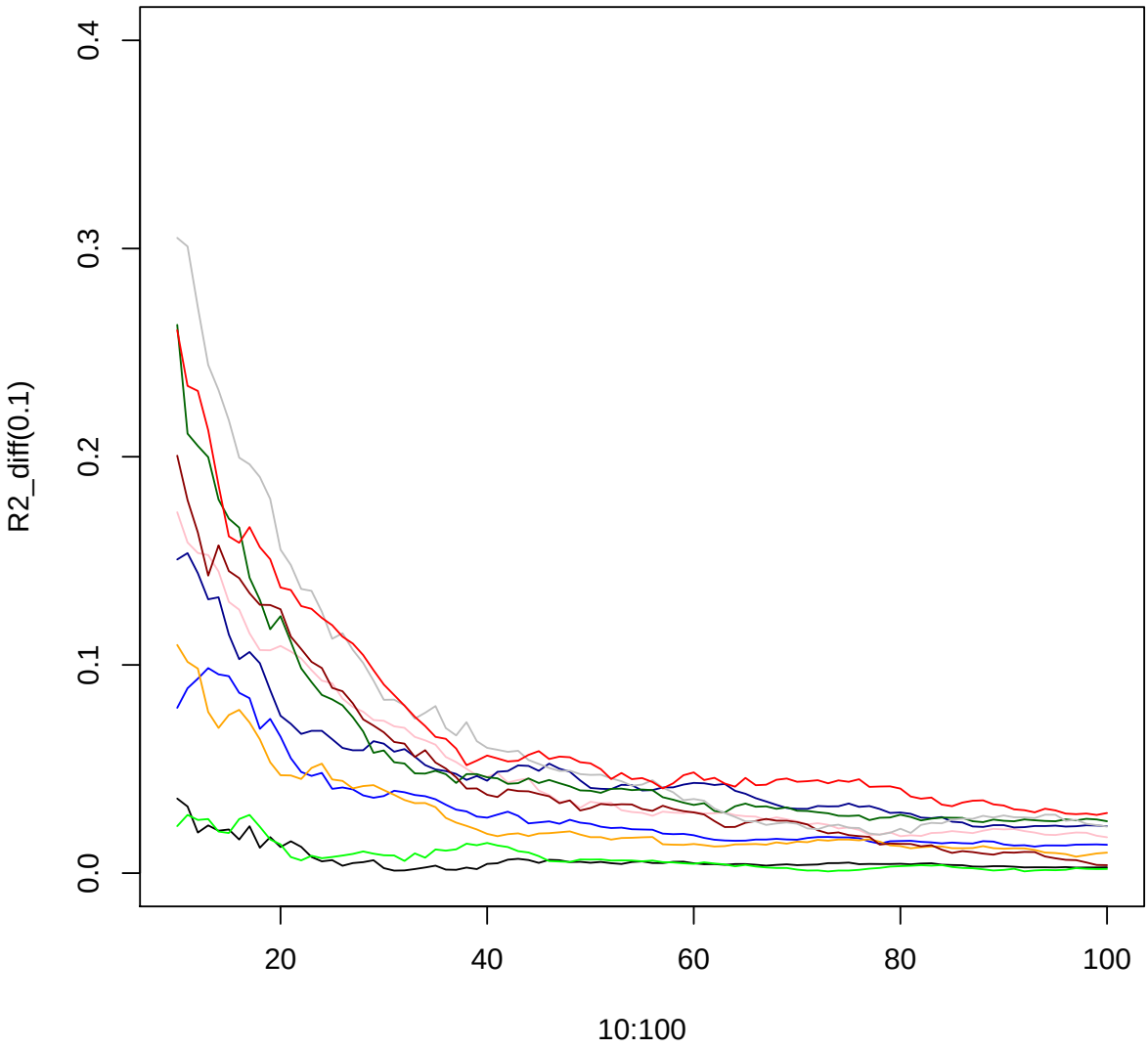


4.3 Differens mellan delar och helhet för olika ρ

Tycks som att differensen mellan summan av R_i^2:n och R^2 ökar med ökande ρ. Vi försöker skapa en graf med differenser för flera olika ρ:n.

```
R2_diff <- function(r2, n.sample = 10:100) {
  ss <- sim_data(r2, n = 10000, p = 5) %>%
          subsamples(n.max = max(n.sample), N = 100)

  R2_tot <- metrics(ss, n.sample = n.sample) %>%
    .$Rsquared %>%
    colMeans()
```

```
  individual_R2 <- function(X) {
    ss1 <- lapply(ss, function(ssi) ssi[, c("Y", X)])
    class(ss1) <- "subsamples"
      metrics(ss1, n.sample = n.sample)$Rsquared %>%
      colMeans()
  }

  (individual_R2("X1") +
    individual_R2("X2") +
    individual_R2("X3") +
    individual_R2("X4") +
    individual_R2("X5")) -
  R2_tot
}

plot(10:100, R2_diff(.1), type = "l", ylim = c(0, .4))
lines(10:100, R2_diff(.2), type = "l", col = "green")
lines(10:100, R2_diff(.3), type = "l", col = "blue")
lines(10:100, R2_diff(.4), type = "l", col = "orange")
lines(10:100, R2_diff(.5), type = "l", col = "pink")
lines(10:100, R2_diff(.6), type = "l", col = "darkblue")
lines(10:100, R2_diff(.7), type = "l", col = "darkgreen")
lines(10:100, R2_diff(.8), type = "l", col = "darkred")
lines(10:100, R2_diff(.9), type = "l", col = "gray")
lines(10:100, R2_diff(1), type = "l", col = "red")
```

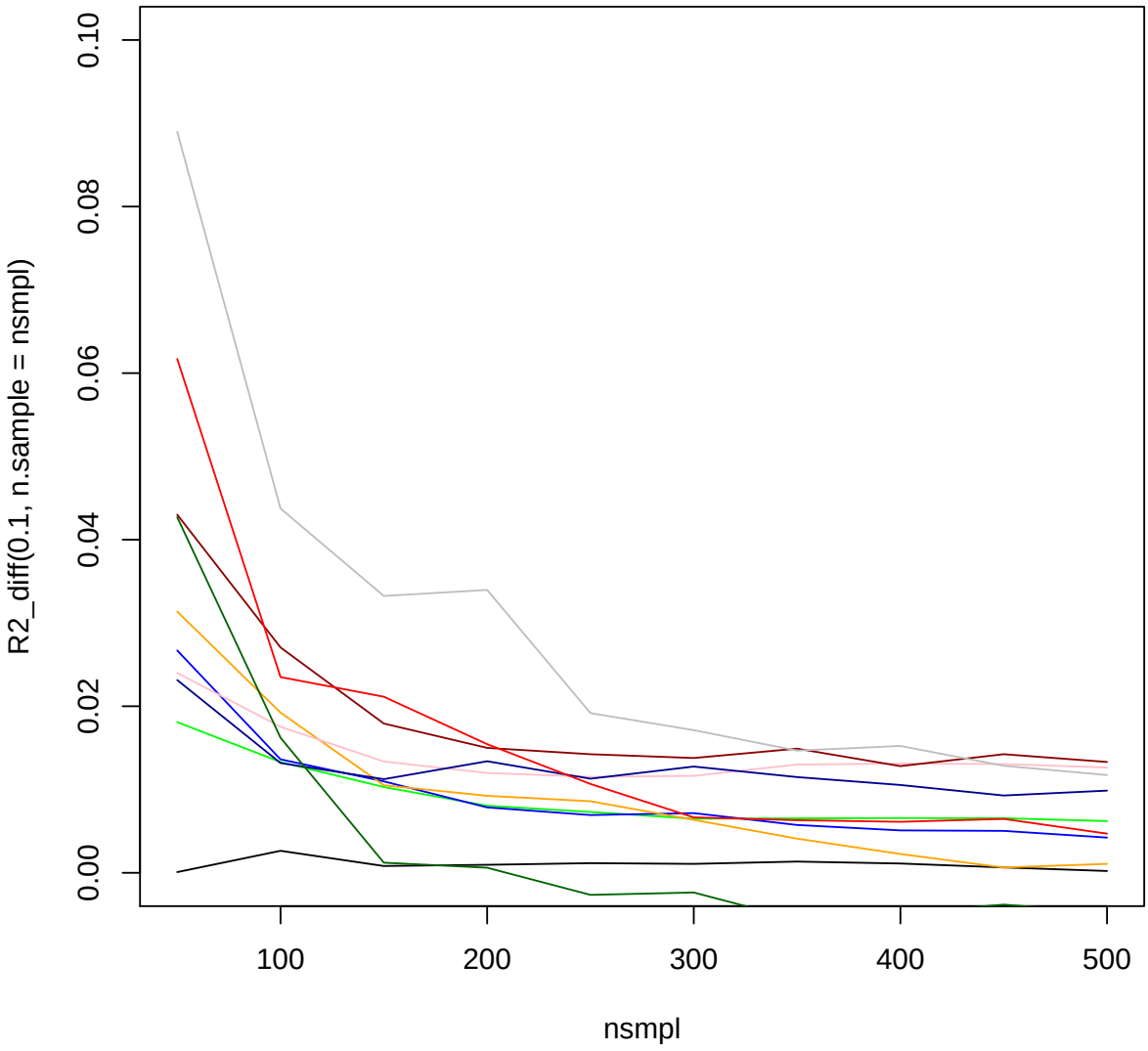


Vad vi ser här är alltså att differensen "naturligtvis" minskar med n men också att den ökar med ρ. Vi ser också att summan av delarna alltid är större än totalen. Finns det ngn relation mellan dessa referenser och $E[r^2]$ (som ju också beror på ρ). Eller om vi använder adjusted R2 istf R2?

Det tycks dessutom som att ev konvergens mot 0 sker väldigt långsamt!? Har vi ens konsistens?

```
nsmpl <- seq(50, 500, 50)
plot(nsmpl, R2_diff(.1, n.sample = nsmpl), type = "l", ylim = c(0, .1))
lines(nsmpl, R2_diff(.2, n.sample = nsmpl), type = "l", col = "green")
lines(nsmpl, R2_diff(.3, n.sample = nsmpl), type = "l", col = "blue")
lines(nsmpl, R2_diff(.4, n.sample = nsmpl), type = "l", col = "orange")
lines(nsmpl, R2_diff(.5, n.sample = nsmpl), type = "l", col = "pink")
lines(nsmpl, R2_diff(.6, n.sample = nsmpl), type = "l", col = "darkblue")
lines(nsmpl, R2_diff(.7, n.sample = nsmpl), type = "l", col = "darkgreen")
lines(nsmpl, R2_diff(.8, n.sample = nsmpl), type = "l", col = "darkred")
```

```
lines(nsmpl, R2_diff(.9, n.sample = nsmpl), type = "l", col = "gray")
lines(nsmpl, R2_diff(1, n.sample = nsmpl), type = "l", col = "red")
```



Det tycks här ganska tydligt att vår bias inte är helt additiv. Dock blir differensen rel stabil för $n > 40$ och försumbart ostabil efter ca $n > 300$. Vi ser ui den senare grafen att det också är möjligt med negativ differens (vilket5 dock fortsatt tycks väldigt ovanligt). Att vi får en större diff för större ρ kanske trots allt är naturligt om vi kan koppla det till variansen av r enligt (Hotelling 1953, 212) men borde vi inte på det stora hela i så fall se en lite mner symmetrisk diff åtm då $\rho \approx 0.5$?

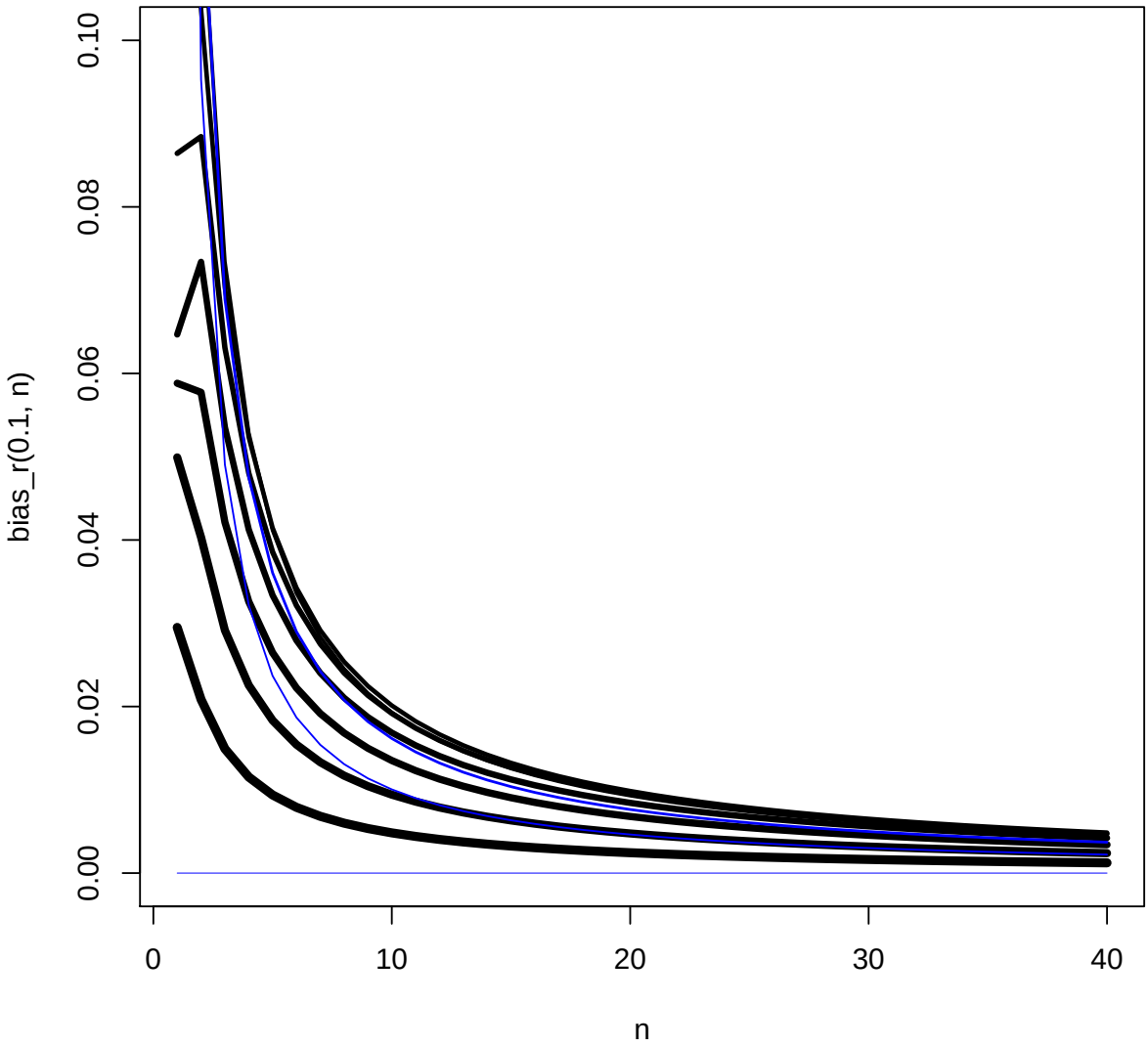
5 2016-02-18

Vi börjar med att införa några funktioner för r enl (Hotelling 1953, (p. 212)). Observera dock att vi här trunkerar de oändliga serierna väldigt snabbt. De är assymptotiska men med ganska dålig konvergens.

```
# Bias enl lambda_1
bias_r <- function(rho, n) {
   -(1 - rho ^ 2) *
    (
      -rho / (2 * n) +
        (rho - 9 * rho ^ 3) / (8 * n ^ 2) +
        (rho + 42 * rho ^ 3 - 75 * rho ^ 5) / (16 * n ^ 3)
    )
}

# sd enl sigma_r
sd_r <- function(rho, n) {
  (1 - rho ^ 2) / (sqrt(n)) *
    (1 +
       (11 * rho ^ 2) / (4 * n) +
       (-192 * rho ^ 2 + 479 * rho ^ 4) / (32 * n ^ 2)
    )
}

var_r <- function(...) sd_r(...) ^ 2
```

Gör några plottar för bias. Tjockaste linjen är lägsta ρ. Vi ser att mönstret vänder men inte är helt symmetriskt.

```
n <- 1:40
plot(n, bias_r(.1, n), type = "l", lwd = 10 / 2, ylim = c(0, .1))
lines(n, bias_r(.2, n), type = "l", lwd = 9 / 2)
lines(n, bias_r(.3, n), type = "l", lwd = 8 / 2)
lines(n, bias_r(.4, n), type = "l", lwd = 7 / 2)
lines(n, bias_r(.5, n), type = "l", lwd = 6 / 2)
lines(n, bias_r(.6, n), type = "l", lwd = 5 / 2)
lines(n, bias_r(.7, n), type = "l", lwd = 4 / 2)
lines(n, bias_r(.8, n), type = "l", lwd = 3 / 2, col = "blue")
lines(n, bias_r(.9, n), type = "l", lwd = 2 / 2, col = "blue")
lines(n, bias_r(1, n), type = "l", lwd = 1 / 2, col = "blue")
```

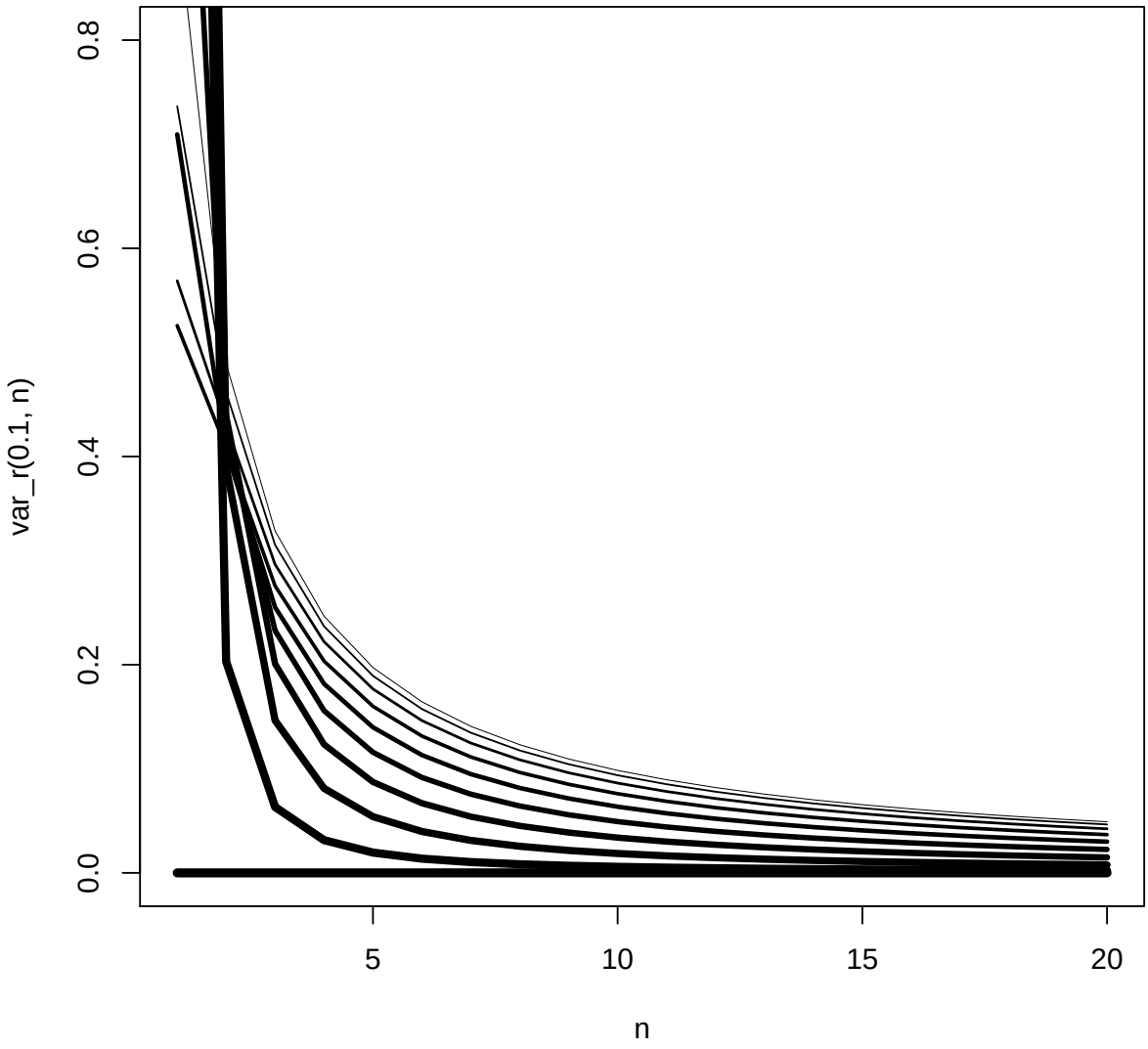


Vi ser alltså att vi har större bias för små n och stora ρ.

Vi undersöaker variansen på samma sätt. Här sätter vi istället tjockare linjer för större ρ (för att kunna urskilja linjerna).

```
n <- 1:20
plot(n, var_r(.1, n), type = "l", lwd = 1 / 2, ylim = c(0, .8))
lines(n, var_r(.2, n), type = "l", lwd = 2 / 2)
lines(n, var_r(.3, n), type = "l", lwd = 3 / 2)
lines(n, var_r(.4, n), type = "l", lwd = 4 / 2)
lines(n, var_r(.5, n), type = "l", lwd = 5 / 2)
lines(n, var_r(.6, n), type = "l", lwd = 6 / 2)
lines(n, var_r(.7, n), type = "l", lwd = 7 / 2)
lines(n, var_r(.8, n), type = "l", lwd = 8 / 2)
lines(n, var_r(.9, n), type = "l", lwd = 9 / 2)
```

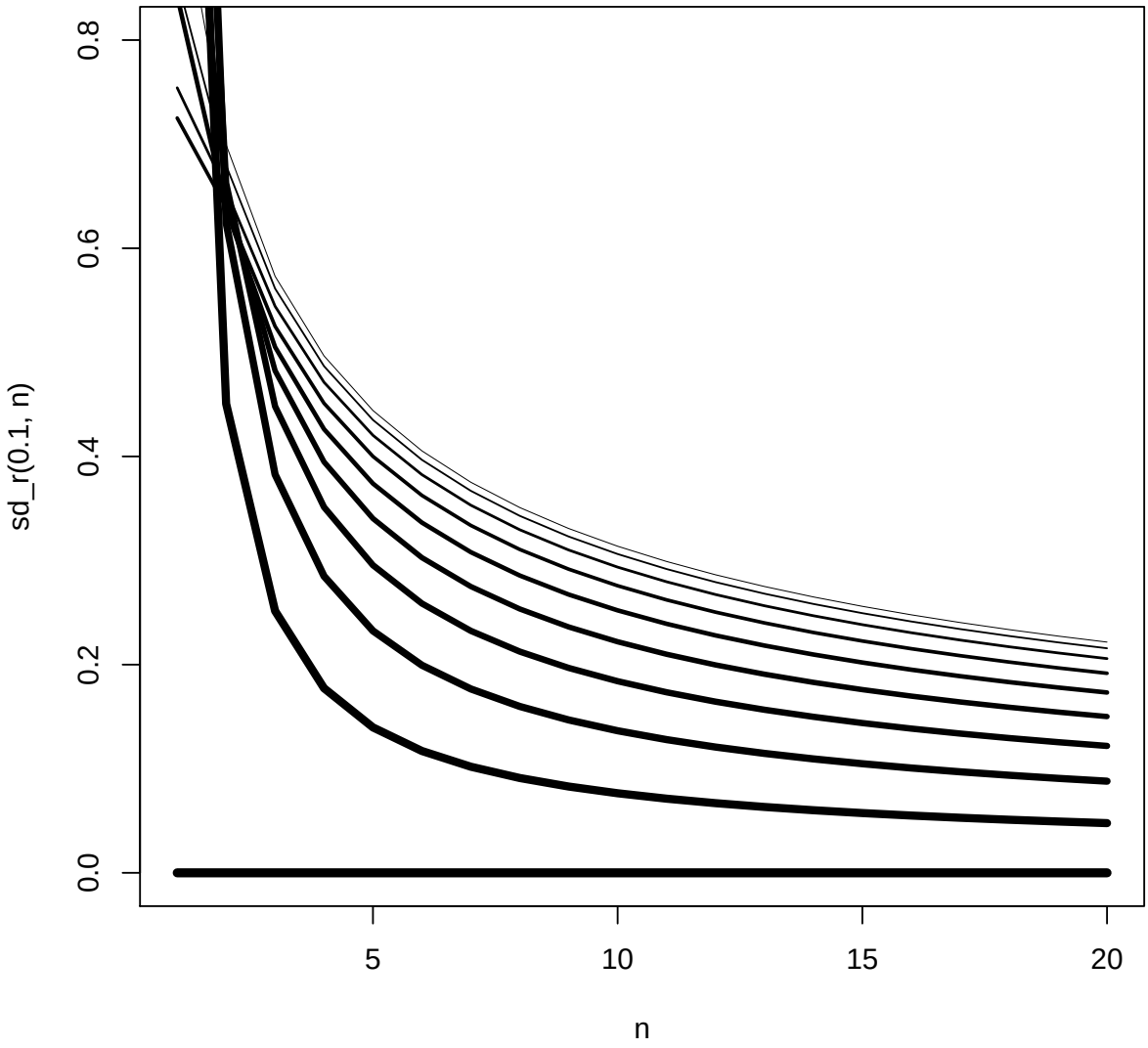
```
lines(n, var_r(1, n), type = "l", lwd = 10 / 2)
```



Här ser vi att större ρ ger mindre varians. För tydlighets skull kompletterar vi med standardavvikelser också.

```
n <- 1:20
plot(n, sd_r(.1, n), type = "l", lwd = 1 / 2, ylim = c(0, .8))
lines(n, sd_r(.2, n), type = "l", lwd = 2 / 2)
lines(n, sd_r(.3, n), type = "l", lwd = 3 / 2)
lines(n, sd_r(.4, n), type = "l", lwd = 4 / 2)
lines(n, sd_r(.5, n), type = "l", lwd = 5 / 2)
lines(n, sd_r(.6, n), type = "l", lwd = 6 / 2)
lines(n, sd_r(.7, n), type = "l", lwd = 7 / 2)
lines(n, sd_r(.8, n), type = "l", lwd = 8 / 2)
```

```
lines(n, sd_r(.9, n), type = "l", lwd = 9 / 2)
lines(n, sd_r(1, n), type = "l", lwd = 10 / 2)
```



(Och slutsatsen härav är förstås densamma.)

5.1 Läsning (Hotelling 1953)

Avs 9 konstaterar att Fischers z-transform är asymptotisk och funkar inte så bra för små n (för extremt små n kan det bli väldigt fel). Hotelling söker ett bättre alternativ men konstaterar att inget finns. Han söker också en annan funktion av r som är oberoende av n och unbiased i förhållande till ρ. Han konstaterar att nite heller en sådan finns utan att en sådan funktion måste bero på n och ett förslag, $z**$ som också ska funka för små n presenteras i slutet av avsnitt 10 (viket också avslutar artikeln).

Nämns i kommentar från Dr Irwin att Karl Pearson och Filon redan 1898 visade att standard error of r in

samples of n varied as 1 - r2. Sedan tog det lång tid innan man visade att integralen av detta uttryck = Fishers z.

Mr F J Anscombe menar har själv ingen erfarenhet av att använda Fishers z eller de tabellverk som Hotelling refererar till. Han ifrågasätter värdet av att studera fördelningen för r etc då han inte finns några direkta tillämpningar därav. Han ber dock Hotelling försvara sig :-)

Enl SN används z en hel del men framförallt för att beräkna p-värden för R2 och än mer då man använder imputation i dessa sammanhang.

5.2 Läsning av (R. Fisher and Fisher 1915)

Räknar ut ett flertal moment för r mha geometriska formler. Skriver s 519 ang första momentet:

> In effect for high values of r, where ρ^{2}" is nearly equal to unity, the form of the curve is nearly constant ...

I artikeln utvecklar Fisher formlerna stegvis för små n och för udda och jänma n etc. Han transformerar också r till ett t vars egenskaper han undersöker och allra sist nänms hans kända z.

Det noteras att fördelningen av r är väldigt skev för ρ nära +-1

5.3 Läsning av (R. A. Fisher 1921)

Känns lite oklart vad det egentligen är jag läser. Tycks börja med ett förord som lagts till senare men dessutom tycks inledningen vara skriven av ngn annan? Bör kanske dubbelkollas?

På s 207 står att r överskattar ρ (vilket ju är vad vi ser också).

Referenser

Fisher, R A. 1921. "On the probable error of a coefficient of correlation deduced from a small samlpe."

Fisher, R.a., and R.a. Fisher. 1915. "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population." *Biometrika* 10 (4): 507–21. doi:10.2307/2331838.

Hogben, David. 1968. "The distribution of the sample correlation coefficient with one variable fixed." *Journal of Research of the National Bureau of Standards, Section B: Mathematical Sciences* 72B (1): 33. doi:10.6028/jres.072B.007.

Hotelling, Harold. 1953. "New Light on the Correlation Coefficient and its Transforms Author(s): Harold Hotelling." *Journal of the Royal Statistical Society. Series B (Methodological),* 15 (2): 296–193–232.

Olkin, Ingram, and J.W. Pratt. 1958. "Unbiased estimation of certain correlation coefficients." *The Annals of Mathematical Statistics* 29 (1): 201–11. doi:10.2307/2237306.

Warren, W. G. 1971. "Correlation or Regression: Bias or Precision." *Applied Statistics* 20 (2): 148. doi:10.2307/2346463.