Arbetslogg 2016 vecka 11

Erik Bulow

14 mars 2016

Contents

1 Förberedelser

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

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

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

2 2016-03-14

2.1 Läsning av (Olkin and Finn 1995)

Utvecklar metoder för att skapa konfidensintervall runt R^2.

Vid första ögonkastet tycks artikeln fokuserad på ρ men så är inte fallet. Ställer upp ett par olika situationer där man vill jämföra olika typer av ρ^2, t ex multiple mot simple och i olika samples etc. Teoretisk ansats men med tillämpning på riktig data.

OBS! Bygger på normalfördelningsantagande som kanske funkar för stora stickprov etc men det är vi ju egentligen inte intreserade av! Skriver i conslusion att deras metoder kan rekommenderas med viss försiktighet för $n \in [60, 200]$ och först därefter med större säkerhet.

2.2 Läsning av (Yin and Fan 2001)

Jämför olika formler för adjusted R^2 (nämner även empiriska metoder men fokuserar inte på det). Slutsats att (Wherry 1931) inte är den bästa metoden utan de rekommenderar istället Pratt eller Browne. Skriver dock att (Wherry 1931) är den mest använda formeln i SAS och SPSS (och vi vet ju själva att den också används i R). De undersöker empiriskt med monte-carlo.-simulering och tycks ha ett väldigt brett spektrum av paramtervärden att slumpa utifrån.

Skriver att olika studier visar olika resultat för vilken justering som är den bästa. Finns t ex resultat som pekar åp att Browne bäst för just multipel korrelation.

Poängterar att man i mpnga studier blandat ihop (Ezekei 1929) och (Wherry 1931) och iobland även ρ^2 med ρ_c^2, vilket leder till felaktiga resultat. Nämner också att olika studier jmfrt olika metoder, vilket gör att man saknar helhetsbild samt att utvärdering mha olika data set inte nödvändigtvis heller underlättar jämförelser. Förespråkar att använda genererad syntetisk data.

Nämner att R^2 används både för förståelse mn också för predektion men att den överskattas i båda fallen, dvs att man tenderar få et lägre värde då samma modell tillämpas på nytt data set.

Skriver att man i princip alltid utgår från fixed modell i regression även om det fnins teoeri för random models. Denna teori är dock alltför komlpex för att ha fått ngt praktiskt genomslag. Detta bidrar dock också till överskattning av ρ^2 då det innebär att man ignorerar en källa till variation. Och denna variation ökar desutom extra mycket med fler variabler som antas fixa men som eg inte är det.

Nämner att om R_c, ρ_c baseras på samma modell men på dett nytt data set gäller: $E[R_c] \approx \rho_c < \rho < E[R]$

Samlpe multilpe correlation coefficient används som regel både frö R och R_c. Det är dock känt att överskattningen ökar och att vi därmed måste man "shrink" eller "correct" R för att åstadkomma adjusted R.

Nämner att det finns åtm 15 olika justerinfsformler.

Menar här liksom även tidigare källa att Smiths formel enl (Ezekei 1929) och beskrivet somså även av (Wherry 1931):

$$\hat{R}^2 = 1 - \frac{N}{N-p}(1-R^2)$$

Även den formell som ibland kallas Wherrys formell 1:

$$\hat{R}^2 = 1 - \frac{N-1}{N-p-1}(1-R^2)$$

föreslågs av Ezekei.

Om dessa formler skrivs:

> [...] cited with different names, listed here in decreasing order of frequency: the Wherry formula (Ayabe, 1985; Kennedy, 1988; Krus & Fuller, 1982; Schmitt, 1982; Stevens, 1996), the Ezekiel formula (Huberty & Mourad, 1980; Kromrey & Hines, 1996), the WhenyMcNemer formula (Newman et al., 1979), and the CohedCohen formula (Kennedy, 1988). The Wherry formula-2 was also cited in one study as an estimator for cross-validation (Kennedy, 1988). This formula is currently being implemented by popular statistical packages for computing the adjusted R2 in multiple regression procedures (e.g., SAS/STAT User's Guide, 1990; SPSS User's Guide, 1996).

Wherry formula-2:

$$\hat{R}^2 = 1 - \frac{N-1}{N-p}(1-R^2)$$

Denna formel presenterades verkligen av (Wherry 1931) men har i sin tur också kallats t ex McNemer formula och den har också misstagits för Wherry-1.

Tre olika approximationer till (Olkin and Pratt 1958) redovisas och om denna skrivs:

> These formulas were cited as the Olkin and Pratt formula in several studies (Ayabe, 1985; Claudy, 1978; Huberty & Mourad, 1980; Krus & Fuller, 1982) and were cited as the Herzberg formula in one study (Cummings, 1982).

Det som sedan kallas Pratts formula är ytterligare en approximation av (Olkin and Pratt 1958) och som är väldigt lik de tidigare.

$$\hat{R}^2 = 1 - \frac{(N-3)(1-R^2)}{(N-p-1)}\left[1 + \frac{2(1-R^2)}{N-p-2.3}\right]$$

Denna formel presenterades första gången i personlig kommunikation men finns beskriven i bl a (Claudy 1978).

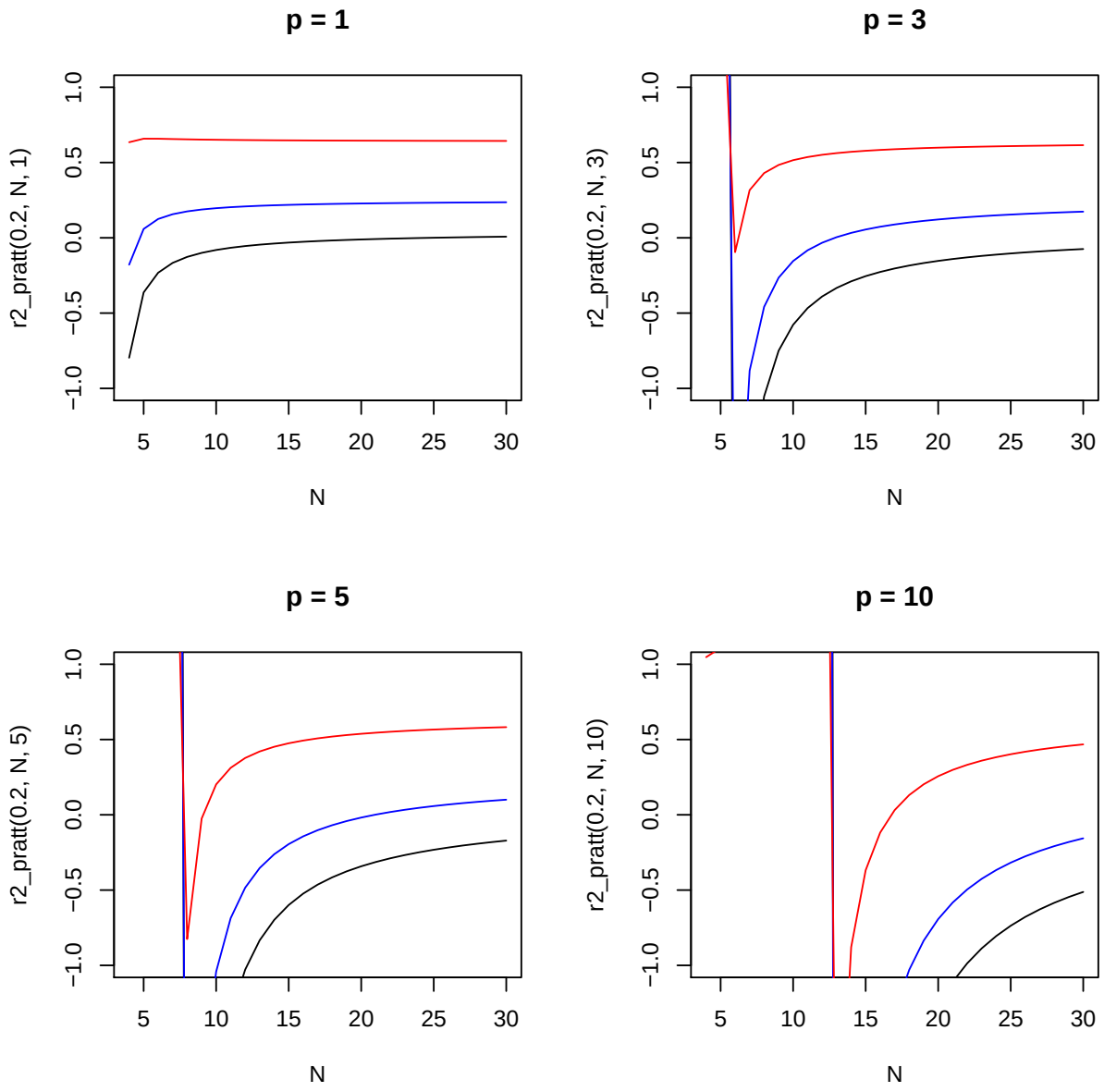
```
r2_pratt <- function(R, N, p) {
  1 - (
    ((N - 3) * (1 - R ^ 2) / (N - p - 1)) *
    (1 + (2 * (1 - R ^ 2)) / (N - p - 2.3))
  )
}
N <- 4:30

par(mfrow = c(2,2))
plot(N, r2_pratt(.2, N, 1), type = "l", ylim = c(-1, 1), main = "p = 1")
lines(N, r2_pratt(.5, N, 1), type = "l", col = "blue")
lines(N, r2_pratt(.8, N, 1), type = "l", col = "red")

plot(N, r2_pratt(.2, N, 3), type = "l", ylim = c(-1, 1), main = "p = 3")
lines(N, r2_pratt(.5, N, 3), type = "l", col = "blue")
lines(N, r2_pratt(.8, N, 3), type = "l", col = "red")

plot(N, r2_pratt(.2, N, 5), type = "l", ylim = c(-1, 1), main = "p = 5")
lines(N, r2_pratt(.5, N, 5), type = "l", col = "blue")
lines(N, r2_pratt(.8, N, 5), type = "l", col = "red")

plot(N, r2_pratt(.2, N, 10), type = "l", ylim = c(-1, 1), main = "p = 10")
lines(N, r2_pratt(.5, N, 10), type = "l", col = "blue")
lines(N, r2_pratt(.8, N, 10), type = "l", col = "red")
```



Vi ser av bilderna att vi också borde ha en lägre gräns för $/p$ kanske för att det alls ska vara meningsfullt att titta på detta.

Kallar (Claudy 1978):s formel för Claudy formula-3.

Identifierar också 9 st formler för R_c (cross-validity coefficient) som jag inte återger lika noggrant. En del av dessa formler utgår i sin tur från ρ^2 men skattar då denna i sin tur via atningen (Wherry 1931) eller (Olkin and Pratt 1958).

Här såekuleras även kring att en formel kan ha missförstås pga tryckfel (blandat ihop + och -).

Simulerar för $\rho = .2, .5, .8, p = 2, 4, 8, N = 20, 40, 60, 100, 200$.

Nämner att (Claudy 1978) konstaterat att multikolinearitet inte haft ngn större effekt i sammanhanget men detta tas ändå med som en faktor i denna undersökning. Intercorrelation mellan independent variables sattes till $.10, .30, .50$ (samma för alls kombinationer av variabler).

Resulterade i designt experiment med $3 \times 3 \times 5 \times 3 = 135$ kobinationer. Upprepades 500 ggr per kombination => 67500 replicates totalt.

All data multivariat normalfördelad.

Beskriver samplingsproceduren i SAS.

Jämförde mean och sd för alla metoder baserat på de 500 repetitionerna. Antar att skattning unbiased om mean inom $R \in [\rho \pm 0.01]$. Presenterar andel unbiased för olika metoder. Uppställningen är tydlig och klar och här framkommer att Pratt bäst och att den vanliga formlen inte särskilt bra. Bäst resultat ges då N/p stor.

Observera att (Skidmore and Thompson 2011) refererar även till denna artikel men att deras resultat är att den vanliga formeln är OK men även de menar att Pratt egentligen är bäst! Skidmore har även $N = 10$ och inkluderar även skeva fördel Osäker på hur denna skillnad uppstår. En skillnad är förstås att Skidmore inte bara tittar på normalfördelad data utan även skeva fördelningar etc. Man gör även lite noggrannare analys av bias (mer än att bara konstatera huruvida ett estimat kan betraktas som unbiased eller inte).

Här gjordes också en variansanalys av bias med slutsatser. Både enskilda och interaktinssammanslagna faktorer visar sig kunna förklara väldigt små andelar av den totala variationen men viktigaste till minst viktiga är $N, \rho^2, N\rho^2, p, Np, N\rho^2 p$. Andelen förklarad varians är större för de empiriska metoderna. Allra viktigast är dock förhållandet N/p och inte ngn enskild parameter.

Skriver att den vanliga formeln lika bra som övriga endast då $N/p \approx 100$. (Skidmore and Thompson 2011) är alltså lite snällare och säger att den vanliga formeln är OK (men säger inte att den skulle vara jättebra).

Observera att det enda vi kollar i denna artikel ju är andelen unbiased enl def medan Skidmore undersöker mer än så (vilket de också själva påpekar).

På det hela taget en välskriven och mkt intressant artikel.

2.3 Läsning av (Bobko 2001)

OBS! Om ρ, inte ρ^2

Detta är ett bokkapitell ur en bok som ev är ganska grundläggande och som handlar enbart om correlation och regression. Kanske kan vara en poäng att referera till denna för en allra första introduktion föär ovana läsare?

Gör skillnad på två olika skrivsätt där:

$$r_{X,Y} = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2 \sum (Y - \bar{Y})^2}}$$

kallas conceptual formula och där dess algebraiska motsvarighet:

$$r_{X,Y} = \frac{\sum XY - (\sum X)(\sum Y)/n}{\sqrt{[\sum X^2 - (\sum X)^2/n][\sum Y^2 - (\sum Y)^2/n]}}$$

kallas computational formula. Den första lättare att tolka, den andra att beräkna.

Påpekar att r inte är robust utan att outliers kan påverka dess värde opropertionerligt mkt. Detta är väl antagligen också värt att nämnas i sammanhanget då små stickprov diskuteras.

Jag läser inte färdigt denna text!

2.4 Läsning av (Shieh 2010)

Kommer här fram till att vanliga r trots allt kan vara bättre än unbiased versioner. Obs ρ, ej ρ^2. Undersöker:

$$RMSE = \sqrt{MSE(\hat{\rho}, \rho)} = \sqrt{E[(\hat{\rho} - \rho)^2]} = Bias(\hat{\rho}, \rho)^2 + Var(\hat{\rho})$$

.

Fig 1 och 2 ger rätt snygga illustrationer över förhållande mellan ρ och bias resp RMSE. Lätt att se mönstret! Olika linjer för olika n, $n = 20, 50, 100$.

Tycker att sample r duger för $\mid \rho \mid \leq .6$. Detta motiveras bl a med att det är beräkningsmässigt fördelaktigt jmfrt med olika adjusted versioner.

2.4.1 För ρ^2

OBS! Gäller r^2, ej R^2, dvs simple correlation, ej multiple.

$Bias(r^2, \rho^2) > 0$ for $0 \leq \rho \leq .70$ and $Bias(r^2, \rho^2 2) < 0$ for $\rho \geq .75$.

Finner att för adjusted enl (Wherry 1931) $\hat{\rho}^2_E$ (egen beteckning) och enl Pratt $\hat{\rho}^2_{PA}$ gäller (då $p = 1$):

> According to these findings, $\hat{\rho}^2_E$ is advantageous in MSE for small $\rho < .30$, r^2 dominates for $.30 \leq \rho \leq .85$ and $\hat{\rho}^2_{PA}$ performs best for large $\rho > .85$.

2.5 Läsning av (Wang and Thompson 2007)

Denna artikels upplägg liknar väldigt mkt (Skidmore and Thompson 2011) men för lite färre formler. Dock finns även här olika föfrdelningar med olika kurtosis etc. Skidmore refererade till denna artikel och utgav sig just för att vara en förbättring i förhållande till denna.

Slutsats att R^2 bara är marginellt bias och att Smiths och (Ezekei 1929) duger bra för korrigering.

Poängterar att många tidsskrifter nu kräver rapportering av effect size. Skriver att alla effektmått baserade på OLS kommer att överestimera sitt sanna värde då man får en sorts over fitting ty man får in sample bias.

Så här kan man citera de olika justerade:

> For example, in the R2 arena, the six primary correction candidates were proposed by Ezekiel (1929, 1930), Smith (as cited in Ezekiel, 1929, p. 100), Wherry (1931), Olkin and Pratt (1958), Pratt (personal communication, October 20, 1964, cited in Claudy, 1978), and Claudy (1978).

Annan bra förklarande text till varför storleken av ρ påverkar skattningen av sig själv:

> Although the reasons why sample size and the number of measured variables impact sampling error are intuitively straightforward, least obvious is the reason population effect size impacts sampling error in estimating effect sizes. Thomp- son (2002) explained, As an extreme heuristic example, pretend that one was conducting a bivariate r2 study in a situation in which the population r2 value was 1.0. In this population scat- tergram, every person's asterisk is exactly on a single regression line. In this in- stance, even if the researcher draws ridiculously small samples, such as n = 2 or n = 3, and no matter which participants are drawn, the researcher simply cannot incorrectly estimate the variance-accounted-for effect size. That is, any two or three or four people will always define a straight line in the sample scattergram, and thus r2 will always be 1.0. (p. 68)

Bobko, Philip. 2001. "A Review of the Correlation Coefficient and its Properties." In *Correlation and Regression: Applications for Industrial Organizational Psychology and Management.*, 12–42. doi:10.4135/9781412983815.

Claudy, J. G. 1978. "Multiple Regression and Validity Estimation in One Sample." *Applied Psychological Measurement* 2 (4): 595–607. doi:10.1177/014662167800200414.

Ezekei, Mordecai. 1929. "The Application of the Theory of Error to Multiple and Curvilinear Correlation." *Journal of the American Statistical Association* 24 (165): 99–104.

Olkin, Ingram, and Jeremy D. Finn. 1995. "Correlations redux." *Psychological Bulletin* 118 (1): 155–64.

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.

Shieh, Gwowen. 2010. "Estimation of the simple correlation coefficient." *Behavior Research Methods* 42 (4): 906–17. doi:10.3758/BRM.42.4.906.

Skidmore, Susan Troncoso, and Bruce Thompson. 2011. "Choosing the Best Correction Formula for the Pearson r 2 Effect Size." *The Journal of Experimental Education* 79 (3): 257–78. doi:10.1080/00220973.2010.484437.

Wang, Zhongmiao, and Bruce Thompson. 2007. "Is the Pearson r 2 Biased, and if So, What Is the Best Correction Formula?" *The Journal of Experimental Education* 75 (2): 109–25. doi:10.3200/JEXE.75.2.109-125.

Wherry, R. 1931. "A new formula for predicting the shrinkage of the coefficient of multiple correlation." *The Annals of Mathematical Statistics* 2 (4): 440–57. http://www.jstor.org/stable/2957681$/backslash$npapers2://publication/uuid/F3D4916B-BB98-4094-A459-DF4387AC9610.

Yin, Ping, and Xitao Fan. 2001. "Estimating R 2 Shrinkage in Multiple Regression: A Comparison of Different Analytical Methods." *The Journal of Experimental Education* 69 (2): 203–24. doi:10.1080/00220970109600656.