

# Arbetslogg 2016 vecka 11

*Erik Bulow*

*14 mars 2016*

## Contents

<b>1</b>	<b>Förberedelser</b>	<b>1</b>
<b>2</b>	<b>2016-03-21</b>	<b>1</b>
2.1	Läsning av (Algina and Moulder 2001) . . . . .	1
2.2	Läsning av (Zimmerman, Zumbo, and Williams 2003) . . . . .	2
2.3	Läsning av (Shieh 2008) . . . . .	3
<b>3</b>	<b>2016-03-22</b>	<b>3</b>
3.1	Läsning av “Winning the publication game” . . . . .	3
<b>4</b>	<b>2016-03-23</b>	<b>4</b>
4.1	Läsning av (Alf and Graf 2002) . . . . .	4
4.2	Läsning av (Barbiero and Ferrari 2016) . . . . .	4
	<b>Referenser</b>	<b>5</b>

## 1 Förberedelser

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

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

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

## 2 2016-03-21

### 2.1 Läsning av (Algina and Moulder 2001)

Om stickprovsstorlek för KI för multipel korr coeff. Avser KI för dskillnad i  $R^2$ , dvs  $\Delta R^2$  mellan olika modeller vid tillägg av ytterligare kovariat. Om man lägger till kovariat  $X_i$  kallas  $\Delta R^2$  squared semipartial correlation coefficient for  $X_i$ .

Skriver att det under de senaste åren blivit allt vanligare att tidskrifter inom beh science kräver rapportering av effekt sizes och ibland även KI av dessa. Radar upp ett par tidsskrifter som infört sådana krav. Nämner äv en task force 1999 med dessa rekommendationer. Läger ännu inte in off referens men finns här:

Wilkinson, L., & APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. American Psychologist, 54, 594-604.

Simuleringsstudier med olika  $k, n, \rho_r^2, \rho_f^2$  där  $\rho_r^2$  för reduced model, dvs model utan  $X_i$  enl ovan och  $\rho_f^2$  för den fulla modellen inkl denna. Endast multivariat normalfördelad data. Totalt 10143 parameterkombinationer. Rättfärdigar stickprovsstorlek gm referens. Varje kombination upprepas 50000 ggr. 95 % KI skapades vid varje rep. Ref till artikel som ger gränser för att utvärdera om en skattad nivå är hög eller inte etc.

**OBS!** Förklarar också varför det räcker att slumpa data där varje variabel har varians = 1 då detta lätt kan generaliseras. Kanske kan vara bra att refera till. Förklarar också på ganska bra sätt hur kovariansmatrisen och dess element konstrueras.

Skattningen av KI tar ej hänsyn till att underliggande fördelning skev, vilket ger visst fel. Dessutom blir det lite för brett.

Med stickprovsstorlekar under 600 kommer skapat KI inte upp till en konfidensgrad på 0.95 även om så är avsikten. För  $n > 600$  blir det bättre men fortfarande långt ifrån bra. KI ger ofta underskattning av verkligt resultat. OBS! Detta gäller om man vill hitta ngn gemensam gräns som funkar i de flesta situationer. Då t ex differensen är större krävs mindre stickprov för att upptäcka detta. Artikeln innehåller omfattande tabeller som ger olika värden för olika parameterkombinationer.

Bekräftar också det vanliga att större  $p$  (här  $k$ ) kräver större  $n$ .

## 2.2 Läsning av (Zimmerman, Zumbo, and Williams 2003)

Handlar om  $r$ , ej  $R^2$ .

Undersöker bias samt korrektionsformler. Simuleringar. Testar med Olkin-Pratt etc och även Fisher Z.

Anger att redan (Fisher 1915) fann att  $E[r] = \rho - \frac{\rho(1-\rho^2)}{2n}$  och därpå föreslåg en unbiased korrektions-formel:  $\hat{\rho} = r \left[ 1 + \frac{1-r^2}{2n} \right]$ . Denna kallas "Fisher approximate unbiased estimator". Sedan kom (Olkin and Pratt 1958) med

$$\hat{\rho} = r \left[ 1 + \frac{1-r^2}{2(n-3)} \right]$$

. Dessa formler leder till (dirivera etc) att största bias (oberoende av  $n$ ) uppnås för  $\rho = \pm .577$  (vars magnitud i sin tur avgörs av  $n$ ). Poängterar att Fishers  $z$  introducerades först med tanke på skattningens varians, inte mean. Poängterar att medan man ofta finner Fishers  $Z$  på formen:

$$Z = \frac{1}{2} \ln \left[ \frac{1+r}{1-r} \right]$$

anges formeln åt andra hållet:

$$r = \frac{e^Z - e^{-Z}}{e^Z + e^{-Z}}$$

mer sällan (trots att denna används för att bli skapade konfidensintervall etc).

Ger också förhållandet mellan Pearsons och Spearmans korrelationskoefficienter samt anger att det även föreslagits att även Spearmans koef transformeras mha Fishers  $Z$ . Även mha ytterligare utv formel. Undersöker även då data inte bivariat normalfördelad. Finner bl a att Fishers  $Z$  inte är robust och ger felaktigt resultat för icke normalfördelad data.

Slutsats att korrektion bör tillämpas för små stickprov eftersom resultatet annars kan bli missledande. Dock marginellt problem åtminstone då stickprov stort.

These consequences [bias av  $r$ ] appear to be more severe than ones typically associated with non-normality in  $t$  and  $F$  tests of differences in location.

## 2.3 Läsning av (Shieh 2008)

Av abstract att döma ytterligare en valideringsstudie som jämför olika metoder för bias-justering. Även denne med slutsats att Pratt bäst.

Antar N-data.

Föreslår modifierade versioner av bef adjusted values via  $\max(0, r\hat{\rho}^2)$  och betecknar dessa med ett plus, t ex  $\rho_X^{2+}$  där X anger vilken justeringsformel som används. Jag implementerar nu dessa i R-funktionen `adjusted_r2` via argument `min0`.

Ger också approximation till  $f(R^2)$  via ordinal (central) betafördelning.

Utvärderar de olika justeringsformlerna och kommer fram till att (Olkin and Pratt 1958) bäst och bättre ju fler termer som ingår innan den oändliga summan trunckas. Dock är Pratts formel också nästan lika bra. Dessa båda är bättre än  $\hat{\rho}_{ML}^2$  enl (Alf and Graf 2002).

Att Pratt var bättre enl (Yin and Fan 2001) motiveras med att den byggde på slumpad data medan denna artikel bygger på jämförelser av teoretiska fördelningar, vilket bör vara mer exakt.

De modifierade versionerna som ger enbart tal  $\geq 0$  ger mer bias än enl ursprungliga formler. Största skillnader för små  $\rho$ . Dock syns inga sådana skillnader om man istf bias jämf MSE.

Rekommendationen blir efter sammanvägning av minsta bias/MSE och beräknings svårighet att använda  $\hat{\rho}_P^{2+}$ , dvs positivt korrigerade enl Pratt!!!

Nämner själva att det skulle vara önskvärt undersöka detta även för icke normalfördelad data!!!

## 3 2016-03-22

- Utforskar nya markdown-format etc och testar dessa i Teds ryggstudie.
- Blir introducerad till nytt projekt avseende adverse events av Szilard. Startar eget projekt och därmed egen arbetslogg för det arbetet.

### 3.1 Läsning av “Winning the publication game”

Rekommenderar att man skapar tankekarta utifrån fyra huvudgrenarna/frågorna:

- Why did we start?
- What did we do?
- What did we found?
- What does it all mean?

Empiriskt har det visats sig att artiklar tenderar ha följande antal stycken per avsnott:

Avsnitt	# stycken
Intro	2
Metod	7
Resultat	7
Diskussion	6

s 44 ff ger bra konkreta råd till vad de olika delarna bör innehålla och hur detta kan struktureras. Fyra meningar är viktigast och de kan flöja ett ganska fast mönster:

1. Första i intro: What we looked at
2. Sista i intro: What we did
3. Första i diskussion: What we found
4. Sista i diskussion: what it means - THE MESSAGE!

Se fig 5.6.

Innehåller konkret uppställning över hela artikelns föreslagna upplägg.

## 4 2016-03-23

Fortsatte läsa bok enl ovan.

### 4.1 Läsning av (Alf and Graf 2002)

Föreslår en ny ML-skattning av  $R^2$ . Har dock visats enl (Shieh 2008) att denna inte kan ersätta t ex Pratt etc.

Utgår från fördelningsfunktion enl (Fisher 1928) och använder Excel.

Poängterar att  $R^2$  inte är en ML-skattning av  $\rho^2$ , vilket däremot den nu konstruerade  $\hat{\rho}_{(ML)}^2$  är.

Skriver att (Ezekiel and Mordecai 1929) var unbiased enbart för  $\rho^2 = 0$  medan medan (Olkin and Pratt 1958) alltid unbiased.

Väljher medvetet att visa just den situation för vilken den egna skattningen är den bästa.

Skattningen är biased men faller inom parameterrummet (dvs ger inga negativa tal).

Utgår återigen från multivariat N-data. ANvänder exempaldata från betyg inom luftvapnet (data som tycks återkomma, exempelvis från (Raju et al. 1999)).

Nämner också att resultat från just (Raju et al. 1999) visade att skattningarna är ganska robusta, dvs funkar även om data skett etc. Poängterar dock att det endast rör ett exempel och att man borde utvärdera fler ggr med lite olika förutsättningar etc.

Själva artikeln helt deskriptiv. Alla formler etc i appendix.

Artikeln känns på det hela taget lite oklar och tror i övrigt inte vi behöver tillmäta den alltför stort intresse (inledningen är dock ådagogiskt skriven).

### 4.2 Läsning av (Barbiero and Ferrari 2016)

Beskriver R-paket för simulering av diskret korrelerad data, vilket jag tänker kan vara relevant för egna simuleringar. OBS! Helt nytt! Står att manuskriptet accepterats men osäkert om det publicerats annat än elektroniskt etc?

```
library(GenOrd)
```

```
## Loading required package: mvtnorm
```

```
## Loading required package: Matrix
```

```
## Loading required package: MASS
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##      select
```

```
p <- 5
r2 <- 0.2
n <- 1000

cumprob <- c(0.25, 0.5, 0.75)
marginal <- replicate(p + 1, cumprob, simplify = FALSE)
corrcheck(marginal) # check ok
```

```
## [[1]]
## 6 x 6 Matrix of class "dsyMatrix"
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1   -1   -1   -1   -1   -1
## [2,]   -1    1   -1   -1   -1   -1
## [3,]   -1   -1    1   -1   -1   -1
## [4,]   -1   -1   -1    1   -1   -1
## [5,]   -1   -1   -1   -1    1   -1
## [6,]   -1   -1   -1   -1   -1    1
##
## [[2]]
## 6 x 6 Matrix of class "dsyMatrix"
##      [,1] [,2] [,3] [,4] [,5] [,6]
## [1,]    1    1    1    1    1    1
## [2,]    1    1    1    1    1    1
## [3,]    1    1    1    1    1    1
## [4,]    1    1    1    1    1    1
## [5,]    1    1    1    1    1    1
## [6,]    1    1    1    1    1    1
```

```
Sigma <- diag(p + 1)
Sigma[1, -1] <- Sigma[-1, 1] <- sqrt(r2 / p)
# generate a sample of size n
m <- ordsample(n, marginal, Sigma)
m <- as.data.frame(m)
summary(lm(V1 ~ ., m))$r.squared
```

```
## [1] 0.1997067
```

Jag testar att inkludera detta som del i `sim_data`-funktionen i paketet.

Själva artikeln är längre än vad jag ordkar läsa men paketet i sig tycks bra!

## Referenser

Alf, E. F., and R. G. Graf. 2002. "A New Maximum Likelihood Estimator for the Population Squared Multiple Correlation." *Journal of Educational and Behavioral Statistics* 27 (3): 223–35.

doi:10.3102/10769986027003223.

Algina, J., and B. C. Moulder. 2001. "Sample Sizes for Confidence Intervals on the Increase in the Squared Multiple Correlation Coefficient." *Educational and Psychological Measurement* 61 (4): 633–49. doi:10.1177/00131640121971400.

Barbiero, Alessandro, and Pier Alda Ferrari. 2016. "An R Package for the Simulation of Correlated Discrete Variables." *Communications in Statistics - Simulation and Computation* 0918 (March): 0–0. doi:10.1080/03610918.2016.1146758.

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

Fisher, R A. 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.

———. 1928. "The General Sampling Distribution of the Multiple Correlation Coefficient." *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 121 (788): 654–73. <http://rspa.royalsocietypublishing.org/content/121/788/654.abstract>.

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.

Raju, N S, R Bilgic, J E Edwards, and P F Flee. 1999. "Accuracy of Population Validity and Cross-Validity Estimation: An Empirical Comparison of Formula-Based, Traditional Empirical, and Equal Weights Procedures." *Applied Psychological Measurement* 23 (2): 99–115. doi:10.1177/01466219922031220.

Shieh, Gwown. 2008. "Improved Shrinkage Estimation of Squared Multiple Correlation Coefficient and Squared Cross-Validity Coefficient." *Organizational Research Methods* 11 (2): 387–407.

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

Zimmerman, D, B Zumbo, and R Williams. 2003. "Bias in estimation and hypothesis testing of correlation." *Psicologica* 24 (24): 133–58. <http://www.redalyc.org/articulo.oa?id=16924109SL>  
<http://redalyc.uaemex.mx/redalyc/html/169/16924109/16924109.html> \backslash\$nh<http://www.uv.es/psicologica/articulos1.0>