Arbetslogg 2016 vecka 10

Erik Bulow

07 mars 2016

Table of Contents

# Förberedelser

# Try it out!  
memory.limit(50000)

## [1] 50000

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

# 2016-03-07

## Läsning av (Cattin 1980)

Handlar bl a om multiple correlation coefficient i CV. Står att skattningar för regressionen sker via OLS. Gör skillnad på fixed och random model men skriver om båda.

Är långt ifrån enda källan men också här ges den ganska pedagogiska formeln för som vi kanske kan återanvända:

.

Nämner att man tidigare funnit att adjusted enl (Wherry 1931) (dvs (Ezekei 1929)) har en bias som mest uppgår till $.1 / N $ om fixed (men tar inte hänmsyn till ... kanske därmed en referens att kolla upp?)

Nämner också att (Olkin and Pratt 1958) är biased ty oändlig serir trunkeras.

Nämner att det även finns flera källor som utvecklat adjusted-versioner spec för cross-validatiln men att även dessa tenderar vara biased. Här rekommenderas olika versioner för fixed resp randmo regression.

Gör också egna simuleringar i fallet då , dvs för vanliga . Finner att bias är ganska liten men att korrigernig kan behövas då .

Förespråkar att justeras mha ngn föreslagen formell hellre än via CV då detta anges ge mindre bias (och förstås mindre beräkningsintensivt).

Poängterar vikten av att OLS används vid skattning och menar att terorin falerar vid t ex stepwise linear regression och liknande (ty prediktorerna måste väljas på förhand).

## Läsning av (Crocker 1972)

Behandlar multiple correlation coefficient (dock och inte ). Poängterar att (enl (**???**))

Detta betyder att kan hamna nära 1 för stora och små . Detta kanske kan vara viktigt då det samtidigt är pedagogiskt.

Nänmer att ref [3] ger konfidensintevall för för olika och att detta även utvecklats i ref [6].

Noterar att .

## Läsning av (**???**)

Tycks vara ngt slags original för sample-fördelning av multilpe correlation coefficient.

Ficher skriver att han blev tvungen att betrakta helt nya fördelningar. Dessa var dessutom olika för olika parametervärden men efter hand insågs att det fanns ett mönster som förenade dem.

Påpekar att om så kommer för godtycklig linjärkombination . Därmed reduceras problemet att finna den multipla korrelationskoefficienten till att hitta korrelationskoefficienten mellan två variabler.

Artikeln är extremt formelrik men en viktig slutsats är att den multipla korrelationen ej beror på hela korrelationsmatrisen mellan alla variabler utan bara på den multilpa korrelationen i populationen varifrån sampling sker.

Dock är själva fördelningsformeln oerhört krånglig och utvecklas olika för olika parametervärden. Känns inte smo att detta kan ha ngn smo helst praktisk nytta i dess här föreslagna form. Nyttjar bla också Bessel-funktioner. Kollade också bland de artiklar som refererar till denna men hittade ingen som tycks ha utvecklat metoden (även om det fnins gott om referenser).

## Läsning av (**???**)

Handlar om multiple correlation coefficient med samples från N. Behandlar väntevärde och varians av sådana .

och för ges då

och för

där (dvs hälften av antalet kovariater) och (avrundat till heltal).

Påtalas också att (Fisher 1924) gav den ungefärliga approximationen:

och att detta är en ganska bra approximation åtm då n stort.

Vi kan väl här konstatera att den bias som här presenteras tycks ara den bias för vilken (Ezekei 1929) justerar!? Dock hänvisade E själv till tidigare opublicerade källor så kan inte se exakt att det var just därför.

F.ö. har vi väl sedan tidigare liknande resultat för det icke mulipla fallet och nu får vi ngt som liknar detta.

**OBS!!!** Detta känns väl som ett ganska intressant och viktigt resultat att ta med sig!?

Vi får också enl (19):

I och med dessa uttryck skulle vi alltså kunna kolla att den bias vi får överrensstämmer med detta! :-)

Nänmer också att det finns en approximation på detta uttryck sedan tidigare men visar att den inrte är tillräcklig utan att detta exakta uttryck krävs, åtminstone för små stickprov.

Sedan beräknas även motsvarande för och till artikeln finns ett editorial appendix med tabellverk över olika och .

Enligt appendix ges formlerna istället direkt map enl (i) och (ii). Nänmer att olika förf använder olika beteckningar. T ex Fisher , Wishart appendixet och vi och att dessa behöver transformeras en aning mellan de olika skrivsätten.

På det hela taget en viktig artikel känns det som.

## Läsning av (**???**)

Låt beteckna ett sample av observationer dragna slumpmässigt från icke-singulär -variat normalfördelning. Då är

den vanliga korrelationskoefficienten mellan kovariaterna och . Låt sedan vara determinanten av korrelationsmatrisen av de enkla korrelationerna och dess första kofaktor. Då ges den multipla korrelationskoefficienten mellan och som den ickenegativa kvadratroten:

Därefetr ges delvis en formel för konfidensnitervall (uttrycks dock inte helt explicit) samt tabelluppgifter för denna beronde på stickprovsstorlek och antal kovariater.

## Läsning av (**???**)

Skriver explicit att storleken av bias för unadjusted kan vara tillräckligt stor för att orsaka rejäla tolkningsproblem.

Ger en approximation av och beräknar biasen för olika givna . Observera att detta gjordes då det kanske fortfranade var lite svårare att använda den exakta formeln, vilket ju enligt ovan egentligen är att föredra. Vi skulle ju kunna komplettera dessa beräkningar med värden från den exakta formeln.

Biasen blir allra värst då .

Väldigt bra och pedagogisk artikel. Saknar dock illustrerande grafer, vilket vi skulle kunna tillföra.

Nämner att för adjusted gäller (approximativt):

dvs bias > 0 om och < 0 om . Denna bias är dock väldigt liten, den beror inte på och blir som mest . Största bias uppstår då . Bias = 0 då .

Poängterar att det inte räcker med stort n för att undvika bias utan att det krävs att förhållandet mellan p och n är bra.

## Läsning av (Ozer 1985)

Förklarar och kritiserar tolkning av mha Venn-diagram (refererar till folk som gjort det tidigare). I denna tolkning (som också uttrycks algebraiskt) mäts korrelation som delmängder av element som förekommer i bvåde X och Y (dvs diskreta fall).

Känns lite off-topic men kanske kan vara värt att nänma som en alternativ förklaringsmodell etc. Läser inte färdigt.

## Fundernigar kring hur själva rtikeln kan skrivas

Det går att modifiera template för Word-dokument som genereras av Knitr: <https://vimeo.com/89562453>

Det finns även en del färdiga -mallar i paketet rticles som kan väljas via File > New file > R Markdown.... Man kan även skapa egna templates enligt: <http://rmarkdown.rstudio.com/developer_document_templates.html>

Har vi tur så kanske den tidsskrift vi vill submitta till erbjuder template i ngt lättanvänt format. Elsviewer-artiklar har t ex en mall i rticles-paketet.

# 2016-03-08

## Läsning am betafördelning på Wikipedia

<https://en.wikipedia.org/wiki/Beta_distribution>

OBS! Berör den vanliga, centrerade. Mode (antimode få ) kan beräknas men median saknar closed form. Finns olika förenklade formler för median givna i artikeln.

Medelvärde ges av:

Om .

Det bör alltså ganska intressant att underseröka för vilka värden betafördelningen slår över från U-shaped till den "vanliga formen". Tror också att detta har nämnts ngnstans i litteraturen men kan tyvärr inte minnas var.

Variansen ges av:

Man kan också parametrisera fördelningen mha :

Betafördelningen utvecklades av Pearson men kallades då Pearson-fördelning typ 1 och hade 4 parametrar. Dock går det att transformera denna fördelning till vanlig beta (på ngt sätt).

Betafördelningen tycks ha nämnts första gången 1911.

Parametrarna kan lättast skattas mha momentmetoden (det var f.ö. en skism mellan Pearson och Fisher just angående huruvida man skulle använda detta eler maximum likelihood, vilket dock tycks mer komplicerat).

## Googlande

Finns en relevant fråga på SO som kan knytas till formel för ickecentralitetsparametern i ickecentrala betafördelningen: <http://stats.stackexchange.com/questions/58107/conditional-expectation-of-r-squared/58133#58133>

Bygger dock på ganska avancerad matematik som jag har lite svår att ta till mig. Refererar också till: (**???**) vars avsn 30 behandlar ickecentral betafördelning men in te ger ngn bra formel för . Hjälp för tolkning av SO-posten: <http://www.math.uah.edu/stat/expect/Matrices.html> Med hjälp av dessa formler borde vi kunna få en formel för fördelningen av . Dock görs inte detta i själva frågan utan här gör man istället en approximation för ett upper bound av . Är osäkler på varför. Man får ju en analytisk formel för ickecentral beta och denna i sin tur har en closed form för dess mean!?

Dock kan också noteras att beror på väntevärdet av X. Att vi ovan sett att betafördelningen ger en bra approximation till fördelningen kan nog rimligtivs bero på att vi har väntevärde = 0 för den data vi simulerat. Resultatet kan nog därmed förväntas bli annorlunda med andra väntevärden. Kanske ngt att udersöka iofs men kanske ett stickspår.

**OBS!!!** Noterar nu att ju faktiskt beror på , dvs på stickprovet, vilket ju är det resultat vi redan sett tidigare :-( Detta var alltså ett blindspår :-( ... men jag tror nu att vi kan ha tillräcklig underbyggnad för att anta att det inte går att skatta på annat sätt än via samlpe-data. Vi släpper detta spår.

Hittar inga fler frågor på SO som berör vad vi är intresserade av.

Hittar f.ö. att det tydligen finns problem med de algoritmer som används för beräkning av ickecentral beta, se (**???**) som också utvecklar en alternativ metod. Behandlar även den algoritm som används i R. Teknisk artikel. Har bara skummat.

hej och hå

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | setosa | versicolor | virginica | p | test |
| n | 50 | 50 | 50 |  |  |
| Sepal.Length (mean (sd)) | 5.01 (0.35) | 5.94 (0.52) | 6.59 (0.64) | <0.001 |  |
| Sepal.Width (mean (sd)) | 3.43 (0.38) | 2.77 (0.31) | 2.97 (0.32) | <0.001 |  |
| Petal.Length (mean (sd)) | 1.46 (0.17) | 4.26 (0.47) | 5.55 (0.55) | <0.001 |  |
| Petal.Width (mean (sd)) | 0.25 (0.11) | 1.33 (0.20) | 2.03 (0.27) | <0.001 |  |
| Species (%) |  |  |  | <0.001 |  |
| setosa | 50 (100.0) | 0 ( 0.0) | 0 ( 0.0) |  |  |
| versicolor | 0 ( 0.0) | 50 (100.0) | 0 ( 0.0) |  |  |
| virginica | 0 ( 0.0) | 0 ( 0.0) | 50 (100.0) |  |  |

## Läsning av (**???**)

# Referenser

Cattin, Philippe. 1980. “Estimation of the Predictive Power of a Regression Model.” *Journal of Applied Psychology* 65 (4): 407–14. doi:[10.1037//0021-9010.65.4.407](https://doi.org/10.1037//0021-9010.65.4.407).

Crocker, Douglas C. 1972. “Some Interpretations of the Multiple Correlation Coefficient.” *The American Statistician* 26 (2). Taylor & Francis: 31–33. doi:[10.1080/00031305.1972.10477345](https://doi.org/10.1080/00031305.1972.10477345).

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.

Fisher, R a. 1924. “The distribution of the partial correlation coefficient.”

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](https://doi.org/10.2307/2237306).

Ozer, Daniel J. 1985. “Correlation and the coefficient of determination.” *Psychological Bulletin* 97 (2): 307–15. doi:[10.1037/0033-2909.97.2.307](https://doi.org/10.1037/0033-2909.97.2.307).

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>.