Arbetslogg 2016 vecka 10

Erik Bulow

07 mars 2016

Contents

1 Förberedelser

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

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

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

2 2016-03-07

2.1 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 R^2 som vi kanske kan återanvända:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i + \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i + \bar{Y}_i)^2}$$

.

Nämner att man tidigare funnit att adjusted R^2 enl (Wherry 1931) (dvs (Ezekei 1929)) har en bias som mest uppgår till \$.1 / N \$ om x fixed (men tar inte hänmsyn till $\rho \ldots$ 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 $p = 1$, dvs för vanliga r^2. Finner att bias är ganska liten men att korrigernig kan behövas då $\rho \leq .4, N < 50$.

Förespråkar att R^2 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).

2.2 Läsning av (Crocker 1972)

Behandlar multiple correlation coefficient (dock r och inte R^2). Poängterar att (enl (Wishart, Kondo, and Elderton 1931))

$$E[R^2|\rho = 0] = \frac{p}{n-1}$$

Detta betyder att $E[R^2]$ kan hamna nära 1 för stora p och små n. Detta kanske kan vara viktigt då det samtidigt är pedagogiskt.

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

Noterar att $R^2|\rho = 0 \sim F = \frac{R^2/p}{(1-R^2)/(n-p-1)}$.

2.3 Läsning av (R. A. Fisher 1928)

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 $Y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$ så kommer $cor(Y, \hat{Y}) = \xi(\mathbf{x})$ för godtycklig linjärkombination ξ. 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).

2.4 Läsning av (Wishart, Kondo, and Elderton 1931)

Handlar om multiple correlation coefficient med samples från N. Behandlar väntevärde och varians av sådana R^2.

$$\bar{R}^2 = 1 - \frac{b}{a+b} F(1, 1, a+b+1, \rho^2)$$

och för ρ ges då

$$\bar{R}^2 = \frac{a}{a+b}$$

och för $\rho = 1$

$$\bar{R}^2 = 1$$

där $a = p/2$ (dvs hälften av antalet kovariater) och $b = (n-p-1)/2$ (avrundat till heltal).

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

$$E[R^2] = \bar{R}^2 = a - \frac{b}{a+b}(1-\rho^2)$$

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):

$$\sigma^2_{R^2} = \frac{b(b+1)(1-\rho^2)^2}{(a+b)(a+b+1)} F(2, 2, a+b+2, \rho^2) - \frac{b^2(1-\rho^2)^2}{(a+b)^2} F^2(1, 1, a+b+1, \rho^2)$$

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 R och till artikeln finns ett editorial appendix med tabellverk över olika n och p.

Enligt appendix ges formlerna istället direkt map n, p enl (i) och (ii). Nänmer att olika förf använder olika beteckningar. T ex Fisher n_1, n_2, Wishart a, b appendixet N, n och vi n, p och att dessa behöver transformeras en aning mellan de olika skrivsätten.

På det hela taget en viktig artikel känns det som.

2.5 Läsning av (Kramer 1963)

Låt $X_{ij}(i = 1, 2, \ldots, k; j = 1, 2, \ldots, n)$ beteckna ett sample av n observationer dragna slumpmässigt från icke-singulär k-variat normalfördelning. Då är

$$r_{hm} = \frac{n\sum_{j=1}^{n} X_{hj}X_{mj} - (\sum_{j=1}^{n} X_{hj})\sum_{j=1}^{n} X_{mj}}{\sqrt{[n\sum_{j=1}^{n} X_{hj}^2 - (\sum_{j=1}^{n})^2]\,[n\sum_{j=1}^{n} X_{mj}^2 - (\sum_{j=1}^{n} X_{mj})^2]}}$$

den vanliga korrelationskoefficienten mellan kovariaterna X_h och X_m. Låt sedan P vara determinanten av korrelationsmatrisen av de enkla korrelationerna och P' dess första kofaktor. Då ges den multipla korrelationskoefficienten mellan X_i och $(X_2, \ldots, X_k)$ som den ickenegativa kvadratroten:

$$R = \sqrt{1 - \frac{P}{P'}}$$

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

2.6 Läsning av (Montgomery and Morrison 1973)

Skriver explicit att storleken av bias för unadjusted R^2 kan vara tillräckligt stor för att orsaka rejäla tolkningsproblem.

Ger en approximation av $E[R^2|n, k, \rho^2]$ och beräknar biasen för olika givna n, k, ρ. 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 $\rho = 0$.

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

Nämner att för adjusted R^2 gäller (approximativt):

$$bias(\bar{R}^2) = -\frac{\rho^2(1-\rho^2)(1-2\rho^2)}{n}$$

dvs bias > 0 om $\rho \geq 1/2$ och < 0 om $\rho \leq 1/2$. Denna bias är dock väldigt liten, den beror inte på p och blir som mest $.1/n$. Största bias uppstår då $\rho = .2, .8$. Bias $= 0$ då $\rho = 0, 1/2, 1$.

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

2.7 Läsning av (Ozer 1985)

Förklarar och kritiserar tolkning av r^2 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.

2.8 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 LaTeX-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.

3 2016-03-08

3.1 Läsning am betafördelning på Wikipedia

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

OBS! Berör den vanliga, centrerade. Mode (antimode få $\alpha, \beta < 1$) kan beräknas men median saknar closed form. Finns olika förenklade formler för median givna i artikeln.

Medelvärde ges av:

$$\mu = E[X] = \frac{1}{1 + \frac{\beta}{\alpha}}$$

Om $\alpha = \beta \Rightarrow \mu = 1/2$.

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:

$$var(X) = E[(X - \mu)^2] = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Man kan också parametrisera fördelningen mha $\mu, \nu = \alpha + \beta (\nu > 0)$:

$$\alpha = \mu\nu, \beta = (1 - \mu)\nu$$

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 α, β 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).

$$\hat{\alpha} = \bar{x}\left(\frac{\bar{x}(1 - \bar{x})}{\bar{v} - 1}\right), \beta = (1 - \bar{x})\hat{\alpha}, \text{ if } \bar{v} < \bar{x}(1 - \bar{x})$$

3.2 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: (Walck 2007) vars avsn 30 behandlar ickecentral betafördelning men in te ger ngn bra formel för λ. 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 R^2. Dock görs inte detta i själva frågan utan här gör man istället en approximation för ett upper bound av $E[R^2]$. 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 λ 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 λ ju faktiskt beror på X, dvs på stickprovet. Detta innebär ju att vi i praktiken är tillbaka i vår sedan tidigare kända situation. Dock har vi här ett beroende på hela X, dvs en designmatris som vi kanske kan ta från vårt ursprungliga stora sample. Således har vi kanske trots allt inte ett beroende på den enskilda stickprovet!?

Det som sägs är:

> Consider the simple linear model:
>
> $$\boldsymbol{y} = X'\boldsymbol{\beta} + \epsilon$$
>
> where $\epsilon_i \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$ and $X \in \mathbb{R}^{n \times p}$, $p \geq 2$ and X contains a column of constants.
>
> $R^2 \sim \mathrm{B}(p-1, n-p, \lambda)$ where $\mathrm{B}(p-1, n-p, \lambda)$ is a non-central Beta distribution with non-centrality parameter λ with
>
> $$\lambda = \frac{||X'\beta - \mathrm{E}(X)'\beta 1_n||^2}{\sigma^2}$$

Dock misstänker jag här att X' kan vara sammanblandat med X och att det således borde vara $X\beta$ istf $X'\beta$. Har vi inte β borde vi istället kunna skatta $X\beta$ med hattmatrisen, dvs $X(X'X)^{-1}X'Y$ ... eller är det lika bra att börja om från början med data som helt följer modellen?

OBS! Denna (eller åtm liknande formel finns ockspå som (12) i (Helland 1987). Är väl därmed bättre att utgå från den som faktiskt publicerad referens.

OBS!!! X härrör fortfarande till just aktuellt sample ty λ växer med n :-(

Dock kan vi anta att centralitetsparametern = 0 då $E[X] = 0$ och då approximera med vanlig betafördelning. (Men gäller nog inte föfr lite mer komplicerade fördelningar, gissar att det inte räcker att bara standardisera resp parameter ... eller?)

3.3 Läsning av (Helland 1987)

Om tolkning av R^2 i regression. Argumenterar för att R^2 bara kan tolkas korrekt just då kovariaterna är random (då detta är bästa sättet att få en heltäckande bild av X). Skriver att det finns en del statistiker som avråder från att i huvud taget titta på R^2. Föreslår approximativt konfidensintervall för populationskoefficienten för korrelation. Observerar att adjusted R^2 faller inom intervallet men inte den vanliga. Utgår från modeller med intercept och ickekorrelerade fel. Bygger på matrisformler.

Ger också ρ^2 på formen:

$$\rho^2 = \frac{\sum_{i=i}^{p} \beta_i x_i}{var(y)}$$

Härleder också ickecentralitetsparametern

$$\lambda = \frac{\beta' X_0' X_0 \beta}{\sigma^2}$$

Detta sägs ge en konditional distribution av R^2 givet X_0 för random X. För en unconditional fördelning krävs dock fördelningsantagande för X. Detta beräknades redan av(R. A. Fisher 1928) men är för krångligt för att kunna användas. En approximation har dock givits av en Gurland 1968 (ej läst):

$$k = \frac{\rho^2}{1-\rho^2}, a = \frac{(n-k)k(k+2)+p}{(n-1)k+p}, v = \frac{(n-1)k+p}{a} \Rightarrow \frac{R^2}{1-R^2} \approx \frac{(n-1)k+p}{n-p-1} F_{v,n-p-1}$$

Denna approximation har sedan visat sig fungera bra. Utifrån denna approximation konstrueras sedan konfidensintervall. Den formel som då föreslås beror dock på ρ. Numeriska metoder används och konvergens uppnås ofta efter 3-4 itterationer. Resultatet härav blir väldigt bra överrensstämmande med tidigare teoretiskt uträknade motsvarande värden. Artikelns beskrivning av metoden är antagligen tillräcklig för att själv kunna återimplementera den men det känns ändå lite krångligt.

3.4 Läsning av (Rodgers and Nicewander 1988)

Innehåller en del historia. Artikeln skrevs för att fira att det var ca 100 år sedan regression infördes. Återkommer till de kändisar vi sett sedan tidigare men i organiserad form. Redan 1920 skrev f.ö. Pearson "Note on the history of correlation".

Skriver också att konceptet både är ett av de mest använda men också mest missbrukade inom statistik.

Artikeln presenterar sedan 13 olika tolkningar av r men bara under vissa förenklade förutsättningar, såsom endast bivariat fördelning:

1. den anliga algebraiska formeln
2. som standardiserad kovarians
3. som lutningen (slope) i regressionsmodell
4. geomertiskt medelvärde
5. roten av proportion of variability accounted for
6. mean cross product of standardized variables
7. vinkeln mellan två standardiseerade regressionslinjer
8. funktion av vinkeln mellan de två variabelvektorerna
9. ...

En del av de övriga känns lite väl teoretiska ...

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.

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.

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. 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.

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

Helland, Inge S. 1987. "On the Interpretation and Use of R2 in Regression Analysis." *Biometrics* 43 (1). [Wiley, International Biometric Society]: 61–69. doi:10.2307/2531949.

Kramer, K H. 1963. "Tables for Constructing Confidence Limits on the Multiple Correlation Coefficient." *Journal of the American Statistical Association* 58 (304). [American Statistical Association, Taylor & Francis, Ltd.]: 1082–5. doi:10.2307/2283334.

Montgomery, David B, and Donald G Morrison. 1973. "A Note on Adjusting R2." *The Journal of Finance* 28 (4): 1009–13.

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.

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

Rodgers, Joseph Lee, and W. Alan Nicewander. 1988. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1): 59–66. doi:10.2307/2685263.

Walck, Christian. 2007. "Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists." http://www.stat.rice.edu/{~}dobelman/textfiles/DistributionsHandbook.pdf.

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.

Wishart, Author J, T Kondo, and E M Elderton. 1931. "The Mean and Second Moment Coefficient of the Multiple Correlation Coefficient, in Samples from a Normal Population." *Biometrika* 22 (3/4): 353–76.