PROGRAMMA VOOR HET SCHATTEN VAN DE CORRELATIE-COEFFICIENT IN INKOMPLETE DATASETS

Ulbe Brouwer[1]), Pieter Vijn[2])

I INLEIDING

Regelmatig komt het voor dat onderzoekers de populatie correlatie ρ willen schatten op grond van inkomplete observaties voor de x en y variabele. (NB. we beperken ons hier tot het bivariate geval).

De meest gangbare aanpak is het berekenen van de steekproefcorrelatie gebaseerd op de komplete paren (x,y) (In SPSS is dit de bekende pairwise deletion optie). Het zal duidelijk zijn dat in dit geval, vaak moeizaam, verkregen informatie wordt weggegooid. Ook andere schattings-methoden zoals aangegeven door bv. Boas (1976) en Frane (1978) leveren doorgaans geen bevredigende resultaten op.

In het aan het Technisch Centrum FSW (in samenwerking met het Psychologisch Laboratorium UvA) ontwikkelde programma RMISDAT worden Bayesiaanse schatters voor ρ geïntroduceerd. Voor het berekenen van deze schatters wordt *alle* informatie, in de data aanwezig, gebruikt. Simulatie onderzoeken (Brouwer en Vijn, 1979A, 1979B) hebben aangetoond dat deze Bayesiaanse schatters in praktische situaties bruikbaar kunnen zijn.

II BAYESIAANSE SCHATTERS VOOR ρ

Het voert te ver de gehanteerde rekenformules hier expliciet te vermelden (zie hiervoor Brouwer en Vijn, 1979). We willen in dit onderdeel slechts globaal aangeven welke schatters er in RMISDAT worden berekend. Uitgegaan wordt van een bivariate normale verdeling met ontbrekende x en/of y scores.

[1]) Technisch Centrum FSW, Universiteit van Amsterdam

[2]) Psychologisch Laboratorium, Universiteit van Amsterdam

Daartoe vatten we (x,y) samen in de rijvector $z' = (x,y)$.
Er wordt nu aangenomen dat

$$z|\mu_z, \Sigma = N(\mu_z, \Sigma)$$

met $\mu_z' = (\mu_x, \mu_y)$

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \rho\,\sigma_x\,\sigma_y \\ \rho\sigma_y\sigma_x & \sigma_y^2 \end{pmatrix}$$

Met de regel van Bayes volgt nu dat de aposteriori dichtheid van $\theta' = (\mu_x, \mu_y, \sigma_x^2, \sigma_y^2, \rho)$ gelijk is aan

$$p(\theta|z) \sim l(z|\theta)\, p(\theta),$$

waarin $l(z|\theta)$ de 'likelihood' van z gegeven θ en $p(\theta)$ de apriori dichtheid van θ is.
We nemen bovendien aan dat

(i) $p(\theta) = p(\mu_x)\; p(\mu_y)\; p(\sigma_x^2)\;\; p(\sigma_y^2)\;\; p(\rho)$

(ii) μ_x, μ_y, $\ln \sigma_x^2$ en $\ln \sigma_y^2$ uniform verdeeld zijn

(iii) $p(\rho) \sim (\rho-\rho_l)^{a-1}(\rho_u - \rho)^{b-1}$

Hierin zijn ρ_l en ρ_u de theoretisch mogelijke onder- resp. bovengrens voor ρ. De parameters a en b geven de sterkte van de prior informatie weer. Voor bv. a = b = 1 heeft men een uniforme prior voor ρ over het interval $[\rho_l, \rho_u]$. Hoe groter a en/of b, hoe meer voorkennis men heeft ten aanzien van ρ. Een milde informatieve symmetrische prior voor ρ rond $\rho = 0.50$ zou b.v. als volgt gespecificeerd kunnen worden: $\rho_l = .00$, $\rho_u = 1.00$, a = b = 3. (D.w.z. in dit geval gebruiken we de informatieve beta-prior: $p(\rho) \sim \rho^2(1-\rho)^2$.

De volgende twee bayesiaanse schatters voor ρ worden berekend:

$$\max \quad p(\rho \mid \kappa = \hat{\kappa},\ \mu_x = \bar{x},\ \mu_y = \bar{y},\ z) \qquad (1)$$

en

$$\max \quad p(\rho \mid \kappa = \hat{\kappa},\ \mu_y = \bar{y}, z) \qquad (2)$$

waarin

$\hat{\kappa} = \frac{s_y^2}{s_x^2}$ = quotient van steekproef varianties van y resp. x

$\bar{x}$ = gem. van x scores

$\bar{y}$ = gem. van y scores

De eerder genoemde simulatie onderzoeken hebben aangetoond dat numeriek uit integreren van k uit (2) geen noemenswaardige verbeteringen ten aanzien van de schatting van ρ oplevert in het gebied $|\rho| \leq .70$. Deze laatste modale schatter is nu dan ook niet in het huidige programma opgenomen. (NB. dit mede om technische redenen: nummerieke integratie is hier veelal zeer tijdrovend).

III HET PROGRAMMA RMISDAT

Het programma RMISDAT kan maximaal slechts 100 paren (x,y) verwerken. Grotere datasets zijn niet zo interessant meer omdat in die gevallen de steekproef correlatie, over de komplete paren berekend, al een voldoende nauwkeurige schatter zal zijn. Een voorbeeld:

Bij N = 100 met 20% ontbrekende y-waarden is de gemiddelde afwijking tussen de produkt moment correlatie over alle 100 cases (r_c) en de product moment correlatie (r_{ic}) over de 80 overgebleven cases voor 100 gesimuleerde steekproeven (uit een populatie met ρ = .50) slechts .03 met een spreiding van 0.015. Bij kleinere steekproeven kan echter het verschil tussen r_c en r_{ic} wel aanzienlijk zijn.

Vooral voor steekproeven ter grootte van ruwweg N = 40, en zeg 20 tot 50% ontbrekende paren levert RMISDAT een substantieel betere schatter voor ρ op. Dit geldt in nog sterkere mate voor de zogenaamde "restriction of range" situatie (case I). Dan ontbreken y scores indien de bijbehorende x score groter (of kleiner) is dan een bepaalde criterium waarde x_0. In de praktijk van het psychologisch onderzoek komt deze situatie regelmatig voor.

Stel dat 50 personen een vorderingen toets maken (x score) en dat de 50% personen met de laagste x scores tevens een aanvullende toets moet afleggen (y score).
Gevraagd: de correlatie tussen x en y.
In dit soort situaties kunnen bayesiaanse schatters tot aanzienlijke verbeteringen in de schatting voor ρ leiden.
Het programma berekent tevens de Pearson schatter R voor ρ (in "restriction of range, case I", Gullickson, 1974).

$$R = \frac{\frac{S_x}{s_x} r}{\sqrt{1+r^2\left[\frac{S_x^2}{s_x^2}-1\right]}} \qquad (4)$$

met S_x: steekproef standaard deviatie voor x
s_x: steekproef standaard deviatie voor x met $x < x_0$
r : steekproef correlatie over de komplete paren

Deze schatter R wordt in case I situaties als regel toegepast. Simulatie onderzoek toonde aan (Brouwer en Vijn, 1978) dat R echter niet uniform beter is dan de steekproefcorrelatie r berekend uit de komplete paren. Als vuistregel kan gelden:

verwachting voor ρ	te kiezen schatter
$\lvert\rho\rvert \leq .50$	r
$\lvert\rho\rvert > .50$	R

Eerder genoemd onderzoek heeft uitgewezen dat de bayesiaanse schatter (2) voor relatief lage tot middelgrote ρ ($\rho < .70$) een betere schatter voor ρ is dan R.
De vele simulaties die met betrekking tot de bayesiaanse schatters (1) en (2) zijn uitgevoerd, hebben uitgewezen dat, gebruikmakend van een uniforme prior op $[-1, +1]$:

- (2) altijd beter is dan de klassieke produkt moment correlatie over de komplete paren (x,y).
- (1) bij relatief lage ρ een betere schatter is dan (2)
- voor relatief middelhoge tot hoge ρ (2) beter is dan (1).

Afhankelijk van de verwachting van de onderzoeker met betrekking tot de waarde van ρ zal hij een keuze moeten maken uit R, (1) of (2).
Als vuistregel zou kunnen gelden (onder de aanname van een uniforme prior op $[-1,+1]$):

verwachting voor ρ	te kiezen schatter
$\lvert\rho\rvert \leq .30$	(1)
$.30 < \lvert\rho\rvert < .70$	(2)
$\lvert\rho\rvert \geq .70$	R

Wat betreft de schatter voor ρ die ontstaat door κ nummeriek uit te integreren, kan gesteld worden dat deze schatter beter is dan R voor $\lvert\rho\rvert \leq .85$.
Het blijft jammer dat de bayesiaanse schatters niet uniform beter zijn dan de 'klassieke' schatters. Maar ook al verwerpt een onderzoeker de bayesiaanse schatters, dan zal hij toch op grond van vooraf informatie over ρ een keuze moeten maken tussen de klassieke schatters. Indien hij verwacht dat ρ niet zo groot zal zijn en tot r besluit, dan kan hij beter de bayesiaanse schatter (1) kiezen. Deze schatter is zonder vooraf informatie over ρ beter dan r. De bayesiaanse schatters worden natuurlijk aantrekkelijker indien op grond van voorgaand onderzoek de a priori informatie over ρ substantieel is.

RMISDAT geeft voor de 2 modale bayesiaanse schatter voor ρ de kansdichtheid in tabellarische en in een print plot weer. Daarnaast wordt tevens de kumulatieve kans verdeling geprint (en geplot). Hierdoor is het mogelijk geworden concrete kansuitspraken te doen t.a.v. ρ. Het is b.v. mogelijk de kans af te lezen dat $\rho < 0$ of $\rho > .4$. Het ligt in de bedoeling in latere versies van RMISDAT de HDR (highest density regions) intervallen te laten berekenen. Ook zal het dan mogelijk zijn uitspraken te doen over de kans dat ρ in een bepaald gebied ligt. In komende versies van RMISDAT zal eveneens κ numeriek uit geïntegreerd worden.

We geven tot slot nog een klein getallen voorbeeld:

Voor 40 observaties is de correlatie tussen x en y gelijk aan .26. De bayesiaanse schatters (1) en (2) zijn ook bij benadering gelijk aan .26 (bij een uniforme prior voor ρ over $[-1+1]$). De 10 y-waarden met de hoogste bijbehorende x scores worden nu als ontbrekend opgegeven (m.a.w. 25% ontbrekende waarden in restriction of range, case I).

De steekproef correlatie (over de 30 komplete paren) is nu .18 R = .20, bayesschatter (1) = .21 en (2) = .22 (bij een uniforme prior voor ρ). Een lichte informatieve prior over -.25 tot +.75 met a = b = 3 geeft een lichte verbetering in de schatter (1). Ze wordt nu nl. .23.

REFERENTIES

Boas, J.

A note on the estimation of the covariance between two random variables using extra information in the separate variables.

Statistica Neerlandica, 21, 1969.

Brouwer, U. en Vijn, P.
De Empirische power van 2 procedures in "Restriction of Range" (case I).
Tijdschrift voor Onderwijs Research, 78, 3, 1978.

Brouwer, U. en Vijn, P.
Ontbrekende waarnemingen en bayesiaanse schatters voor de correlatie coëfficiënt.
Intern rapport Technisch Centrum FSW, febr. 1979.

Brouwer, U. en Vijn, P.
Bayesiaanse schatters voor de correlatie coëfficiënt in "restriction of range" (geval 1).
Aangeboden aan: Tijdschrift voor Onderwijs Research, 1979.

Frane, J.W.
Missing data and BMDP.
Technisch rapport, 1978.

Gullickson, A. en Hopkins, K.
Interval estimation of correlation coefficients corrected for restriction of range.
Educational and Psychological Measurement, 36, 1976.

Programmabeschrijving RMISDAT,
Technisch Centrum FSW Universiteit van Amsterdam,
lp. nr. 66, 1979.