Interne notater

STATISTISK SENTRALBYRÅ

80/32

22. oktober 1980

LINEÆR REGRESJON UTEN KONSTANTLEDD

LITT ELEMENTÆR INFORMASJON

Av

Herdis Thorén Amundsen

INNHOLD

INNLEDNING

I enkelte prosjekter er det aktuelt å finne minste kvadraters regresjonskoeffisientene i en lineær regresjon der konstantleddet er satt lik null, altså av formen

$$y = b_1x_1 + b_2x_2 + \ldots + b_kx_k + \text{restledd}. \qquad (1)$$

De nødvendige formler vil en kunne utlede av de generelle løsninger i videregående lærebøker.

Siden vi ikke har funnet noen lett tilgjengelig elementær beskrivelse, er en del punkter belyst i dette notatet.

S.E. Brun har gjort oppmerksom på problemstillingene og J. Ouren har bistått med prøvekjøring av TROLL-programmet.

1. Formelapparatet

Beregning av regresjonskoeffisientene, $b_1, b_2, \ldots, b_k$, samt standardavvik osv., svarer til å beregne de tilsvarende størrelsene i en vanlig regresjon, men en setter inn null istedenfor gjennomsnittene $\bar{y}, \bar{x}_1, \ldots, \bar{x}_k$ i alle varianser og kovarianser som inngår i formlene. (Dette gjelder selv om gjennomsnittene i observasjonsmaterialet ikke er lik null.) Dessuten blir "antall frihetsgrader" (df) i t- og F-tester noe endret, se nedenfor.

For et materiale med observasjonssett $(y_i, x_{1i}, \ldots, x_{ki})$ for $i = 1, 2, \ldots, n$, bruker vi altså

$$\sum_{i=1}^{n} y_i^2 \quad \text{istedenfor} \quad \sum_{i=1}^{n} (y_i-\bar{y})^2,$$

$$\sum_{i=1}^{n} y_i x_{1i} \quad \text{istedenfor} \quad \sum_{i=1}^{n} (y_i-\bar{y})(x_{1i}-\bar{x}_1),$$

$$\sum_{i=1}^{n} x_{1i} x_{ji} \quad \text{istedenfor} \quad \sum_{i=1}^{n} (x_{1i}-\bar{x}_1)(x_{ji}-\bar{x}_j) \text{ for } j = 1, 2, \ldots, k.$$

Eksempelvis har vi for en regresjon mhp én variabel,

$$y = bx + \text{restledd}, \qquad (2)$$

med observasjonsparene (x_i, y_i) for $i = 1, 2, \ldots, n$, at

$$b = \frac{\sum_i y_i x_i}{\sum_i x_i^2}$$

istedenfor $\frac{\sum_i (y_i-\bar{y})(x_i-\bar{x})}{\sum_i (x_i-\bar{x})^2}$ i vanlig regresjon med konstantledd.

Restvariansen blir estimert ved

$$\frac{1}{n-1} \sum_i (y_i-bx_i)^2,$$

altså divisjon med (n-1) istedenfor (n-2).

Variansen for b blir estimert ved

$$\frac{\frac{1}{n-1} \sum_i (y_i-bx_i)^2}{\sum_i x_i^2}, \quad \text{osv.}$$

Gjør vi de vanlige forutsetningene om uavhengighet og normal fordeling med samme teoretiske varians for alle y-variable, så kan vi teste om den teoretiske regresjonskoeffisienten er null ved hjelp av

$$t = b\sqrt{\frac{(n-1)\sum_i x_i^2}{\sum_i (y_i-bx_i)^2}},$$

eller ved

$$F = t^2.$$

Her gjelder nå t-fordelingen med (n-1)df, respektive F-fordelingen med 1 og (n-1)df. (I vanlig regresjon har vi (n-2)df.)

Den vanlige korrelasjonskoeffisienten,

$$r = \frac{\sum_i (y_i-\bar{y})(x_i-\bar{x})}{(\sum_i (x_i-\bar{x})^2 \cdot \sum_i (y_i-\bar{y})^2)^{0.5}},$$

kan ikke brukes som "mål" for hvordan regresjonen (2) passer til data (unntatt hvis $\bar{x}$ og $\bar{y}$ faktisk er null for våre observasjoner).

For en regresjon med k høyresidevariable, som (1), og n > k observasjonssett, får vi tilsvarende

$$\text{restvarians} = \frac{1}{n-k} \sum_i (y_i - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2 = \frac{SSR}{n-k} , \qquad (3)$$

der SSR er restkvadratsummen, jfr. punkt 2 nedenfor om TROLL-programmet.

For testing av en enkelt regresjonskoeffisient må vi bruke t-fordelingen med (n-k)df (istedenfor (n-k-1)df når vi har med et konstantledd i tillegg til de k regresjonskoeffisientene).

Den vanlige F-observator for testing av alle de k regresjonskoeffisientene under ett, blir nå

$$F = \frac{\sum_i y_i^2 - \sum_i (y_i - b_1 x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2}{\sum_i (y_i - b_i x_{1i} - b_2 x_{2i} - \ldots - b_k x_{ki})^2} \cdot \frac{n-k}{k} = \frac{\sum_i y_i^2 - SSR}{SSR} \cdot \frac{n-k}{k} , \qquad (4)$$

med k og (n-k)df.

Andre testobservatorer må modifiseres tilsvarende.

2. Bruk av TROLL for regresjon uten konstantledd

Enkelte standard regresjonsprogrammer har vist seg mindre egnet for regresjon uten konstantledd.

Standard regresjonsprogrammet i TROLL ser ut til å være bra, bortsett fra et par punkter. En får

korrekte beregninger for

- regresjonskoeffisientene, b_j (B, C, osv.),
- standardfeilen (ST ER) på disse,
- t_j-verdiene (T-STAT),
- restkvadratsum (SSR),
- reststandardavvik (SER), altså kvadratroten av (3) ovenfor,
- kovariansmatrisen for regresjonskoeffisientene,
- gjennomsnittene for de variable (LHS MEAN = $\bar{y}$ og MEAN, som er $\bar{x}_1$, $\bar{x}_2$ osv.).

En får

meningsløs beregning for

. "determinasjonskoeffisienten" (RSQ), svarende til R^2 i vanlig regresjon,

. F-verdi (kalt "F" nedenfor),

. PARTIAL og BETA koeffisientene som ikke har den vanlige betydning.

Beregningen av RSQ og "F" svarer til formlene

$$RSQ = \frac{\Sigma(y_i-\bar{y})^2 - SSR}{\Sigma(y_i-\bar{y})^2}$$

og

$$"F" = \frac{RSQ}{1-RSQ} \cdot \frac{n-k}{k-1} \quad . \qquad (6)$$

Sammenlikning med (3) og (4) ovenfor viser hvor feilene ligger. Vi kan si at denne RSQ sammenlikner kvadratsummen av avvikene rundt regresjonslinjen med kvadratsummen av avvikene rundt $\bar{y}$. Dette har ikke mening når vi a priori setter konstantleddet lik null. (Hvis vi er usikre på om vi bør utelate konstantleddet, kan vi jo ta en titt på RSQ og "F", men selve testen må vi foreta på annen måte.) Denne "F" er ikke F-fordelt. Både RSQ og "F" kan f.eks. bli negative, i dette tilfelle gir programmet advarselen: WARNING 2040. VARIANCE OF RESIDUALS IS GREATER THAN VARIANCE OF OBSERVED LEFTHAND SIDE. Og dette er jo et tegn på at regresjon uten konstantledd ikke passer særlig bra. (I regresjon med én x-variabel opplyser programmet at "F" har 0 og (n-1)df, men dividerer dog ikke med 0.)

Hvis vi ønsker en F-test av samtlige k koeffisienter under ett, kan vi regne ut F ifølge (4), enten direkte, da må vi altså sørge for å få ut Σy_i^2 i tillegg til de øvrige data, eller via (6). Vi finner

$$F = \frac{1}{k} \left("F"(k-1) + \frac{n\bar{y}^2(n-k)}{SSR} \right).$$

For k = 1 erstattes (k-1) med 1 i denne formelen. For k = 2 er det kanskje like lett å beregne

$$F = \frac{t_1^2 + t_2^2}{1-r_{12}^2} \cdot \frac{n-2}{2n},$$

der t_1 og t_2 er de to T-STAT-verdiene og

$$r_{12}^2 = \frac{(\text{kovar}(b_1,b_2))^2}{\text{var } b_1 \cdot \text{var } b_2}$$

er beregnet ut fra "COVARIANCE MATRIX" for b_1(B) og b_2(C) i utskriften.

Vi må også huske at for regresjon uten konstantledd blir summen av restleddene ($y^{observert} - y^{beregnet}$) ikke lik null. I TROLL angir SR hva summen blir.

3. Sammenlikning av regresjoner uten konstantledd for flere observasjonsmaterialer

Anta at vi har regresjoner av formen (1) fra m ulike utvalg (m fylker, f.eks.), der utvalg nr. q har n_q observasjonssett, q = 1, 2, ..., m. Vi ønsker å undersøke om disse er slik at de kunne være generert ut fra de samme teoretiske regresjoner, dvs. at vi kan beregne en felles regresjon av formen (1) for alle de m datasettene under ett.

For å undersøke dette, beregner vi den felles regresjonen og sammenlikner restkvadratsummen, SSR, med summen av restkvadratsummene for de m regresjonene. Under de vanlige forutsetningene om uavhengighet, normal fordeling og samme teoretiske restvarians i hele materialet, vil

$$F = \frac{SSR - \Sigma SSR_q}{\sum_{q=1}^{m} SSR_q} \cdot \frac{n-km}{k(m-1)} \qquad (7)$$

være F-fordelt med k(m-1) og (n-km)df når nullhypotesen om felles teoretiske koeffisienter er riktig. Her er SSR_q restkvadratsummen i regresjon nr. q, q = 1, 2, ..., m og

$$n = \sum_{q=1}^{m} n_q .$$

Vi forkaster altså nullhypotesen når den observerte F er større enn f.eks. 95-prosent fraktilen i nevnte fordeling.

Draper & Smith (1966) har en test for sammenlikning av m enkelte regresjonskoeffisienter i "Exercise E", s. 39-40. Vår SSR blir enklere å beregne fordi vi ikke har ulike konstantledd å ta hensyn til.

Hvis vi må forkaste nullhypotesen, vil vi kanskje forsøke å samle utvalgene (fylkene) i grupper der nullhypotesen kan gjelde innen hver gruppe. Vi kan ha to ulike situasjoner her.

(i) Vi foretar grupperingen ut fra a priori overlegninger, f.eks. om bosetting, næringsstruktur, geografi eller hva vi nå mener kan ha betydning. Da kan vi beregne en F_g for hver gruppe ut fra (7). Vi setter n_g istedenfor n, antall utvalg (fylker), m_g, i gruppen istedenfor m, samt restkvadratsummen SSR_g beregnet ut fra den felles regresjon for denne gruppen. Vi har da en F_g med $k(m_g-1)$ og (n_g-km_g)df for hver gruppe for testing av gruppens "homogenitet".

Vi kan få en noe sterkere test hvis vi er noenlunde sikre på at de teoretiske variansene for de ulike gruppene ikke er forskjellige. Da kan vi bruke

$$F_g^1 = \frac{SSR_g - \sum_{q=1}^{m_g} SSR_{qg}}{\sum_{q=1}^{m} SSR_q} \cdot \frac{n-km}{k(m_g-1)} \qquad (8)$$

istedenfor F_g. F_g^1 har $k(m_g-1)$ og $(n-km)$df.

I begge tilfelle bør vi bruke lavere forkastningssannsynlighet for hver enkelt test enn det nivået vi ønsker totalt. Med G grupper og nivå 5 prosent, kan vi bruke $\frac{5}{G}$ prosent for den enkelte test. Har vi G = 2, bruker vi altså 97,5 prosent fraktilene istedenfor 95 prosent, og med G = 5 kan vi bruke 99 prosent fraktilene i F-fordelingen.

En felles test for alle gruppene får vi ved å bruke

$$F_G = \frac{\Sigma SSR_g - \Sigma SSR_q}{\Sigma SSR_q} \cdot \frac{n-km}{k(m-G)} \cdot \qquad (9)$$

Vi må forkaste hypotesen om "homogenitet" innen alle gruppene hvis vi finner at $F_G \geq F_{0,95}$ (f.eks.), der $F_{0,95}$ er 95 prosentfraktilen i F-fordelingen med k(m-G) og (n-km)df. Hvis vi får forkasting, beregner vi F_g^1 ifølge (8) for de enkelte gruppene, og plukker ut de som har $F_g^1 \geq F_{0,95}$. (Vi sammenlikner her ikke med den fraktilen som er nevnt under (8) ovenfor, og som vil være noe høyere.) Det må finnes minst én slik gruppe, for hvis $F_g^1 < F_{0,95}$ i alle gruppene, så vil også $F_G < F_{0,95}$.

Den siste testen, med nivå 5 prosent, er antagelig sterkere enn testsettet under (8) med nivå $\frac{5}{G}$ prosent, i alle fall når G ikke er for stor.

(ii) Vi foretar grupperingen, ikke ut fra a priori viten, men ved å "se på" de m enkeltregresjonene og samle dem i grupper der regresjonskoeffisientene for tilsvarende x_j-er ser mest mulig "like" ut. En måte å gå fram på er da å regne ut verdien av F_G i formel (9). Vi skal her kalle den $F_G^{(ii)}$, fordi den ikke vil ha samme fordeling som under (i). Dette skyldes fremgangsmåten ved grupperingen. For et gitt tall C vil i alminnelighet sannsynligheten for at $F_G^{(ii)} \leqq C$ være større enn sannsynligheten for at $F_G \leqq C$ (vi "lager" lavere $F_G^{(ii)}$-verdier enn vi ville få ved trekking fra en F_G-fordeling, med mindre vi er uhyre kløfnete til å gruppere sammen utvalgene).

Hvis vi nå forkaster nullhypotesen om "homogene grupper" når vi finner

$$F_G^{(ii)} \geqq F_{0,95}, \tag{10}$$

så må vi regne med å ha et sannsynlighetsnivå lavere enn 5 prosent. Vi har også lavere teststyrke enn under (i), dvs. det er vanskeligere å få forkastet nullhypotesen hvis den er gal.

Hvis (10) er oppfylt, må vi nokså trygt kunne slutte at en eller flere av gruppene passer dårlig, og plukke ut de som har $F_g^1 \geqq F_{0,95}$, som omtalt under (9) ovenfor.

Hvis vi ikke får forkasting, hvis altså

$$F_G^{(ii)} < F_{0,95}, \tag{11}$$

så kan det likevel tenkes at vi har en eller et par $F_g^1 \geqq F_{0,95}$. Disse bør vi antagelig se nærmere på. Forkasting av nullhypotesen også for disse, vil gjøre nivå og teststyrke noe høyere enn om vi bare forkaster under (10).

(I prinsippet kan vi tenke oss en minimeringsprosess ved hjelp av EDB for å finne en "minimal" $F_G^{(ii)}$. Da ville vi også - i prinsippet - kunne finne fordelingen av $F_G^{(ii)}$ og få kontroll over testnivået. I praksis vil dette støte på atskillige vanskeligheter, både matematiske, spørsmålet om størrelsen på G, om de fremkomne gruppene har noen mening ut fra vårt problem, etc.)

For å bøte på problemet med lav teststyrke, bør vi kanskje velge et høyere nominelt nivå enn 5 prosent, f.eks. 10 prosent, og erstatte $F_{0,95}$ ovenfor med $F_{0,90}$.

(iii) Hvis det er ett, eller et par, utvalg i materialet som skiller seg markert ut, slik at de(t) ikke kan parres sammen med noen av de øvrige, så må vel dette (disse) tas ut som særskilt(e) "gruppe(r)" med $m_g = 1$, og testene i (i), resp. (ii), utføres på resten av materialet.

4. Spørsmål om ulike teoretiske varianser for de m utvalgene

Testmetodene foran forutsetter at vi kan regne med at alle de m resp. m_g observasjonssettene har samme teoretiske varians. Anta at vi på forhånd har grunn til å tro at ett, f.eks. nr. m, eller flere bestemte som vi kan gi numrene r+1, r+2, ..., m, av settene har større varians enn de øvrige. Da har vi mulighet for å teste dette ved hjelp av restkvadratsummene i enkeltregresjonene. Vi kan se om

$$F_m = \frac{SSR_m}{\frac{m-1}{\sum_{q=1} SSR_q}} \cdot \frac{n - n_m - (m-1)k}{n_m - k}$$

er større enn (f.eks.) 95 prosentfraktilen i F-fordelingen med (n_m-k) og $(n-n_m-(m-1)k)$df, når det gjelder ett utvalg.

I annet tilfelle sammenlikner vi

$$F_{(r)} = \frac{\frac{\sum_{r+1}^{m} SSR_q}{m-r}}{\frac{\sum_{1}^{r} SSR_q}{r}} \cdot \frac{\sum_{1}^{r}(n_q-k)}{\sum_{r+1}^{m}(n_q-k)}$$

med 95 prosentfraktilen med $\sum_{r+1}^{m}(n_r-k)$ og $\sum_{1}^{r}(n_q-k)$df. Vi forutsetter da at variansene innen hver av de to gruppene er like.

Testene vil imidlertid gjelde approksimativt også om alle de teoretiske variansene innen hver av de to grupper ikke er like store.

Hvis vi er meget usikre m.h.t. likheten av variansene, har vi en mulighet for å teste om alle m er like forutsatt at alle $n_q \geq 2k+2$. En testmetode som da kan tilpasses vårt problem er beskrevet av Scheffé (1959) ch. 3.8.

Imidlertid er det slik at moderat ulikhet mellom de teoretiske variansene ikke behøver å ødelegge de F-testene vi har omtalt foran. Hvis det er like mange observasjoner i utvalgene, altså $n_1 = n_2 = \ldots = n_m = \frac{n}{m}$, og ikke altfor store ulikheter mellom variansene, så blir ikke testens nivå og styrke særlig berørt (Scheffé, ch. 10,4). Det sanne nivået kan bli noe høyere enn det nominelle, f.eks. 5 prosent, slik at vi bør velge et noe lavere nominelt nivå, f.eks. 2,5 prosent, for å være på "den sikre siden".

Litteratur

Draper, N.R. and Smith, H. (1966). Applied Regression Analysis. Wiley.

Scheffé, H. (1959). The Analysis of Variance. Wiley.

Sverdrup, E. (1967). Basic Concepts of Statistical Inference II. North-Holland.

Rao, C.R. (1968). Linear Statistical Inference and Its Applications. Wiley.