Center for Statistik
Handelshøjskolen i København

MPAS

Tue Tjur

November 2006

Multipel regression med laggede responser som forklarende variable

Ved en *tidsrække* forstås i almindelighed et datasæt, der indeholder en observation pr. tidsenhed (f.eks. pr. dag, pr. time eller pr. uge). Ordet "observation" skal her læses i udvidet betydning, der kan sagtens være tale om, at man til hvert tidspunkt observerer flere forskellige ting, hvoraf en (eller flere) måske vil optræde som stokastiske variable i de modeller man benytter, andre som forklarende variable (eller faktorer) med givne værdier.

Statistisk set er det væsentlige træk ved tidsrækker, som adskiller dem fra datasæt af mere traditionel type, at man kun undtagelsesvis kan analysere dem ordentligt ved hjælp af modeller, som antager at observationerne er stokastisk uafhængige.

Antag for eksempel at (y_t), $t = 1, \ldots, T$, er observationer dag for dag over en eller anden periode af temperaturen kl. 12 middag i København. Vi interesserer os for statistiske metoder, der kan bruges til forudsigelse af denne temperatur et døgn frem. Til brug for dette har vi diverse andre variable (x_t), (z_t), ... som indeholder oplysninger af meteorologisk karakter eller oplysninger som på anden måde kan indeholde information om temperaturen. Det kunne være skydække, vindstyrke og nedbør (i København eller andre steder), datoens gennemsnitstemperatur over de sidste ti år, eller variable der indgår i en årstidstrend (f.eks. rene svingninger med periode et år).

Hvis man i en sådan situtation forsøger sig med en traditionel regressionsmodel af typen

$$y_t = \gamma + \beta x_t + \delta z_t + \cdots + \sigma u_t \,,$$

hvor u_t'erne er uafhængige N(0,1), har man allerede begået to fejl:

(1) Hvis modellen skal bruges til forudsigelse, kan det jo ikke nytte noget, at skydække eller vindhastighed på *samme* dag optræder som forklarende variabel, for så kan vi slet ikke regne forudsigelsen ud dagen før. I en model, som skal bruges til forudsigelse et døgn frem, må der kun indgå forklarende variable på højre side, som stammer fra dagen før (eller endnu tidligere). Altså, hvis f.eks. x_t først kan registreres på dag t,

$$y_t = \gamma + \beta x_{t-1} + \delta z_t + \cdots + \sigma u_t \,.$$

Tilsvarende, hvis vi ville forudsige temperaturen en hel uge frem, måtte vi naturligvis nøjes med at bruge forklarende variable, hvis værdier foreligger en uge før.

Bemærk, at nogle af de variable vi har omtalt godt kan indgå med deres samtidige værdier. Det gælder for eksempel datoens gennemsnitstemperatur over en tiårsperiode, og variable som indgår i en periodisk tidstrend. I vores notation underforstås altså, at x_t er en variabel som først kan registreres samtidig med responsen, medens, z_t kendes på forhånd (eller i hvert fald dagen før).

Hvis vi kan få sådan en model til at passe, ligger det lige for at bruge den estimerede middelværdi (altså højre side uden fejlled, med estimatorer indsat) som den bedste forudsigelse af selve observationen.

(2) Men det kan vi nok ikke, for det er slet ikke realistisk at forvente, at observationerne i denne model er uafhængige. Tværtimod, al erfaring taler for, at temperaturen en given dag vil være stærkt positivt korreleret med temperaturen dagen før, også selvom de andre meteorologiske variable muligvis vil korrigere for en del af denne afhængighed. Hvis man tegner residualerne som funktion af tiden vil man typisk se, at de zig–zagger op og ned langsommere end de burde hvis de var uafhængige. Dette fænomen kaldes *autokorrelation* (altså "selvkorrelation"; temperaturen er "korreleret med sig selv" i den forstand at temperaturer målt på tidspunkter tæt ved hinanden er (positivt) korrelerede).

Men netop denne korrelation betyder jo, at temperaturen i går muligvis er den allervigtigste oplysning vi har til forudsigelse af temperaturen i dag. Man fristes derfor til at inddrage temperaturen i går som forklarende variabel, og skrive

$$y_t = \gamma + \alpha y_{t-1} + \beta x_{t-1} + \delta z_t + \cdots + \sigma u_t$$

Man kunne endda tage temperaturen to dage før, tre dage før osv. med på højre side. Men nu er vi ude i noget, som i hvert fald ikke er en sædvanlig multipel regressionsmodel. Umiddelbart ser det meget, meget forbudt ud. En modelspecifikation af denne art, hvor de *laggede* — dvs. "forsinkede" — responser optræder som forklarende variable på højre side giver kun mening, hvis den fortolkes rigtigt:

Ideen er, at den *betingede* fordeling af responsen y_t, givet de tidligere responser $y_{t-1}, y_{t-2}, \ldots$, antages at være den normale fordeling med middelværdi $\gamma + \alpha y_{t-1} + \beta x_{t-1} + \delta z_t + \ldots$ og varians σ^2.

Med denne fortolkning viser det sig til gengæld, at man får en model, der i praksis kan håndteres og fortolkes næsten som en multipel regressionsmodel, idet man stort set kan ignorere det faktum at en eller flere af de forklarende variable på højre side er laggede versioner af selve responsvektoren.

Den vigtigste begrundelse for dette er, at likelihoodfunktionen ser ud præcis som i en multipel regression. For at indse dette må vi først bemærke, at det er nødvendigt at opfatte den første observation y_1 som givet (ikke–stokastisk). Vi har jo ikke en observation af y_0, som gør

det muligt at lade responsen y_1 indgå i datasættet sammen med alle tilhørende værdier af de forklarende variable. Men y_1 kan bruges som den første værdi af den forklarende variabel, som i ligningen hedder y_{t-1}, dvs. til at forklare y_2, som så bliver den første egentlige respons. Tilsvarende, hvis vi på højre side vil medtage den to gange laggede respons y_{t-2}, så er vi nødt til at opfatte y_1 og y_2 som givne, osv. Dette er sådan set ikke anderledes end i modellen, hvor skydække eller vindhastighed fra dagen før eller to dage før optræder på højre side. Helt generelt er man jo i enhver regressionsmodel tvunget til at se bort fra observationer, for hvilke en eller flere forklarende variable er uoplyste.

Efter denne bemærkning kan vi opskrive likelihoodfunktionen ved brug af kædereglen for betingede fordelinger (bedst kendt for sandsynlighedsfunktioner i det diskrete tilfælde, men den gælder uændret for tætheder i det kontinuerte tilfælde). I ord: Tætheden i det observerede punkt (som er likelihoodfunktionens værdi som funktion af de ukendte parametre) kan opskrives som

den marginale tæthed for den første observation, ganget med

den betingede tæthed for den anden givet den første, ganget med

den betingede tæthed for den tredje givet de to første, ganget med

osv. osv., ganget med

den betingede tæthed for den sidste givet alle de foregående.

I vores model kommer det til at se sådan ud:

$$L(\gamma, \alpha, \ldots, \sigma^2) = \prod_t \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_t - (\gamma + \alpha y_{t-1} + \ldots))^2}{2\sigma^2} \right)$$

hvor produktet skal tages over de t for hvilke alle variable, inklusive de laggede, er veldefinerede (f.eks. for $t = 2, \ldots, T$ hvis kun de én gang laggede observationer er med på højre side).

Rent algebraisk er dette en ganske almindelig multipel–regressions–likelihood, som kan maksimeres på sædvanlig måde ved mindste kvadraters metode. Så når det gælder beregningen af ML–estimatorer og kvotienttesstørrelser kan man bruge et standardprogram til håndtering af lineære normalfordelingsmodeller. De laggede versioner af responsvektoren, der optræder som forklarende variable på højre side, skal behandles præcis som om de var egentlige "eksterne" forklarende variable. Bortset fra, at de altså skal konstrueres først og tildeles passende navne (f.eks. `Y1, Y2` osv., hvis responsen hedder `Y`).

Heraf følger ikke at de fordelinger, som man plejer at benytte i forbindelse med multipel regression (T–fordelinger i forbindelse med sikkerhedsgrænser, T– og F–fordelinger i forbindelse med kvotienttest for modelreduktion) kan bruges på samme måde som man er vant til. Men man kan vise, at disse fordelinger under passende realistiske forudsætninger

er gode approksimationer til de faktiske fordelinger af teststørrelserne. Og man kan i øvrigt også give andre gode argumenter for at bruge præcis de samme metoder, som anvendes i multipel regression. Specielt kan testet for $\alpha = 0$ — som jo i dette tilfælde (forudsat at kun et enkelt lag inddrages) kan fortolkes som et test for hypotesen om *ingen autokorrelation* — foretages som et T–test på sædvanlig måde. Hvis denne hypotese godkendes, er vi tilbage i en sædvanlig lineær model for uafhængige observationer.

Fra et praktisk synspunkt er konklusionen således, at man i modeller til forudsigelse i tidsrækkedata har lov til at tage laggede versioner af responsvektoren med som forklarende variable på højre side. Vi har her forklaret det i tilfælde af lineære normalfordelingsmodeller, men samme argument kan bruges hvis der f.eks. er tale om binære responser, som vi ønsker at forudsige ved hjælp af en logistisk regressionsmodel.

Angående fortolkning af parameterestimaterne skal siges, at så længe man tænker på prediktionssituationen, hvor den relevante fordeling af y_t er den betingede fordeling, givet alle de foregående observationer, så er fortolkningen stort set som den plejer at være i en multipel regressionsmodel. De beregnede fittede værdier skal forstås som prediktioner af næste observation ud fra alle de foregående, og residualplots o.l. skal fortolkes på samme måde som i en lineær model. Specielt skal residualerne som funktion af tiden ligne en følge af uafhængige normalfordelte variable.

Hvis man derimod begynder at interessere sig for den *marginale* fordeling af y_t bliver det mere indviklet. Det er for eksempel ikke trivielt at indse, at denne fordeling faktisk er en normal fordeling — og det gælder heller ikke altid. Men det gælder under antagelser, som ligner dem vi vil gøre i det følgende, så lad os som udgangspunkt antage, at y_t's marginale fordeling er normal.

Betragt ligningen

$$y_t = \gamma + \alpha y_{t-1} + \beta x_{t-1} + \delta z_t + \cdots + \sigma u_t$$

som var vores oprindelig specifikation af modellen. Denne ligning er faktisk en helt legitim opskrivning af modellen, når man (ud over at u_t'erne skal være uafhængige N(0,1)) forudsætter at u_t er stokastisk uafhængig af de tidligere observationer $y_{t-1}, y_{t-2}, \ldots$. Man kan bruge ligningen til at simulere tidsrækker af denne type ved hjælp af almindelige uafhængige N(0,1) (simulerede) variable, når blot den første værdi y_1 vælges i henhold til den marginale fordeling, som vil blive udledt i det følgende.

Lad $\omega_t^2 = \mathrm{var}(y_t)$ betegne variansen på y_t — altså den *marginale varians*, i modsætning til *prediktionsvariansen* σ^2, som i analysen og fortolkningen af modellen mere eller mindre har overtaget den rolle, som

variansen spiller i sædvanlige lineære modeller. Ved at tage variansen på begge sider af modelspecifikationen får vi (idet u_t jo er stokastisk uafhængig af y_{t-1}, så vi kan benytte additionsreglen for varianser på højre side)

$$\omega_t^2 = \alpha^2\omega_{t-1}^2 + \sigma^2 \,.$$

Ved rekursivt at benytte denne ligning, idet vi ovenfor substituerer $\alpha^2\omega_{t-2}^2 + \sigma^2$ for ω_{t-1}^2, dernæst $\alpha^2\omega_{t-3}^2 + \sigma^2$ for ω_{t-2}^2 osv., får vi efter p trin

$$\omega_t^2 = \alpha^{2p}\omega_{t-p}^2 + (1 + \alpha^2 + \alpha^4 + \cdots + \alpha^{2p-2})\sigma^2$$

Af denne ligning følger at $\omega_t^2 \geq (1 + \alpha^2 + \alpha^4 + \cdots + \alpha^{2p-2})\sigma^2$. Hvis vi forestiller os, at modellen skal kunne føres vilkårligt langt tilbage i tid (hvilket vel er et rimeligt krav i de fleste anvendelser i stil med den vi har skitseret) må vi åbenbart kræve, at kvotientrækken $1+\alpha^2+\alpha^4+\ldots$ er konvergent, dvs. at $\alpha^2 < 1$.

Under antagelsen $\alpha^2 < 1$ fås for $p \to \infty$ (forudsat at bidraget fra leddet $\alpha^{2p}\omega_{t-p}^2$ forsvinder i grænsen, hvilket gælder under den ret beskedne antagelse at ω_{t-p}^2 holder sig begrænset)

$$\omega_t^2 = (1 + \alpha^2 + \alpha^4 + \ldots)\sigma^2$$

eller, ifølge formlen for summation af en kvotientrække,

$$\omega_t^2 = \frac{\sigma^2}{1 - \alpha^2} \,.$$

Variansfunktionen bliver således automatisk konstant, når blot det forudsættes at den er begrænset og at modellen kan udvides vilkårligt langt bagud i tid. Den sidste formel viser, hvordan forholdet mellem den marginale varians og prediktionsvariansen afhænger af *autoregressionskoefficienten* α.

Sæt nu $\mu_t = \mathrm{E}y_t$. Ved at tage middelværdi på begge sider i modelspecifikationen får vi så

$$\mu_t = \gamma + \alpha\mu_{t-1} + \beta x_{t-1} + \delta z_t + \ldots$$

Hvis vi her på højre side substituerer $\gamma+\alpha\mu_{t-2}+\beta x_{t-2}+\delta z_{t-1}+\ldots$ for μ_{t-1}, derefter i den fremkomne ligning det tilsvarende udtryk for μ_{t-1} osv., så får vi i grænsen (under et beskedent krav om konvergens, som er opfyldt hvis blot følgerne (μ_t), (x_t) og (z_t) er begrænsede)

$$\mu_t = \gamma\sum_{p=0}^{\infty}\alpha^p + \beta\sum_{p=0}^{\infty}\alpha^p x_{t-1-p} + \delta\sum_{p=0}^{\infty}\alpha^p z_{t-p} + \ldots$$

Det betyder at y_t's marginale middelværdi har samme form som den betingede middelværdi, blot er de forklarende variable (også det underforståede 1–tal i konstantleddet $\gamma = \gamma 1$) erstattet med "baglæns glidende gennemsnit" af de oprindelige forklarende variable, med koefficienter $1, \alpha, \alpha^2 \ldots$. I eksemplet betyder det, at temperaturen på dag t er påvirket (indirekte via sin afhængighed af y_{t-1}) af vindhastigheden på *alle* tidligere dage, dog (forudsat at α ikke er meget tæt på 1 eller -1) på en sådan måde, at effekten af vindhastigheden for lang tid siden er betydningsløs i forhold til det bidrag der stammer fra vindhastigheden i går.

Hvis modellen kun indeholder et konstantled får vi specielt

$$\mu_t = \gamma \sum_{p=0}^{\infty} \alpha^p = \frac{\gamma}{1-\alpha} \, .$$

I dette tilfælde er alle observationerne således identisk fordelte,

$$y_t \sim N\left(\frac{\gamma}{1-\alpha}, \frac{\sigma^2}{1-\alpha^2}\right) \, .$$

Denne tidsrække kaldes en *autoregressiv proces af orden 1* (forkortet AR(1)). Den siges at være *stationær*, fordi fordelingen (ikke alene de endimensionale marginale fordelinger, men også den simultane fordeling af et helt fragment $(y_1, \ldots, y_t)$) er invariant under translation af tidsskalaen.

Tilsvarende er en autoregressiv proces af orden 2 (AR(2)) en stationær tidsrække, givet på formen

$$y_t = \gamma + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \sigma u_t$$

— og så fremdeles.

Hermed nærmer vi os den traditionelle teori for tidsrækker, som langt hen ad vejen beskæftiger sig med tidsrækker uden egentlig middelværdistruktur. Det er en omfattende og kompliceret teori, som først efter lang tids hårdt arbejde fører frem til modeller af den type vi har omtalt her.