Robuste Kovariansmatricer for OLS Estimation af Parametrene i en Regressionsmodel

Forelæsningsnoter til Finansiel Økonometri

Jesper Lund

mail@jesperlund.com
http://www.jesperlund.com

21. marts 2006

1 Indledning og motivation

Denne forelæsningsnote beskriver en kovariansmatrix for OLS estimatoren i en regressionsmodel, som tager højde for autokorrelation og heteroskedasticitet i regressionsmodellens fejlled.[1] Sådanne kovariansestimatorer kaldes ofte for HAC estimatorer, hvor HAC er en forkortelse for **H**eteroskedasticity and **A**utocorrelation **C**onsistent. Et andet navn for metoden er **robuste** kovariansmatricer, hvor "robust" hentyder til at estimatoren tager højde for autokorrelation og heteroskedasticitet.

De klassiske forudsætninger for OLS estimation i en regressionsmodel er som bekendt at fejlleddet har konstant varians (homoskedasticitet), og at der ikke er autokorrelation i fejlleddet. Konsekvenserne af brud på disse forudsætninger er at OLS estimatoren ikke er efficient (der findes en bedre estimator blandt mængden af lineære estimatorer), og at standardafvigelserne for OLS estimaterne er biased, dvs. t-tests og konfidensintervaller vil være forkerte. HAC estimatoren løser det sidstnævnte problem, idet standardafvigelser korrigeres for autokorrelation og heteroskedasticitet af ukendt form.[2]

2 OLS estimation med de klassiske forudsætninger

Vi betragter regressionsmodellen

$$y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t, \tag{1}$$

hvor $\mathbf{x}_t$ er en $K \times 1$ vector af forklarende variable. Typisk er en af disse forklarende variable et konstantled. De klassiske forudsætninger for OLS estimation er

$$\begin{aligned} E(\varepsilon_t) &= 0 && (2)\\ \mathrm{Var}(\varepsilon_t) &= \sigma^2 && (3)\\ \mathrm{Var}(\varepsilon_t, \varepsilon_{t-j}) &= 0 \text{ hvis } j > 0 && (4) \end{aligned}$$

for alle t. Derudover skal ε_t være uafhængig af den stokastiske vektor $\mathbf{x}_t$.[3]

OLS estimatoren for $\boldsymbol{\beta}$ er

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t y_t\right)\\ &= \boldsymbol{\beta} + \left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t\varepsilon_t\right), \end{aligned} \tag{5}$$

[1] Forelæsningsnoten kan eventuelt læses sammen med Appendix A.3 i Campbell *et al* (1997).

[2] GLS (generalized least squares), der er den efficiente lineære estimator, kræver derimod at formen for heteroskedasticitet og/eller autokorrelation specificeres. GLS har ikke samme robusthed som OLS kombineret med HAC estimatoren.

[3] Vi antager **ikke** at ε_t er uafhængig af $\mathbf{x}_{t+j}$ for alle j, idet $\mathbf{x}_t$ kan bestå af laggede værdier af y_t, hvis man bruger OLS til at estimere en AR(p) model (autoregression).

hvor den anden linie fremkommer ved at indsætte $y_t = \boldsymbol{\beta}'\mathbf{x}_t + \varepsilon_t$ i OLS formlen. For at udlede fordelingen for OLS estimatoren kan man opfatte

$$\boldsymbol{\Omega}_{\boldsymbol{X}} \equiv \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \tag{6}$$

som en konstant $K \times K$ matrix (symmetrisk), og $\mathrm{Cov}(\hat{\boldsymbol{\beta}})$ kan beregnes ud fra

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \mathrm{Cov}\left(\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t\right) \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1}. \tag{7}$$

Hvis forudsætningerne (2)–(4) er opfyldt, og $\mathbf{x}_t$ er uafhængig af ε_t, er $\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t$ en sum af ukorrelerede stokastiske vektorer, og

$$\mathrm{Cov}\left(\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t\right) \;=\; \sum_{t=1}^{T} E\left(\mathbf{x_t} \mathbf{x_t}' \varepsilon_t^2\right) \;=\; \sigma^2 \left(\sum_{t=1}^{T} \mathbf{x_t} \mathbf{x_t}'\right), \tag{8}$$

da alle kovariansled mellem $\mathbf{x}_t\varepsilon_t$ og $\mathbf{x}_i\varepsilon_i$ for $t \neq i$ er 0, og ε_t har konstant varians. Hvis vi indsætter ligning (8) i (7), kan vi forenkle udtrykket til

$$\begin{aligned} \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) &= \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \sigma^2 \left(\sum_{t=1}^{T} \mathbf{x_t} \mathbf{x_t}'\right) \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \\ &= \sigma^2 \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1}, \end{aligned} \tag{9}$$

hvilket er den formel som ses i lærebøger om OLS. Den ukendte parameter σ^2 skal naturligvis estimere ved den empiriske varians $s^2 = \frac{1}{T-K} \sum_{t=1}^{T} e_t^2$, hvor $e_t = y_t - \hat{\boldsymbol{\beta}}' \mathbf{x_t}$ er residualerne fra OLS regressionen.

3 OLS med heteroskedasticitet eller autokorrelation

Brud på forudsætning (3) og (4) kaldes hhv. heteroskedasticitet og autokorrelation. Konsekvensen af disse forudsætningsbrud er at vi ikke kan forenkle udledningen af $\mathrm{Cov}\left(\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t\right)$ i ligning (8). Det vil sige, at vi er nødt til at anvende en anden og mere kompliceret formel.

Hvis der kun er tale om heteroskedasticitet, dvs. $E(\varepsilon_t \varepsilon_i) = 0$ for $t \neq i$ gælder stadig, kan følgende estimator udledt af White (1980) anvendes:

$$\mathrm{Cov}\left(\sum_{t=1}^{T} \mathbf{x}_t \varepsilon_t\right) = \sum_{t=1}^{T} e_t^2 \mathbf{x_t} \mathbf{x_t}'. \tag{10}$$

Kovariansmatricen for OLS estimatoren bestemmes ved at indsætte (10) i (7), hvilket giver

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1} \left(\sum_{t=1}^{T} e_t^2 \mathbf{x_t} \mathbf{x_t}'\right) \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1}. \tag{11}$$

Med White's formel behøver vi ikke at antage at $E[\varepsilon_t^2 \mid \mathbf{x}_t] = E[\varepsilon_t^2] = \sigma^2$ som i (8). Formelt er det altså heteroskedasticitet betinget af regressorerne $\mathbf{x}_t$ som kovariansmatrix estimatoren (10) korrigerer for.[4]

Hvis der er autokorrelation i fejlleddet, vil udtrykket for $\mathrm{Cov}\left(\sum_{t=1}^{T} \mathbf{x}_t\varepsilon_t\right)$ indeholde et antal kovariansled mellem $\mathbf{x}_t\varepsilon_t$ og $\mathbf{x}_{t-j}\varepsilon_{t-j}$ for $j > 0$. Når der er autokorrelation i fejlleddet, kan man udlede følgende formel for $\mathbf{\Omega}_T \equiv \mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right)$

$$\mathbf{\Omega}_T = \mathbf{S}_0 + \sum_{j=1}^{L} w_j \left(\mathbf{S}_j + \mathbf{S}_j'\right), \tag{12}$$

hvor w_j er skaleringsfaktor (vægt) som beskrives nedenfor, og $K \times K$ matricen $\mathbf{S}_j$ er defineret ved

$$\mathbf{S} = \sum_{t=j+1}^{T} e_t e_{t-j} \mathbf{x}_t \mathbf{x}_{t-j}'. \tag{13}$$

Kovariansmatricen for OLS estimatoren er således:

$$\mathrm{Cov}\left(\hat{\boldsymbol{\beta}}\right) = \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1} \mathbf{\Omega_T} \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right)^{-1}. \tag{14}$$

Bemærk at White-formlen (11) fremkommer som et specialtilfælde af (14), hvis $L = 0$, dvs. $\mathbf{\Omega}_T = \mathbf{S}_0$.

Vi mangler at beskrive hvordan L og w_j fastsættes. Det afhænger blandt andet af vores eventuelle forhåndsantagelser om strukturen af autokorrelationen i fejlleddet. Vi diskuterer to muligheder nedenfor.

3.1 Autokorrelation af generel form

Den mest generelle situation er selvfølgelig, at vi slet ikke antager noget som helst om autokorrelationsstrukturen, men blot ønsker en estimator som er robust over for en (stort set) vilkårlig form for autokorrelation i fejlleddet. Selv om det måske lyder som en umulig opgave (med uendeligt mange ekstra parametre som skal estimeres), kan den asymptotiske statistiske teori faktisk løse opgaven.

Newey & West (1987) viser, at hvis $L = l(T)$ er en langsomt voksende funktion af T, og hvis $w_j = 1 - \frac{j}{L+1}$, vil (14) være en konsistent estimator for kovariansmatricen for OLS estimatet.[5] Vægtene $w_j = 1 - \frac{j}{L+1}$ sikrer at kovariansmatricen (14) er positiv semi-definit,[6] eller sagt med andre ord: vi risikerer ikke negativ varianser.

[4] White's kovariansmatrix estimator er implementeret i mange standardprogrammer til lineær regression, herunder `PROC REG` i SAS, hvis option `ACOV` anvendes i `MODEL` statement'et, jf. `SAS/STAT` manualen vedr. `PROC REG`.

[5] Lag-længde funktionen $L = l(T)$ skal formelt tilfredsstille betingelserne $\lim_{T\to\infty} l(T) = \infty$ og $\lim_{T\to\infty}[l(T)/T^{1/4}] = 0$, se Theorem 2 i Newey & West (1987). Der er i øvrigt skrevet adskillige artikler om metoder til at fastlægge lag-længden L i HAC formler som (12).

[6] En matrix $\mathbf{\Sigma}$ er positiv definit, hvis $\mathbf{z}'\mathbf{\Sigma}\mathbf{z} > 0$ for alle $\mathbf{z}$. Hvis betingelsen kun er $\mathbf{z}'\mathbf{\Sigma}\mathbf{z} \geq 0$, er matricen $\mathbf{\Sigma}$ positiv semi-definit.

3.2 Autokorrelation af MA(q) formen

Hvis ε_t følger en MA(q) proces som

$$\varepsilon_t = u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q} \tag{15}$$

vides det som bekendt at ε_t og ε_{t-j} er ukorrelerede for $j > q$. Autokorrelation i fejlleddet i en regressionsmodel opstår blandt på grund af brugen af overlappende observationer, og i den situation vil fejlleddet have en moving average struktur.

Når ε_t er MA(q) kendes strukturen i autokorrelationen, og det er ikke nødvendigt at anvende den første generelle metode til at bestemme L. Den teoretisk korrekte korrektion for autokorrelation fremkommer ved at benytte $L = q$ og vægtene $w_j = 1$.

Ulempen ved at anvende $w_j = 1$ er imidlertid, at man ikke er garanteret at kovariansmatricen er positiv semi-definit.[7] Den nemmeste måde at undgå dette problem på er at bruge vægtene $w_j = 1 - \frac{j}{L+1}$, da det sikrer en positiv semi-definit kovariansmatrix, jf. Newey & West (1987).

Det har dog den ulempe, at argumentet for at bruge $L = q$ forudsatte at $w_j = 1$. En pragmatisk løsning for en MA(q) process er derfor at sætte $L = 2q$ og bruge Newey-West vægtene $w_j = 1 - \frac{j}{L+1}$ for at sikre at kovariansmatricen er positiv (semi)-definit.

4 Newey-West kovariansmatricen vha. SAS

I dette afsnit beskrives hvordan SAS kan anvendes til at beregne Newey-West kovariansmatricen med vægtningen $w_j = 1 - \frac{j}{L+1}$.

Det kan desværre ikke gøres direkte i `PROC REG` som med White-metoden til korrektion for heteroskedasticitet, men man kan udnytte at OLS er en GMM estimator, og `PROC MODEL` giver mulighed for at benytte Newey-West formlen ved GMM estimation. En nærmere beskrivelse af GMM ligger dog uden for rammerne af denne forelæsningsnote.

I eksemplet nedenfor estimeres den lineære regressionsmodel

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t \tag{16}$$

med OLS, og standardafvigelserne på $\hat{\beta}_i$ beregnes vha. Newey-West formlen med $L = 6$. Der er angivet en del kommentarer i SAS koden, som gerne skulle forklare hvad de enkelte statements laver, men uden et basalt kendskab til GMM er det dybest set en "black box" metode til at beregne Newey-West standardafvigelser. Output fra `PROC MODEL` ligner i dette tilfælde (OLS som GMM estimation) output fra `PROC REG`, så det burde være rimeligt selvforklarende.

[7] I mange tilfælde går det godt at anvende $w_j = 1$, men man er nødt til at checke om den resulterende kovariansmatrix rent faktisk er positiv definit. Det kan f.eks. gøres ved at beregne egenværdier, idet alle egenværdier af en positiv definit matrix er større end 0 (strengt positive).

```
/* Eksempel på lineær regresssion med Newey-West estimator.    */
/* Udgangspunktet er et SAS datasæt a, som indeholder respons- */
/* variablen y og to forklarende variable x1 og x2.            */
/*                                                             */
PROC REG DATA = a;
  MODEL y = x1 x2;
RUN;

PROC MODEL DATA = a;
  PARMS b0 b1 b2;           /* Liste med navne på parametre        */
  y = b0 + b1*x1 + b2*x2;   /* Regressionsmodellen angives her     */

  /* Her estimeres modellen vha. GMM, som er identisk med OLS.      */
  /* FIT option KERNEL=(BARTLETT, L+1, 0) sørger for at kovarians-  */
  /* matricen beregnes vha. Newey-West metoden. I eksemplet er L=6. */
  /* Alle forklarende variable skal gentages under INSTRUMENTS.     */
  /*                                                                */
  FIT y / GMM KERNEL=(BARTLETT, 7, 0) VARDEF=N;   /* L+1 = 7 her */
  INSTRUMENTS x1 x2;   /* Angiv alle de forklarende variable her */
RUN;
```

Litteratur

Campbell, J.Y., A.W. Lo & A.C. MacKinlay (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton NJ.

Newey, W. and K.D. West (1987), "A Simple Positive Semi-Definite Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55, 703–708.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817–838.