

Bilagor till "Regression med R och Python - Introduktion och fördjupning"

Per Johansson och Mattias Nordin

Senast uppdaterad 26 mars 2025

Innehåll

Bilaga A: Minsta kvadratmetoden (OLS)	7
A.1 OLS i stickprovet	7
A.2 OLS i populationen	11
A.3 Datagenererande process, exogenitet och strukturella modeller	12
A.4 Inferens med fixa kovariater	14
A.5 OLS med stokastiska kovariater	21
A.6 Regressionsanatomi	29
A.7 Regressionsanalys med beroende feltermen	31
 Bilaga B: Generaliserad minsta kvadratmetoden (GLS)	 45
B.1 GLS och tvärsnittsdata	45
B.2 Paneldata och klustrade data	51
B.3 Viktning vid olika sannolikhetsurval	60
 Bilaga C: Instrumentvariabelmetoder (IV)	 63
C.1 IV med en oberoende variabel	63
C.2 Det allmänna fallet	65
 Referenser	 71

Denna bilaga består av tre delar. Bilaga A behandlar grunderna i linjär regression, medan bilaga B fokuserar på den så kallade generaliserade regressionsmodellen. Slutligen behandlar bilaga C instrumentalvariabelmetoder.

Mycket av resultaten som presenteras här är standardresultat som återfinns i många läroböcker inom ekonometri och regressionsanalys, se exempelvis Greene (2012) och Wooldridge (2010). En skillnad mot dessa böcker är att vi här inte utgår från skattningar av parametrar i strukturella modeller utan i stället kommer parametrarna från linjärprojektioner. Vad detta innebär för hur parametrar definieras beskrivs kort i avsnitt A.2 och avsnitt A.3 i bilaga A.

Genomgående används matrisalgebra varför vi först ger en kort introduktion till detta innan vi börjar med bilaga A.

Matrisalgebra

Matriser betecknas generellt med fet stil och stora bokstäver. I detta avsnitt gäller att \mathbf{A} och \mathbf{B} är två matriser av dimension $n \times K$ (n rader och K kolumner) samt $n \times L$ (n rader och L kolumner). Vektorer betecknas generellt med fet stil och små bokstäver. Om inget annat anges kommer en vektor vara en kolumnvektor. I detta avsnitt är \mathbf{c} en kolumnvektor med K element vilket innebär att den har K rader och en kolumn. Den transponerade (se nedan) vektorn, \mathbf{c}' är då en radvektor med dimension $1 \times K$ (en rad och K kolumner).

Transponering: Detta innebär att en matris eller vektor kan sägas rotera och betecknas med $'$ i dessa bilagor. Om \mathbf{A} har dimension $n \times K$ så har \mathbf{A}' dimension $K \times n$. Det gäller att element A_{ij} (värdet i \mathbf{A} på rad i och kolumn j) finns på rad j och kolumn i i \mathbf{A}' .

För två matriser \mathbf{C} och \mathbf{D} sådan att (\mathbf{CD}) existerar, följer att $(\mathbf{CD})' = \mathbf{D}'\mathbf{C}'$.

Vektor och matrismultiplikation: För att matriser och vektorer ska kunna multipliceras måste de vara konforma. Med detta menas att det antal kolumner för den matris/vektor, säg \mathbf{u} , som multipliceras med en annan, säg \mathbf{F} , måste vara lika med antal rader för \mathbf{F} . Således är $\mathbf{A}'\mathbf{B}$ möjlig, medan \mathbf{AB} inte är möjlig. $\mathbf{C} = \mathbf{A}'\mathbf{B}$ har dimension $K \times L$. $\mathbf{f} = \mathbf{Ac}$ och $\mathbf{g}' = \mathbf{c}'\mathbf{A}'$ är bägge möjliga. \mathbf{f} är då en kolumnvektor av dimension $n \times 1$, medan \mathbf{g}' är en radvektor av dimension $1 \times n$. $h = \mathbf{g}'\mathbf{f}$ har då dimension 1×1 , dvs. en skalär.

Kvadratisk matris: En kvadratisk matris har lika många rader som kolumner.

Symmetrisk matris: En matris som är identisk med sin transponat kallas för en symmetrisk matris. För en symmetrisk matris \mathbf{C} gäller alltså att $\mathbf{C} = \mathbf{C}'$. Ett nödvändigt villkor är att matrisen är en kvadratisk matris.

Diagonalmatris: Detta är en symmetrisk matris med element skilda från noll på diagonalen medan elementen utanför diagonalen är noll.

Identitetsmatris: \mathbf{I} är en diagonalmatris med bara ettor på diagonalen. \mathbf{I}_K är en $K \times K$ identitetsmatris.

Inverterbar matris: Om det inte finns några linjära relationer i en symmetrisk matris är matrisen inverterbar. För en skalär k har vi att $\frac{1}{k} \times k = 1$. Man kan säga att $1/k$ är inversen till k . Det är på samma sätt med inversen av en matris. Låt \mathbf{C} ha dimension $K \times K$. Inversen till \mathbf{C} definieras som $\mathbf{C}^{-1}\mathbf{C} \equiv \mathbf{I}_K$. Hur inversen beräknas i praktiken är inget som behövs för

att förstå innehållet i dessa bilagor. Alla matriser är inte inverterbara. En inverterbar matris sägs ha *full rang*.

Idempotent matris C : En idempotent matris är en kvadratisk matris som har följande egenskap: $C^2 = CC = C$.

Spåret av en matris: Spåret (eng. *trace*) av en matris C betecknas $\text{tr}(C)$ och är summan av elementen i diagonalen av en kvadratisk matris. Således är $\text{tr}(I_n) = n$. Notera att spåret av en skalär är lika med värdet av skalären, exempelvis, $\text{tr}(5) = 5$.

Bilaga A: Minsta kvadratmetoden (OLS)

A.1 OLS i stickprovet

Vi utgår från ett stickprov av data $\{\mathbf{x}_i, y_i\}_{i=1}^n$ där i är ett index för en enhet. \mathbf{x}_i är en $(K+1) \times 1$ vektor av oberoende variabler inklusive en konstant, $\mathbf{x}_i = [1 \ x_{i1} \ x_{i2} \ \cdots \ x_{iK-1} \ x_{iK}]'$. Utfallsvariabeln för enhet i betecknas y_i . Utan några som helst antaganden kan vi skriva sambandet mellan Y och X -variablerna som

$$y_i = b_0 + b_1 x_{i1} + \dots + b_K x_{iK} + u_i = \mathbf{x}_i' \mathbf{b} + u_i(\mathbf{b}), \quad (\text{A.1})$$

där \mathbf{b} är en vektor av parametrar, $\mathbf{b} = [b_0 \ b_1 \ b_2 \ \cdots \ b_{K-1} \ b_K]'$ och $u_i(\mathbf{b})$ är en residual, dvs. det som "blir över", vilket är en funktion av \mathbf{b} . Om vi staplar observation 1 på observation 2, observation 2 på observation 3, och så vidare, kan ekvation (A.1) skrivas mer kompakt i termer av vektor och matriser för alla n observationer samtidigt enligt:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{K1} \\ 1 & x_{12} & \dots & x_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{Kn} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} u_1(\mathbf{b}) \\ u_2(\mathbf{b}) \\ \vdots \\ u_n(\mathbf{b}) \end{bmatrix} \quad (\text{A.2})$$

vilket är

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{u}(\mathbf{b}), \quad (\text{A.3})$$

där \mathbf{y} och $\mathbf{u}(\mathbf{b})$ båda är $n \times 1$ vektorer och \mathbf{X} är en matris av dimension $n \times (K+1)$. För givna värden på \mathbf{b} så bestäms residualerna $\mathbf{u}(\mathbf{b})$ så att ekvation (A.3) alltid gäller. Om vi vill hitta unika värden på \mathbf{b} måste vi således ha ett kriterium som begränsar möjliga värden. Ett sådant kriterium är den kvadratisiska funktionen av residualerna. Vi definierar residualkvadratsumman $\text{RSS}(\mathbf{b})$ som

$$\text{RSS}(\mathbf{b}) = \sum_{i=1}^n u_i(\mathbf{b})^2 = \sum_{i=1}^n \left(y_i - \sum_{k=0}^K x_{ki} b_k \right)^2 = \mathbf{u}(\mathbf{b})' \mathbf{u}(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})' (\mathbf{y} - \mathbf{X}\mathbf{b}). \quad (\text{A.4})$$

OLS-estimatoren, $\hat{\beta}$, är det värde på \mathbf{b} som minimerar $\text{RSS}(\mathbf{b})$. Således

$$\hat{\beta} = \arg \min_{\mathbf{b}} \text{RSS}(\mathbf{b}) \quad (\text{A.5})$$

Minimum fås där derivatan med avseende på \mathbf{b} är noll, givet att andraderivatan är negativ. Detta innebär

$$\frac{\partial \text{RSS}(\mathbf{b})}{\partial b_j} = -2 \sum_{i=1}^n x_{ji} \left(y_i - \sum_{k=0}^K x_{ki} \hat{\beta}_k \right) = 0, \quad (\text{A.6})$$

Vi kan samla alla dessa derivator i en $(K+1) \times 1$ vektor:

$$\begin{aligned} \frac{\partial \text{RSS}(\mathbf{b})}{\partial \mathbf{b}} &= -2 \sum_{i=1}^n \mathbf{x}_i \left(y_i - \sum_{k=0}^K x_{ki} \hat{\beta}_k \right) \\ &= -2 \left(\mathbf{X}' \mathbf{y} - \mathbf{X}' \mathbf{X} \hat{\beta} \right) \\ &= \mathbf{0}. \end{aligned} \quad (\text{A.7})$$

Lösningen till denna, så kallade, normalekvation är $\mathbf{X}' \mathbf{y} = \mathbf{X}' \mathbf{X} \hat{\beta}$. Om $\mathbf{X}' \mathbf{X}$ är inverterbar får vi en explicit lösning för estimatoren:

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \\ &= \left(\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\sum_{i=1}^n \mathbf{x}_i y_i \right) \\ &= \left(\frac{1}{n - (K+1)} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n - (K+1)} \sum_{i=1}^n \mathbf{x}_i y_i \right). \end{aligned} \quad (\text{A.8})$$

Det tredje likhetstecknet är ditsatt för att illustrera att OLS-estimatoren kan ses som att vi *skattar* eller *estimerar* $V(\mathbf{X})$ och $\text{cov}(\mathbf{X}, \mathbf{y})$. Med ett intercept i modellen kan OLS estimatoren ses som en kvot mellan två medelvärdesestimater: $\hat{V}(\mathbf{X})$ och $\widehat{\text{cov}}(\mathbf{X}, \mathbf{y})$ och således $\hat{\beta} = \hat{V}(\mathbf{X})^{-1} \widehat{\text{cov}}(\mathbf{X}, \mathbf{y})$.

Eftersom andraderivatan vid lösningen är positiv har vi ett minimum:

$$\left. \frac{\partial^2 \text{RSS}(\mathbf{b})}{\partial \mathbf{b}^2} \right|_{\mathbf{b}=\hat{\beta}} = 2(\mathbf{X}' \mathbf{X}) > \mathbf{0}. \quad (\text{A.9})$$

Vi kan illustrera vad detta innebär i fallet med enkel linjär regression ($K=1$). Vi har då

$$\mathbf{X}' \mathbf{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix}, \quad (\text{A.10})$$

och

$$\mathbf{X}' \mathbf{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}. \quad (\text{A.11})$$

Vi har

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i} \begin{bmatrix} \sum_{i=1}^n x_i^2 & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & n \end{bmatrix}. \quad (\text{A.12})$$

Eftersom $n \sum_{i=1}^n x_i^2 - \sum_{i=1}^n x_i \sum_{i=1}^n x_i = \sum_{i=1}^n (x_i - \bar{x})^2 = ns_X^2$ kan nu OLS-estimatoren skrivas:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \frac{1}{ns_X^2} \begin{bmatrix} \sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i y_i \sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i \sum_{i=1}^n y_i + n \sum_{i=1}^n x_i y_i \end{bmatrix}. \quad (\text{A.13})$$

Notera att första termen i vektorn kan skrivas $n\bar{y} \sum_{i=1}^n x_i^2 - n\bar{x} \sum_{i=1}^n x_i y_i = n\bar{y} \sum_{i=1}^n x_i^2 - n\bar{x}^2 - n\bar{x} \sum_{i=1}^n x_i y_i - n^2 \bar{y} \bar{x} = n\bar{y} s_X^2 - n\bar{x} s_{XY}$. Detta ger

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} \frac{n\bar{y} s_X^2 - n\bar{x} s_{XY}}{ns_X^2} \\ \frac{ns_{XY}}{ns_X^2} \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{s_{XY}}{s_X^2} \\ \frac{s_{XY}}{s_X^2} \end{bmatrix}. \quad (\text{A.14})$$

För OLS-estimatoren, $\hat{\beta}$, kan vi nu definiera residualen för en viss observation i som $\hat{u}_i = y_i - \mathbf{x}_i' \hat{\beta}$. Om vi studerar ekvation (A.6) för $j = 0$ ser vi att $\sum_{i=1}^n \hat{u}_i = 0$ (detta eftersom $x_{i0} = 1$ för alla i) vilket också innebär att $\bar{\hat{u}} = 0$. I fallet med enkel linjär regression har vi då att

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i \Rightarrow \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \Rightarrow \hat{u}_i^2 = (y_i - \bar{y} - \hat{\beta}_1(x_i - \bar{x}))^2. \quad (\text{A.15})$$

Det är rättframt att använda denna ekvation för att visa att $\sum_{i=1}^n \hat{u}_i^2$ minimeras när

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})/(n-1)}{\sum_{i=1}^n (x_i - \bar{x})^2/(n-1)} = \frac{s_{XY}}{s_X^2}, \quad (\text{A.16})$$

vilket är precis det som visades ovan. Notera att residualerna är definierade som $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta} \Rightarrow \mathbf{y} = \mathbf{X}\hat{\beta} + \hat{\mathbf{u}}$. Vi har då att

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'(\mathbf{X}\hat{\beta} + \hat{\mathbf{u}}) \\ &= \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\hat{\mathbf{u}}. \end{aligned} \quad (\text{A.17})$$

Detta kan bara hålla om $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$, vilket innebär att $\sum_{i=1}^n x_{ij} \hat{u}_i = 0$ för $j = 0, 1, \dots, K$. Alla X -variabler är alltså exakt okorrelerade med residualerna i stickprovet.

Prediktionerna, $\hat{\mathbf{y}}$, kan vi ta fram som

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} = \mathbf{P}\mathbf{y}, \quad (\text{A.18})$$

där $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'$ är en matris av dimension $n \times n$ och kallas för *projektionsmatrisen*, eller *hattmatrisen*. Innebörden är att $\mathbf{X}\hat{\beta}$ är en linjär projektion av \mathbf{y} till den yta som bestäms

av \mathbf{X} . Det är denna projektiionsmatris som är grunden för analys av inflytelserika observationer som diskuteras i kapitel 6 i boken. Diagonalelementen av denna matris är det som kallas *leverage* i avsnitt 6.7. För enhet i så är det alltså det i :te diagonalelementet, p_{ii} , som är leverage för denna observation. Denna säger vad som händer med det predicerade värdet på Y om det sanna värdet förändras marginellt, $p_{ii} = \partial \hat{y}_i / \partial y_i$.

A.1.1 Kvadratsummor

Residualkvadratsumman för OLS är $\text{RSS}(\hat{\beta})$ vilken vi i fortsättningen enbart skriver som RSS. Denna är alltså

$$\text{RSS} = (\mathbf{y} - \mathbf{X}\hat{\beta})'(\mathbf{y} - \mathbf{X}\hat{\beta}) = \hat{\mathbf{u}}'\hat{\mathbf{u}} = \sum_{i=1}^n \hat{u}_i^2. \quad (\text{A.19})$$

I boken definierades även följande kvadratsummor:

$$\text{TSS} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{ESS} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2. \quad (\text{A.20})$$

Låt $\bar{\mathbf{y}}$ vara en vektor av dimension $n \times 1$ där varje element är \bar{y} . Kvadratsummorna kan då skrivas

$$\text{TSS} = (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}), \quad \text{ESS} = (\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})'(\mathbf{X}\hat{\beta} - \bar{\mathbf{y}}). \quad (\text{A.21})$$

TSS kan nu skrivas

$$\begin{aligned} \text{TSS} &= (\mathbf{y} - \bar{\mathbf{y}})'(\mathbf{y} - \bar{\mathbf{y}}) \\ &= (\mathbf{X}\hat{\beta} + \hat{\mathbf{u}} - \bar{\mathbf{y}})'(\mathbf{X}\hat{\beta} + \hat{\mathbf{u}} - \bar{\mathbf{y}}) \\ &= (\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})'(\mathbf{X}\hat{\beta} - \bar{\mathbf{y}}) + (\mathbf{X}\hat{\beta} - \bar{\mathbf{y}})'\hat{\mathbf{u}} + \hat{\mathbf{u}}'(\mathbf{X}\hat{\beta} - \bar{\mathbf{y}}) + \hat{\mathbf{u}}'\hat{\mathbf{u}} \\ &= \text{ESS} + (\mathbf{X}\hat{\beta})'\hat{\mathbf{u}} + \bar{\mathbf{y}}'\hat{\mathbf{u}} + \hat{\mathbf{u}}'\mathbf{X}\hat{\beta} + \hat{\mathbf{u}}'\bar{\mathbf{y}} + \text{RSS}. \end{aligned} \quad (\text{A.22})$$

Vi vet sedan tidigare att $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$, vilket innebär att $(\mathbf{X}\hat{\beta})'\hat{\mathbf{u}} = \hat{\beta}'\mathbf{X}'\hat{\mathbf{u}} = 0$ och $\hat{\mathbf{u}}'\mathbf{X}\hat{\beta} = (\mathbf{X}'\hat{\mathbf{u}})'\hat{\beta} = 0$. Slutligen har vi att $\bar{\mathbf{y}}'\hat{\mathbf{u}} = \hat{\mathbf{u}}'\bar{\mathbf{y}} = \bar{y} \sum_{i=1}^n \hat{u}_i = 0$, eftersom vi tidigare konstaterat att summan av residualerna är noll. Därmed har vi att

$$\text{TSS} = \text{ESS} + \text{RSS}. \quad (\text{A.23})$$

A.1.2 Perfekt multikolinjäritet

Perfekt multikolinjäritet inträffar när det finns ett *linjärt beroende* mellan de ingående variablerna i matrisen \mathbf{X} . Med linjärt beroende menas här inte statistiskt beroende utan detta koncept kommer från *matrisalgebra*. Vi kan tänka på matrisen \mathbf{X} som att den består av $K + 1$ kolumnvektorer, en för varje variabel. Linjärt beroende inträffar om en av dessa kolumnvektorer kan skrivas som en linjärkombination av en eller flera av de andra vektorerna. Exempelvis är

följande matris, \mathbf{A} , inte linjärt oberoende:

$$\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \mathbf{a}_3 \quad \mathbf{a}_4] = \begin{bmatrix} 4 & 0 & 2 & 4 \\ 1 & 1 & 1 & 7 \\ -4 & 2 & -1 & 5 \\ 1 & 5 & 3 & 1 \\ -10 & 4 & -3 & 7 \\ 4 & 0 & 2 & 4 \\ 2 & 0 & 1 & 8 \end{bmatrix}. \quad (\text{A.24})$$

Detta är fallet eftersom $\mathbf{a}_1 = -\mathbf{a}_2 + 2\mathbf{a}_3$. När det finns linjärt beroende i matrisen \mathbf{X} så kommer $\mathbf{X}'\mathbf{X}$ inte att vara inverterbar. Då går det inte att ta fram OLS-skattningen enligt ekvation (A.8).

Ett fall när perfekt multikolinjäritet inträffar, så att $\mathbf{X}'\mathbf{X}$ inte är inverterbar, är när vi har med en full uppsättning dummyvariabler tillsammans med en konstant. Med en konstant så finns det en kolumn i \mathbf{X} som har värdet 1 för alla n rader. Låt säga att vi har en kategorisk variabel med L olika kategorier som tillsammans är heltäckande och ömsesidigt uteslutande (varje observation tillhör en och endast en variabel). Om vi i \mathbf{X} -matrisen tar med L stycken kolumner för dummyvariablerna, $\mathbf{d}_1, \mathbf{d}_2, \dots, \mathbf{d}_L$, sådant att \mathbf{d}_l är en vektor som antar värdet för de rader som tillhör kategori l och noll annars. I så fall gäller att $\mathbf{d}_1 + \mathbf{d}_2 + \dots + \mathbf{d}_L = \mathbf{1}$, vilket är identiskt med konstant-kolumnen. Matrisen \mathbf{X} har då ett linjärt beroende och vi kan inte ta fram OLS-skattningar. Lösningen är att utesluta en av dummyvariablerna, där motsvarande kategori får agera referenskategori.¹

A.2 OLS i populationen

OLS-skattaren, $\hat{\beta}$, för ett stickprov kan ses som skattare av en vektor β i populationen. För att se hur denna definieras utgår vi från

$$Y = \mathbf{x}'\mathbf{b} + U, \quad (\text{A.25})$$

där Y är en slumpvariabel (därav skrivs den med stor bokstav) och $\mathbf{x} = [1 \quad X_1 \quad X_2 \quad \dots \quad X_K]'$ är en vektor av slumpvariabler. \mathbf{b} är en godtycklig $(K+1) \times 1$ vektor av konstanter. Vi kan nu definiera β utifrån följande kriterium

$$\beta = \arg \min_{\mathbf{b}} E((Y - \mathbf{x}'\mathbf{b})^2). \quad (\text{A.26})$$

Detta är alltså populationsanalogen till OLS-problemet för ett stickprov som definierades i ekvation (A.5). För att denna parametervektor ska existera så krävs att $E(Y^2) < \infty$ och $E(X_k^2) < \infty$ för alla $k = 1, \dots, K$. I så fall kan vi ta fram β genom att derivera $E((Y - \mathbf{x}'\mathbf{b})^2)$ med avseende på \mathbf{b} och sätta denna lika med noll:

$$\frac{\partial}{\partial \mathbf{b}} E((Y - \mathbf{x}'\mathbf{b})^2) = -2 E(\mathbf{x}(Y - \mathbf{x}'\beta)) = \mathbf{0}. \quad (\text{A.27})$$

¹Alternativt går det att utesluta konstanten, men vi utgår i denna bilaga, – så länge inget annat anges – från att det finns med en konstant i modellen.

Vi har att

$$E(\mathbf{x}Y) = E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta}. \quad (\text{A.28})$$

Om $E(\mathbf{x}\mathbf{x}')$ är inverterbar får vi då

$$\boldsymbol{\beta} = (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}Y). \quad (\text{A.29})$$

Med denna lösning har vi nu

$$Y = \mathbf{x}'\boldsymbol{\beta} + U. \quad (\text{A.30})$$

Om vi sätter in denna ekvation i (A.29) får vi

$$\begin{aligned} \boldsymbol{\beta} &= (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}(\mathbf{x}'\boldsymbol{\beta} + U)) \\ &= (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}\mathbf{x}')\boldsymbol{\beta} + (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}U) \\ &= \boldsymbol{\beta} + (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}U). \end{aligned} \quad (\text{A.31})$$

Vi får alltså att $E(\mathbf{x}U) = \mathbf{0}$. Då den första raden i \mathbf{x} är 1 (vi utgår hela tiden från att vi har med en konstant i modellen) innebär detta även att $E(1 \cdot U) = E(U) = 0$. Detta i sin tur innebär att $\text{cov}(X_j, U) = 0$ för alla $j = 0, 1, \dots, K$.

Notera att denna modell inte säger något om hur $E(Y|\mathbf{x})$ ser ut, och inte heller att det finns ett kausalt samband mellan dessa variabler. Den säger bara att sambandet existerar i populationen. $\mathbf{x}'\boldsymbol{\beta}$ kallas för en *linjär projektion*. Vi säger att ett samband är linjärt om $E(Y|\mathbf{x})$ går att skriva som en linjär funktion av \mathbf{x} . Om vi tar väntevärdet av (A.30) betingat på \mathbf{x} får vi

$$E(Y|\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta} + E(U|\mathbf{x}). \quad (\text{A.32})$$

Sambandet mellan Y och \mathbf{x} är alltså linjärt om $E(U|\mathbf{x}) = 0$. Notera att detta säger något starkare än att $\text{cov}(X_j, U) = 0$. Detta faktum innebär att det inte finns några garantier att $E(U|\mathbf{x}) = 0$. Om så inte är fallet så kan vi istället tänka på $\mathbf{x}'\boldsymbol{\beta}$ som en linjär approximation av $E(Y|\mathbf{x})$.

A.3 Datagenererande process, exogenitet och strukturella modeller

I boken och den här bilagan definierar vi parametrarna, $\boldsymbol{\beta}$, utifrån givna variabler och en population enligt ekvation (A.29). Däremot har vi inte en explicit modell för hur data har genererats. Anledningen till att vi valt att göra på detta sättet är att boken inte är skriven för ett visst ämne och vissa frågeställningar, utan är tänkt att kunna användas för många olika tillämpningsområden och för både deskriptiva, prediktiva och kausala frågeställningar.

Om man istället utgår från en specifik modell, som i olika kontexter kan kallas för exempelvis en *strukturell modell* eller en *datagenererande process*, så kan parametrarna som man *vill* estimeras skilja sig från $\boldsymbol{\beta}$. Många läroböcker tar också ett sådant perspektiv, så det kan vara hjälpsamt med en jämförelse mellan dessa olika perspektiv.

En strukturell modell kan, exempelvis, se ut på följande sätt:

$$Y = \mathbf{x}'\boldsymbol{\gamma} + V, \quad (\text{A.33})$$

där $\boldsymbol{\gamma}$ är strukturella parametrar – och därför fixa – och V är en felterm. Skillnaden från den linjära regressionsmodellen i ekvation (A.30) är således att $\boldsymbol{\gamma}$ inte är definierade utifrån att de minimerar kvadratfelet utan de antas istället givna utifrån en modell för hur verkligheten fungerar. V definieras av Y, \mathbf{x} och $\boldsymbol{\gamma}$, medan U är implicit definierad via första ordningens villkor: $E(\mathbf{x}U) = \mathbf{0}$. Således, V är en distinkt annan felterm än U .

För att det ska gå att få konsistenta skattningar av $\boldsymbol{\gamma}$ måste $E(\mathbf{x}V) = 0$.² Det är viktigt att trycka på att detta inte betyder samma sak som att $E(\mathbf{x}U) = 0$ (vilket gäller definitionsmsigt).

Vi kan illustrera skillnaden mellan $\boldsymbol{\beta}$ och $\boldsymbol{\gamma}$ med en specifik datagenererande process enligt ekvation (A.33). Låt oss anta att data i populationen har följande multivariata normalfördelning:

$$\begin{bmatrix} X_1 \\ X_2 \\ V \end{bmatrix} \sim \mathcal{N} \left(\begin{bmatrix} 1,2 \\ -0,4 \\ 0,2 \end{bmatrix}, \begin{bmatrix} 1 & 0,3 & 0 \\ 0,3 & 2 & -0,3 \\ 0 & -0,3 & 1,5 \end{bmatrix} \right) \quad (\text{A.34})$$

Vi har vektorn av oberoende variabler $\mathbf{x} = [1 \ X_1 \ X_2]'$ och en parametervektor $\boldsymbol{\gamma} = [0,4 \ 0,3 \ -0,5]'$. Ekvation (A.29) ger oss $\boldsymbol{\beta}$. Med lite uträkningar kan vi komma fram till att

$$E(\mathbf{x}\mathbf{x}') = \begin{bmatrix} 1 & E(X_1) & E(X_2) \\ E(X_1) & E(X_1^2) & E(X_1X_2) \\ E(X_2) & E(X_1X_2) & E(X_2^2) \end{bmatrix} = \begin{bmatrix} 1 & 1,2 & -0,4 \\ 1,2 & 2,44 & -0,18 \\ -0,4 & -0,18 & 2,16 \end{bmatrix}, \quad (\text{A.35})$$

och

$$E(\mathbf{x}Y) = \begin{bmatrix} 1,16 \\ 1,542 \\ -1,674 \end{bmatrix}. \quad (\text{A.36})$$

Slutligen, genom att invertera $E(\mathbf{x}\mathbf{x}')$ och multiplicera ihop får vi

$$\boldsymbol{\beta} = (E(\mathbf{x}\mathbf{x}'))^{-1} E(\mathbf{x}Y) = \begin{bmatrix} 0,481 \\ 0,347 \\ -0,657 \end{bmatrix} \neq \boldsymbol{\gamma} = \begin{bmatrix} 0,4 \\ 0,3 \\ -0,5 \end{bmatrix} \quad (\text{A.37})$$

Anledningen till att dessa två parametervektorer skiljer sig åt är att det finns ett samband mellan \mathbf{x} och V , dvs. $E(\mathbf{x}V) \neq 0$. Om istället $E(\mathbf{x}V) = 0$ så hade de varit desamma.³

²För att få väntevärdesriktiga skattningar måste $E(V|\mathbf{x}) = 0$. Inom ekonometri kallas dessa bägge förutsättningar ibland för svag respektive stark exogenitet (se, exempelvis, Wooldridge 2010).

³För att $\beta_0 = \gamma_0$ krävs dessutom att $E(V) = 0$, men interceptet är sällan så intressant.

A.4 Inferens med fixa kovariater

Vårt mål är att utifrån vårt stickprov av data, med hjälp av $\hat{\beta}$ kunna säga något om β . Till att börja med kommer vi att betinga analysen på kovariaterna (ett annat namn på de oberoende variablerna). Detta innebär att vi betraktar dessa som fixa snarare än stokastiska. Detta är rimligt i fall när forskaren bestämmer värdena på \mathbf{X} i ett experiment. I sådana fall kommer enbart U (och Y genom U) att vara stokastisk. Med observationsdata är det dock inte rimligt att tänka på \mathbf{X} som fix utan den bör betraktas som stokastisk. Som vi kommer se, kommer vi dock ha nytta av de betingade resultaten när vi går vidare till att diskutera de obetingade så vi börjar med att betinga på \mathbf{X} .

För vårt stickprov av data kan ekvation (A.30) skrivas som

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}. \quad (\text{A.38})$$

Vi kan sätta in detta i ekvation (A.8), vilket gör att vi erhåller

$$\begin{aligned} \hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \beta + \mathbf{H}\mathbf{u}, \end{aligned} \quad (\text{A.39})$$

där $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Vi ser således att $\hat{\beta}$ är en funktion av β , men också av kovariaterna och feltermerna som ingår i modellen.

Vi kan nu fråga oss vilka egenskaper $\hat{\beta}$ har. Är den exempelvis väntevärdesriktig? För att undersöka detta kan vi notera att vi kan skriva

$$\mathbb{E}(\hat{\beta}|\mathbf{X}) = \beta + \mathbf{H}\mathbb{E}(\mathbf{u}|\mathbf{X}), \quad (\text{A.40})$$

där vi kan flytta ut \mathbf{H} utanför väntevärdesoperatoren då vi betraktar \mathbf{X} som fix. Slutsatsen är alltså att vi har en väntevärdesriktig estimator om $\mathbb{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$. Som diskuterades ovan så kan vi tänka på populationsanalogen, $\mathbb{E}(U|\mathbf{x}) = 0$ som att det beskriver linjäritet. Men för att även $\mathbb{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ ska gälla så räcker det inte att $\mathbb{E}(U|\mathbf{x}) = 0$ utan det krävs också att stickprovet är slumpmässigt draget från populationen. För att se varför, antag att vi bara skulle dra observationer som har stora värden på feltermen. I så fall så skulle $\mathbb{E}(\mathbf{u}|\mathbf{X}) > \mathbf{0}$ även om $\mathbb{E}(U|\mathbf{x}) = 0$. Men om vi drar observationerna slumpmässigt kommer inte detta ske utan då kommer $\mathbb{E}(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ om $\mathbb{E}(U|\mathbf{x}) = 0$. Vi kommer genomgående, så länge inte något annat anges, utgå från att stickprovet är slumpmässigt draget från populationen.

Den betingade variansen kan vi skriva

$$\begin{aligned} \mathbf{V}(\hat{\beta}|\mathbf{X}) &= \mathbb{E}\left((\hat{\beta} - \beta)(\hat{\beta} - \beta)'|\mathbf{X}\right) \\ &= \mathbb{E}(\mathbf{H}\mathbf{u}\mathbf{u}'\mathbf{H}'|\mathbf{X}) \\ &= \mathbf{H}\mathbb{E}(\mathbf{u}\mathbf{u}'|\mathbf{X})\mathbf{H}' \end{aligned} \quad (\text{A.41})$$

Om feltermen är *sfärisk* (oberoende och homoskedastisk) så gäller att

$$\mathbb{E}(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma_U^2 \mathbf{I}_n. \quad (\text{A.42})$$

Notera att $\mathbf{H}\mathbf{H}' = (\mathbf{X}'\mathbf{X})^{-1}$. Under förutsättningen att feltermerna är sfäriska får vi då

$$V(\hat{\beta}|\mathbf{X}) = \sigma_U^2 \mathbf{I}_n (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{A.43})$$

Under förutsättning att $E(U|\mathbf{x}) = 0$, ett slumpmässigt stickprov från populationen och sfäriska feltermer så går det att visa att OLS-estimatoren är den mest effektiva (lägst varians) estimatoren av alla väntevärdesriktiga estimatorer, något som kallas *Gauss-Markov-teoremet*.

För att visa detta, notera först att OLS-estimatoren kan skrivas $\hat{\beta} = \mathbf{H}\mathbf{Y}$. Eftersom \mathbf{H} är fix innebär detta att OLS-estimatoren är linjär med vikter som ges av \mathbf{H} . Låt oss definiera en annan linjär estimator:

$$\hat{\beta}_a = \mathbf{C}\mathbf{Y} = \mathbf{C}\mathbf{X}\beta + \mathbf{C}\mathbf{u}, \quad (\text{A.44})$$

där $\mathbf{C} \neq \mathbf{H}$ är en matris med konstanter. Eftersom

$$\begin{aligned} E(\hat{\beta}_a|\mathbf{X}) &= E(\mathbf{C}\mathbf{X}\beta + \mathbf{C}\mathbf{u}|\mathbf{X}) \\ &= \mathbf{C}\mathbf{X}\beta + \mathbf{C} E(\mathbf{u}|\mathbf{X}) \\ &= \mathbf{C}\mathbf{X}\beta, \end{aligned} \quad (\text{A.45})$$

så krävs det att $\mathbf{C}\mathbf{X} = \mathbf{I}_n$ för att $\hat{\beta}_a$ ska vara väntevärdesriktig, vilket alltså innebär att

$$\hat{\beta}_a = \beta + \mathbf{C}\mathbf{u}. \quad (\text{A.46})$$

Detta innebär att

$$V(\hat{\beta}_a|\mathbf{X}) = \sigma_U^2 \mathbf{C}\mathbf{C}'. \quad (\text{A.47})$$

Skillnaden mellan de två estimatorerna är då

$$\hat{\beta}_a - \hat{\beta} = (\mathbf{C} - \mathbf{H})\mathbf{y} = \mathbf{D}\mathbf{y}, \quad (\text{A.48})$$

där $\mathbf{D} = \mathbf{C} - \mathbf{H}$ och således $\mathbf{C} = \mathbf{D} + \mathbf{H}$. Detta innebär att

$$V(\hat{\beta}_a|\mathbf{X}) = \sigma_U^2 (\mathbf{D}\mathbf{D}' + \mathbf{D}\mathbf{H}' + \mathbf{H}\mathbf{D}' + \mathbf{H}\mathbf{H}') \quad (\text{A.49})$$

För att $\hat{\beta}_a$ skulle vara väntevärdesriktig krävdes $\mathbf{C}\mathbf{X} = \mathbf{I}_n$. Eftersom $\mathbf{C} = \mathbf{D} + \mathbf{H}$, innebär detta att $(\mathbf{D} + \mathbf{H})\mathbf{X} = \mathbf{I}_n$. I och med att $\mathbf{H}\mathbf{X} = \mathbf{I}_n$ medför det att $\mathbf{D}\mathbf{X} = \mathbf{0}$ och $\mathbf{X}'\mathbf{D}' = \mathbf{0}$. Med $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, innebär detta att $\mathbf{D}\mathbf{H}' = \mathbf{D}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{0}$ och $\mathbf{H}\mathbf{D}' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}' = \mathbf{0}$. Således gäller att

$$V(\hat{\beta}_a|\mathbf{X}) = \sigma^2 (\mathbf{D}\mathbf{D}' + \mathbf{H}\mathbf{H}') = \sigma_U^2 (\mathbf{D}\mathbf{D}' + (\mathbf{X}'\mathbf{X})^{-1}) \geq \sigma_U^2 (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{A.50})$$

Från detta följer att variansen för OLS är lägre än alla andra linjära väntevärdesriktiga estimatorer:

$$V(\hat{\beta} | \mathbf{X}) \leq V(\hat{\beta}_a | \mathbf{X}). \quad (\text{A.51})$$

Detta kallas Gauss-Markov-teoremet.

För att kunna göra inferens behöver vi skatta $V(\hat{\beta}|\mathbf{X})$. För att göra detta behöver vi först skatta $E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \sigma_U^2 \mathbf{I}_n$. En naiv skattare av σ_U^2 skulle vara medelvärdet av \hat{u}_i^2 , $\text{RSS}/n = \hat{\mathbf{u}}'\hat{\mathbf{u}}/n$. Det visar sig dock att vi måste göra en frihetsgradsjustering och dela på $n - (K + 1)$ istället för

n för att få en väntevärdesriktig skattare. För att se detta, notera att vi kan skriva residualerna som

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \\ &= (\mathbf{I}_n - \mathbf{P})\mathbf{y} = \mathbf{M}\mathbf{y} = \mathbf{M}\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{u} \\ &= \mathbf{M}\mathbf{u},\end{aligned}\tag{A.52}$$

där $\mathbf{M} = \mathbf{I}_n - \mathbf{P}$. Denna matris kallas förintarmatris ("annihilator matrix"). Skälet till namnet är att $\mathbf{M}\mathbf{X} = \mathbf{0}$. \mathbf{M} har två viktiga egenskaper: den är symmetrisk ($\mathbf{M} = \mathbf{M}'$) och idempotent ($\mathbf{M} = \mathbf{M}\mathbf{M}$). Den första egenskapen följer av

$$\mathbf{M}' = (\mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' = \mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{M},\tag{A.53}$$

medan den andra egenskapen följer av

$$\begin{aligned}\mathbf{M}\mathbf{M} &= (\mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')(\mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \mathbf{I}_n - 2\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{I}_n - 2\mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{I}_n - \mathbf{X}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\ &= \mathbf{M}.\end{aligned}\tag{A.54}$$

Dessa egenskaper ger tillsammans med ekvation (A.52)

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \mathbf{u}'\mathbf{M}'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u},\tag{A.55}$$

Vi kan nu använda oss av ett trick som involverar spåret av matriser. För det första vet vi att spåret av en skalär är skalären, vilket innebär att $\mathbf{u}'\mathbf{M}\mathbf{u} = \text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u})$. För det andra gäller att spåret är oförändrat under cykliskt skift sådant att för tre matriser \mathbf{A} , \mathbf{B} och \mathbf{C} gäller att $\text{tr}(\mathbf{A}\mathbf{B}\mathbf{C}) = \text{tr}(\mathbf{B}\mathbf{C}\mathbf{A}) = \text{tr}(\mathbf{C}\mathbf{A}\mathbf{B})$. I vårt fall har vi då att $\text{tr}(\mathbf{u}'\mathbf{M}\mathbf{u}) = \text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')$. Detta ger alltså

$$\text{E}(\hat{\mathbf{u}}'\hat{\mathbf{u}}|\mathbf{X}) = \text{E}(\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}).\tag{A.56}$$

Generellt gäller att för en stokastisk matris \mathbf{A} så är $\text{E}(\text{tr}(\mathbf{A})) = \text{tr}(\text{E}(\mathbf{A}))$. Tillsammans med det faktum att \mathbf{M} enbart beror på \mathbf{X} , vilken är fix, och antagandet om sfäriska feltermar så får vi

$$\begin{aligned}\text{E}(\hat{\mathbf{u}}'\hat{\mathbf{u}}|\mathbf{X}) &= \text{E}(\text{tr}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}) \\ &= \text{tr}(\text{E}(\mathbf{M}\mathbf{u}\mathbf{u}')|\mathbf{X}) \\ &= \text{tr}(\mathbf{M}\text{E}(\mathbf{u}\mathbf{u}')|\mathbf{X}) \\ &= \text{tr}(\mathbf{M}\sigma_U^2\mathbf{I}_n) \\ &= \sigma_U^2\text{tr}(\mathbf{M}) \\ &= \sigma_U^2\text{tr}(\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= \sigma_U^2(\text{tr}(\mathbf{I}_n) - \text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')).\end{aligned}\tag{A.57}$$

Vi kan återigen använda den cykliska egenskapen hos spåret så att $\text{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \text{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \text{tr}(\mathbf{I}_{K+1})$. Spåret av en identitetsmatris är lika med antalet rader (eller, ekvivalent, antalet kolumner) hos matrisen, vilken i detta fall är $K + 1$. Vi får alltså

$$\text{E}(\hat{\mathbf{u}}'\hat{\mathbf{u}}|\mathbf{X}) = \sigma_U^2(n - (K + 1)) \Rightarrow \text{E}(s_U^2|\mathbf{X}) = \sigma_U^2,\tag{A.58}$$

där $s_U^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}/(n - (K + 1)) = \text{RSS}/(n - (K + 1))$. Den naturliga skattaren av variansen för $\hat{\beta}$ är

$$\hat{V}(\hat{\beta}|\mathbf{X}) = s_U^2(\mathbf{X}'\mathbf{X})^{-1} \quad (\text{A.59})$$

Denna kallas för den skattade *varians-kovariansmatrisen* (eller bara *kovariansmatrisen*) för $\hat{\beta}$.

Antag nu att feltermen, U , är normalfördelad med medelvärde 0 och varians σ_U^2 . Innebörden av detta är att oberoende slumpmässiga dragningar från U nu kan skrivas:

$$\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_U^2 \mathbf{I}_n). \quad (\text{A.60})$$

Eftersom $\hat{\beta} = \beta + \mathbf{H}\mathbf{u}$, där \mathbf{H} är en matris som är konstant får vi att $V(\hat{\beta}|\mathbf{X}) = \mathbf{H} V(\mathbf{u}\mathbf{u}'|\mathbf{X}) \mathbf{H}' = E(\mathbf{u}\mathbf{u}'|\mathbf{X}) \mathbf{H}'$, och det faktum att $\mathbf{H}\mathbf{H}' = (\mathbf{X}'\mathbf{X})^{-1}$ gör att

$$\hat{\beta}|\mathbf{X} \sim \mathcal{N}(\beta, \sigma_U^2(\mathbf{X}'\mathbf{X})^{-1}). \quad (\text{A.61})$$

Nedan går vi igenom exakta egenskaper för teststatistikor under antagandet att feltermen är normalfördelad. Antagandet kan vara nödvändigt för inferens när vi har små stickprov. Notera att om vi har data från ett experiment kan man testa om det finns en effekt genom ett Fisher-test (kallas också för exakt test eller randomiseringstest) som inte bygger på några antaganden. Vi rekommenderar därför att man gör detta om man har få individer i ett experiment. För mer information, se avsnitt 12.2.2.2 i boken.

A.4.1 Hypotesprövning

A.4.1.1 Enkel hypotesprövning

Vi utgår från följande nollhypotes

$$H_0 : \beta_j = \beta_j^{H_0} \quad (\text{A.62})$$

En intuition för vad detta innebär är att om $\hat{\beta}_j$ är långt från $\beta_j^{H_0}$ så ska vi förkasta H_0 . Frågan är vad är långt? För att kunna svara på frågan används en teststatistika med kända egenskaper. Detta ger ett kriterium för när något ska ses som långt ifrån.

Den statistika som används i regressionsanalys är följande

$$T^{obs} = \frac{\hat{\beta}_j - \beta_j^{H_0}}{\hat{s}_{\hat{\beta}_j}}. \quad (\text{A.63})$$

Här är $\hat{s}_{\hat{\beta}_j} = \sqrt{s_U^2 S_{jj}}$, standardfelet för koefficienten $\hat{\beta}_j$, med S_{jj} diagonalelement j i matrisen $(\mathbf{X}'\mathbf{X})^{-1}$ (dvs. $s_{\hat{\beta}_j}^2$ är diagonalelement j i den skattade kovariansmatrisen i ekvation A.59).

Från ekvation (A.61) vet vi att om H_0 är sann så gäller att

$$\frac{\hat{\beta}_j - \beta_j^{H_0}}{\sqrt{\sigma_U^2 S_{jj}}} \sim \mathcal{N}(0, 1). \quad (\text{A.64})$$

Genom att multiplicera och dividera ekvation (A.63) med σ_U kan denna skrivas

$$T^{obs} = \frac{\hat{\beta}_j - \beta_j^{H_0}}{s_{\hat{\beta}_j}} = \frac{\hat{\beta}_j - \beta_j^{H_0}}{\sigma_U \sqrt{S_{jj}}} \cdot \frac{\sigma_U \sqrt{S_{jj}}}{\sqrt{\frac{s_U^2}{\sigma_U^2}}}. \quad (\text{A.65})$$

Vi känner fördelningen för täljaren. Således, om fördelningen för statistikan i nämnaren är känd kan man kanske hitta fördelningen för T^{obs} i ekvation (A.63).

Vi har

$$\frac{(n - (K + 1))s_U^2}{\sigma_U^2} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma_U^2} = \left(\frac{\hat{\mathbf{u}}}{\sigma_U} \right)' \mathbf{M} \left(\frac{\hat{\mathbf{u}}}{\sigma_U} \right) = \frac{\sum_{i=1}^n \sum_{j=1}^n \hat{u}_i \hat{u}_j M_{ij}}{\sigma_U^2}. \quad (\text{A.66})$$

Här är M_{ij} elementet i rad i och kolumn j i \mathbf{M} . Även om feltermerna är oberoende finns ett beroende mellan residualerna eftersom vi viktar med M_{ij} . Det visar sig att om vi känner värdet på \hat{u}_i för $i = 1, \dots, n - (K + 1)$ kan de $K + 1$ övriga residualerna bestämmas utifrån våra OLS-estimat. Vi säger att \mathbf{M} har rangen $n - (K + 1)$. Detta innebär att matrisen som har dimension $(n \times n)$ har $K + 1$ linjära relationer, vilket innebär att statistikan är χ^2 -fördelad med $n - (K + 1)$ frihetsgrader (fg):

$$\frac{(n - (K + 1))s_U^2}{\sigma_U^2} \sim \chi_{n-(K+1)}^2, \quad (\text{A.67})$$

och således

$$\frac{s_U^2}{\sigma_U^2} \sim \chi_{n-(K+1)}^2 / (n - (K + 1)). \quad (\text{A.68})$$

Från Student (1908) vet vi kvoten av en standardnormalfördelad variabel med kvadratroten av en χ^2 -fördelad variabel är t -fördelad med antal frihetsgrader som ges av χ^2 -fördelningen, förutsatt att de två variablerna är oberoende. Men är de oberoende?

Kovariansen mellan $\hat{\mathbf{u}}$ och $\hat{\beta}$ ges av $\text{cov}(\hat{\beta}, \hat{\mathbf{u}}|\mathbf{X}) = \text{E}((\hat{\beta} - \text{E}(\hat{\beta}|\mathbf{X}))\hat{\mathbf{u}}'|\mathbf{X}) = \text{E}((\hat{\beta} - \beta)\hat{\mathbf{u}}'|\mathbf{X})$. Genom att använda oss av algebran för OLS, dvs. $\hat{\mathbf{u}} = \mathbf{M}\mathbf{u}$ och $\hat{\beta} - \beta = \mathbf{H}\mathbf{u}$, får vi $\text{cov}(\hat{\beta}, \hat{\mathbf{u}}|\mathbf{X}) = \text{E}(\mathbf{H}\mathbf{u}\mathbf{u}'\mathbf{M}'|\mathbf{X}) = \sigma_U^2 \mathbf{H}\mathbf{M} = \mathbf{0}$, där det sista uttrycket följer från förintaregenskapen hos \mathbf{M} . Eftersom både \mathbf{u} och $\hat{\beta}$ är normalfördelade innebär detta att de två variablerna är oberoende. Således gäller att

$$\frac{\hat{\beta}_j - \beta_j^{H_0}}{s_{\hat{\beta}_j}} \sim t_{n-(K+1)}, \quad (\text{A.69})$$

Om vi väljer en signifikans- eller risknivå på $100 \cdot \alpha$ % är beslutskriterierna enligt följande:

- Om $H_1 : \beta_j \neq \beta_j^{H_0}$: förkasta H_0 om $|t| > t_{[n-(K+1), \alpha/2]}$,
- Om $H_1 : \beta_j > \beta_j^{H_0}$: förkasta H_0 om $t > t_{[n-(K+1), \alpha]}$,
- Om $H_1 : \beta_j < \beta_j^{H_0}$: förkasta H_0 om $t < -t_{[n-(K+1), \alpha]}$,

där $t_{[n-(K+1), \alpha]}$ är det kritiska värdet.

Notera att ett p -värde är den minsta möjliga risk när H_0 kan förkastas och inte en signifikansnivå som används för ett beslut.

A.4.1.2 Sammansatt hypotesprövning

Den enkla hypotesprövningen utgår från att vi vill testa om en β -koefficient antar ett specifikt värde. Men hur gör vi om vi vill testa flera koefficienter samtidigt? Vi kan använda oss av det som kallas för ett *Wald-test*, vilket innebär att vi kan skriva nollhypotesen på ett kompakt sätt med linjära restriktioner som

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}. \quad (\text{A.70})$$

\mathbf{R} är en matris av dimension $Q \times K + 1$, där Q är antalet restriktioner, medan \mathbf{r} är en $Q \times 1$ vektor med värden som vi testar att restriktionerna ska vara lika. Säg exempelvis att $K = 3$ och vi vill testa om $\beta_1 = 5$ och $\beta_2 = 1$. Vi får då att

$$\mathbf{R} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \mathbf{r} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}, \quad (\text{A.71})$$

viket ger nollhypotesen

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 1 \end{bmatrix}. \quad (\text{A.72})$$

Om vi istället vill testa hypotesen $H_0 : \beta_1 - \beta_2 = 2$: får vi

$$H_0 : \mathbf{R}\boldsymbol{\beta} = \begin{bmatrix} 0 & 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = [\beta_1 - \beta_2] = [2]. \quad (\text{A.73})$$

För att genomföra testet utgår vi från att $U \sim \mathcal{N}(0, \sigma_U^2)$ och att vi med obundet slumpmässigt urval har $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \sigma_U^2 \mathbf{I}_n)$. Vi vet då att ekvation (A.61) håller, dvs.

$$\hat{\boldsymbol{\beta}} | \mathbf{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma_U^2 (\mathbf{X}'\mathbf{X})^{-1}). \quad (\text{A.74})$$

Då \mathbf{R} och \mathbf{r} är konstanter innebär detta att – under nollhypotesen – gäller att

$$\mathbb{E}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} | \mathbf{X}) = \mathbf{0}, \quad (\text{A.75})$$

$$\mathbb{V}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} | \mathbf{X}) = \sigma_U^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}', \quad (\text{A.76})$$

och

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) | \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma_U^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}'). \quad (\text{A.77})$$

Således gäller att

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\sigma_U^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}) | \mathbf{X} \sim \chi_Q^2. \quad (\text{A.78})$$

Sedan tidigare vet vi $(n - (K + 1))s_U^2/\sigma_U^2 \sim \chi_{n-(K+1)}^2$ och att $\hat{\mathbf{u}}$ och $\hat{\boldsymbol{\beta}}$ är oberoende. Således gäller att

$$\frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\sigma_U^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/Q}{s_U^2/\sigma_U^2} \Big| \mathbf{X} \sim F_{(Q, n-(K+1))}. \quad (\text{A.79})$$

Detta eftersom kvoten av två oberoende χ^2 -fördelade variabler delat med respektive fg är F -fördelad. Vi får alltså följande F -statistika:

$$F^{obs} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' [\sigma_U^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/Q. \quad (\text{A.80})$$

Således – under nollhypotesen – gäller att $F^{obs} \sim F_{(Q, n-(K+1))}$. Om $F^{obs} > F_{[Q, n-(K+1), \alpha]}$ förkastas H_0 , med signifikansnivå $100 \cdot \alpha$ procent.

Detta test kan även tas fram från det som kallas *likelihood ratio*-test. Ett sådant test bygger på att man skattar två modeller: en utan restriktioner (en lång modell) och en med restriktioner (kort modell). Det går att visa att ett sådant test ger test-statistikan

$$F^{obs} = \frac{(\text{RSS}_k - \text{RSS}_l)/Q}{\text{RSS}_l/(n - (K + 1))} = \frac{(R_l^2 - R_k^2)/Q}{(1 - R_l^2)/(n - (K + 1))}, \quad (\text{A.81})$$

där RSS_k och R_k^2 är residualskvadratsumma och förklaringsgrad för en kort modellen (med restriktioner), medan RSS_l och R_l^2 är residualskvadratsumma och förklaringsgrad för en lång modellen (utan restriktioner). I boken visas denna statistika i fallet när G variabler är med i den korta modellen och ytterligare $K - G$ variabler är med i den långa modellen. Antalet restriktioner blir då $Q = K - G$. Precis som var fallet med Wald-testet kan denna statistika användas även när vi har restriktioner som inte enbart bygger på exkluderande av variabler. Exempelvis, för nollhypotesen i ekvation (A.72), där $\beta_1 = 5$ och $\beta_2 = 1$, blir den långa modellen:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i. \quad (\text{A.82})$$

Under nollhypotesen har vi att

$$\begin{aligned} y_i &= \beta_0 + 5x_{1i} + 1x_{2i} + \beta_3 x_{3i} + u_i, \\ y_i - (5x_{1i} + 1x_{2i}) &= \beta_0 + \beta_3 x_{3i} + u_i, \\ \tilde{y}_i &= \beta_0 + \beta_3 x_{3i} + u_i. \end{aligned} \quad (\text{A.83})$$

Dvs. den korta regressionen erhålls genom att definiera en ny variabel $\tilde{y}_i = y_i - 5x_{1i} - x_{2i}$ och skatta en enkel linjär regression med \tilde{y}_i som beroende variabel och x_{3i} som oberoende. Vi kan sedan ta fram F -statistikan enligt ekvation (A.81) genom att ta fram RSS_k och RSS_l (alternativt R_k^2 och R_l^2) från dessa bägge regressioner.

För nollhypotesen i ekvation (A.73) där $\beta_1 - \beta_2 = 2$ har vi

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{1i} + (\beta_1 - 2)x_{2i} + \beta_3 x_{3i} + u_i, \\ y_i + 2x_{2i} &= \beta_0 + \beta_1(x_{1i} + x_{2i}) + \beta_3 x_{3i} + u_i, \\ \tilde{y}_i &= \beta_0 + \beta_1 \tilde{x}_i + \beta_3 x_{3i} + u_i. \end{aligned} \quad (\text{A.84})$$

där $\tilde{y}_i = y_i + 2x_{2i}$ och $\tilde{x}_i = x_{1i} + x_{2i}$. Genom att skapa dessa variabler och skatta denna begränsade (korta) regression och jämföra med den långa regressionen kan vi sedan ta fram F -statistikan enligt ekvation (A.81).

Hur förhåller sig då F -statistikan i ekvation (A.81) till F -statistikan i ekvation (A.80)? Svaret är att det är exakt samma statistika. Wald-testet och likelihood ratio-testet är därmed helt ekvivalenta och samma förkastelsegränser gäller därmed.

A.5 OLS med stokastiska kovariater

Vår populationsmodell specificerades i ekvation (A.30). Från detta tas ett slumpmässigt stickprov av data, $\{y_i, \mathbf{x}_i\}_{i=1}^n$. I föregående avsnitt har vi betraktat \mathbf{x}_i som fixt. Vad detta innebär är att varje gång ett nytt slumpmässigt stickprov dras så erhålls exakt samma matris \mathbf{X} och det enda som är slumpmässigt är det som förklarar \mathbf{y} som inte fångas upp i \mathbf{X} , dvs. \mathbf{u} . Detta är rimligt i de fall där \mathbf{X} bestäms i ett experiment, men blir orimligt i de fall vi har tillgång till observationsdata. Med observationsdata är det rimligt att tänka sig att om vi drar ett nytt stickprov så kommer värdena i \mathbf{X} att bli annorlunda och vi måste därför tänka på både \mathbf{X} och \mathbf{u} som stokastiska.

Ofta ser vi på populationsregressionslinjen, $\mathbf{x}'\boldsymbol{\beta}$ som en approximation av det betingade väntevärdet, $E(Y|\mathbf{x})$. En fråga vi kan ställa oss är om det finns något annat linjärt samband som bättre approximerar $E(Y|\mathbf{x})$? Låt $\mathbf{x}'\mathbf{b}$ vara ett sådant linjärt samband. Notera att vi alltid kan göra följande dekomponering:

$$Y = E(Y|\mathbf{x}) + V, \quad (\text{A.85})$$

så att $E(V|\mathbf{x}) = 0$. Vi har att kvadratavvikelsen mellan Y och $\mathbf{x}'\mathbf{b}$ är

$$\begin{aligned} (Y - \mathbf{x}'\mathbf{b})^2 &= (E(Y|\mathbf{x}) + V - \mathbf{x}'\mathbf{b})^2 \\ &= (E(Y|\mathbf{x}) - \mathbf{x}'\mathbf{b})^2 + V^2 + 2(E(Y|\mathbf{x}) - \mathbf{x}'\mathbf{b}) \cdot V. \end{aligned} \quad (\text{A.86})$$

Eftersom $E(V|\mathbf{x}) = 0$ kan den förväntade kvadratavvikelsen mellan Y och $\mathbf{x}'\mathbf{b}$ skrivas

$$E(Y - \mathbf{x}'\mathbf{b})^2 = E((E(Y|\mathbf{x}) - \mathbf{x}'\mathbf{b})^2) + E(V^2). \quad (\text{A.87})$$

Enligt ekvation (A.26) minimeras $E(Y - \mathbf{x}'\mathbf{b})^2$ då $\mathbf{b} = \boldsymbol{\beta}$. Då $E(V^2)$ inte beror på \mathbf{b} så innebär det att även $E((E(Y|\mathbf{x}) - \mathbf{x}'\mathbf{b})^2)$ minimeras då $\mathbf{b} = \boldsymbol{\beta}$. Detta innebär således att $\mathbf{x}'\boldsymbol{\beta}$ är den bästa linjära approximationen av det betingade väntevärdet i populationen.

A.5.1 Asymptotiska egenskaper för OLS-estimatorn

Vi studerar nu de asymptotiska egenskaperna hos $\hat{\boldsymbol{\beta}}$ (dvs. egenskaper för stora stickprov). En fråga vi kan ställa oss är om $\hat{\boldsymbol{\beta}}$ är en konsistent estimator av $\boldsymbol{\beta}$. Konsistens innebär att $\hat{\boldsymbol{\beta}}$ konvergerar i sannolikhet till $\boldsymbol{\beta}$, vilket vi skriver som $\text{plim}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. För att visa konsistens använder vi oss av Slutskys teorem som implicerar att för konstanter a, b och slumpvariabler X, Y gäller att $\text{plim}(a + bX/Y) = a + b \text{plim}(X)/\text{plim}(Y)$. Tillsammans med ekvation (A.39)

där vi slog fast att $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ har vi

$$\begin{aligned}\text{plim}(\hat{\beta}) &= \text{plim}(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\ &= \beta + (\text{plim}(\mathbf{X}'\mathbf{X}))^{-1} \text{plim}(\mathbf{X}'\mathbf{u}) \\ &= \beta + (\text{plim}(\mathbf{X}'\mathbf{X}/n))^{-1} \text{plim}(\mathbf{X}'\mathbf{u}/n),\end{aligned}\tag{A.88}$$

där det sista likhetstecknet kommer att vi både multiplicerar och dividerar med $1/n$. För slumpmässigt draget stickprov har vi att $\text{plim}(\mathbf{X}'\mathbf{X}/n) = E(\mathbf{x}\mathbf{x}')$ och $\text{plim}(\mathbf{X}'\mathbf{u}/n) = E(\mathbf{x}'U)$. Från ekvation (A.31) visade vi att $E(\mathbf{x}U) = \mathbf{0}$ vilket innebär att $\text{plim}(\hat{\beta}) = \beta$. OLS-estimatorn är alltså en konsistent estimator!

Således, för ett oändligt stickprov så konvergerar $\hat{\beta}$ till β och fördelningen för $\hat{\beta}$ är en punkt. Men då vi i praktiken har ett ändligt stickprov måste vi – för inferens – veta fördelningen när n är ändlig, men stor. Eftersom fördelningen tas fram för fallet när n går mot oändligheten används ett trick som består i att man normerar estimatorn med den ”hastighet” som en estimator konvergerar. Vi är alltså intresserade av fördelningen för $\sqrt{n}(\hat{\beta} - \beta)$. Vi har:

$$\begin{aligned}\sqrt{n}(\hat{\beta} - \beta) &= \sqrt{n}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}) \\ &= (\mathbf{X}'\mathbf{X}/n)^{-1}\mathbf{X}'\mathbf{u}/\sqrt{n}.\end{aligned}\tag{A.89}$$

Eftersom $E(\mathbf{x}U) = 0$ gäller att under slumpmässig stickprovsdragning är $E(\mathbf{x}_i U_i) = 0$ och $E(\mathbf{X}'\mathbf{u}/\sqrt{n}) = \mathbf{0}$. Vidare gäller att $V(\mathbf{x}_i U_i) = E(U^2 \mathbf{x}\mathbf{x}')$. Centrala gränsvärdesatsen implicerar då att

$$\mathbf{X}'\mathbf{u}/\sqrt{n} \xrightarrow{d} \mathcal{N}(\mathbf{0}, E(U^2 \mathbf{x}\mathbf{x}')), \tag{A.90}$$

där \xrightarrow{d} innebär att det till vänster *konvergerar i fördelning* (dvs. följer samma fördelning asymptotiskt) till det som står till höger. Vi kan nu lägga till och dra ifrån $(E(\mathbf{x}\mathbf{x}'))^{-1}\mathbf{X}'\mathbf{u}/\sqrt{n}$ från ekvation (A.89):

$$\sqrt{n}(\hat{\beta} - \beta) = (E(\mathbf{x}\mathbf{x}'))^{-1}\mathbf{X}'\mathbf{u}/\sqrt{n} + \left((\mathbf{X}'\mathbf{X}/n)^{-1} - (E(\mathbf{x}\mathbf{x}'))^{-1}\right)\mathbf{X}'\mathbf{u}/\sqrt{n}. \tag{A.91}$$

Då $\text{plim}((\mathbf{X}'\mathbf{X}/n)^{-1}) = (\text{plim}(\mathbf{X}'\mathbf{X}/n))^{-1} = (E(\mathbf{x}\mathbf{x}'))^{-1}$, samtidigt som $\mathbf{X}'\mathbf{u}/\sqrt{n}$ inte kan ”explodera” så att den blir oändligt stor (eftersom väntevärde är $\mathbf{0}$ med finit varians) så kommer den andra termen *inte* påverka fördelningen av $\sqrt{n}(\hat{\beta} - \beta)$. Vidare eftersom $E(\mathbf{x}\mathbf{x}')$ är symmetrisk och konstant innebär detta att vi har

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (E(\mathbf{x}\mathbf{x}'))^{-1} E(U^2 \mathbf{x}\mathbf{x}') (E(\mathbf{x}\mathbf{x}'))^{-1}). \tag{A.92}$$

I praktiken innebär detta att för ett stort, men finit, stickprov har vi den approximativa variansen

$$V(\hat{\beta}) \approx (E(\mathbf{x}\mathbf{x}'))^{-1} E(U^2 \mathbf{x}\mathbf{x}') (E(\mathbf{x}\mathbf{x}'))^{-1} / n. \tag{A.93}$$

Om feltermen är homoskedastisk gäller att

$$E(U^2 \mathbf{x}\mathbf{x}') = E(U^2) E(\mathbf{x}\mathbf{x}') = \sigma_U^2 E(\mathbf{x}\mathbf{x}'). \tag{A.94}$$

I så fall förenklas uttrycket för variansen till

$$V(\hat{\beta}) \approx \sigma_U^2 \frac{E(\mathbf{x}\mathbf{x}')^{-1}}{n}. \quad (\text{A.95})$$

Med fixa kovariater gällde att $E(s_U^2) = \sigma_U^2$. Detta gäller inte längre med stokastiska variabler, men däremot är s_U^2 en konsistent skattare: $\text{plim}(s_U^2) = \sigma_U^2$. Om vi nu skattar $E(\mathbf{x}\mathbf{x}')$ med $(\mathbf{X}'\mathbf{X})/n$ så blir den naturliga skattaren av koefficienternas varians därmed, givet att feltermen är homoskedastisk, samma som med fixa kovariater enligt ekvation (A.59):

$$\hat{V}(\hat{\beta}) = s_U^2 \frac{(\mathbf{X}'\mathbf{X}/n)^{-1}}{n} = s_U^2 (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{A.96})$$

Med stokastiska variabler gäller att t -statistikan i ekvation (A.63) enbart har en känd fördelning asymptotiskt, i vilket fall den följer en standardnormalfördelning. Även F -testet för sammansatt hypotesprövning (avsnitt A.4.1.2) kan användas, men är bara korrekt asymptotiskt. Den asymptotiska fördelningen för statistikan är $F^{obs} \sim F_{Q,\infty}$ och det gäller då att $QF^{obs} \sim \chi_Q^2$. Därför är det asymptotiska F -testet ekvivalent med ett χ^2 -test.

Med observationsdata är homoskedasticitetsantagandet i ekvation (A.94) ofta inte uppfyllt. Ett sätt att tänka på detta är utgå från dekomponeringen i ekvation (A.85), dvs. $Y = E(Y|\mathbf{x}) + V$. Även i de fall då $V(Y|\mathbf{x}) = \sigma_V^2$ så kommer vi ha heteroskedasticitet om $E(Y|\mathbf{x})$ inte är linjär. Vi har $Y = \mathbf{x}'\beta + U$ vilket innebär att

$$U = E(Y|\mathbf{x}) + V - \mathbf{x}'\beta. \quad (\text{A.97})$$

Variansen av U är då

$$V(U) = \sigma_V^2 + V(E(Y|\mathbf{x}) - \mathbf{x}'\beta). \quad (\text{A.98})$$

Då den andra termen är en funktion av \mathbf{x} (så länge vi har icke-linjäritet, $E(Y|\mathbf{x}) \neq \mathbf{x}'\beta$) innebär det att dekomponeringen i ekvation (A.94) inte kan göras och vi har heteroskedasticitet.

Det är dock lätt att hantera heteroskedasticitet. Så länge feltermerna är $E(u_i u_j \mathbf{x}_i \mathbf{x}_j) = 0$ för alla $i \neq j$ gäller,⁴ oavsett om det är homoskedasticitet eller heteroskedasticitet, att

$$\text{plim} \left(\sum_{i=1}^n u_i^2 \mathbf{x}_i \mathbf{x}_i' / n \right) = E(U^2 \mathbf{x} \mathbf{x}'). \quad (\text{A.99})$$

⁴Notera att vi i boken säger att feltermerna betingat på kovariater ska vara oberoende. Det är väldigt svårt att hitta ett exempel där detta nödvändiga villkor inte håller (dvs., $E(u_i u_j | \mathbf{x}_i \mathbf{x}_j) \neq 0$), medan $E(u_i u_j \mathbf{x}_i \mathbf{x}_j) = 0$. Således i praktiken krävs oberoende. Notera också att i kapitel 11 beskriver vi en HAC estimator för att skatta standardfelet när $E(u_i u_j \mathbf{x}_i \mathbf{x}_j) \neq 0$. Detta är möjligt om vi har en idé om hur beroendet ser ut. Med tidseriedata kan vi anta att beroendet bara finns mellan närliggande observationer av distans p tidpunkter eller mindre.

En naturlig variansskattare skulle vara $(\mathbf{X}'\mathbf{X}/n)^{-1} (\sum_{i=1}^n u_i^2 \mathbf{x}\mathbf{x}'/n) (\mathbf{X}'\mathbf{X}/n)^{-1}/n$. Detta är en konsistent skattare eftersom

$$\begin{aligned} & \text{plim} \left((\mathbf{X}'\mathbf{X}/n)^{-1} \left(\sum_{i=1}^n u_i^2 \mathbf{x}_i \mathbf{x}_i' / n \right) (\mathbf{X}'\mathbf{X}/n)^{-1} \right) \\ &= (\text{plim}(\mathbf{X}'\mathbf{X}/n))^{-1} \text{plim} \left(\sum_{i=1}^n u_i^2 \mathbf{x}_i \mathbf{x}_i' / n \right) (\text{plim}(\mathbf{X}'\mathbf{X}/n))^{-1} \\ &= (\text{E}(\mathbf{x}\mathbf{x}'))^{-1} \text{E}(U^2 \mathbf{x}\mathbf{x}') (\text{E}(\mathbf{x}\mathbf{x}'))^{-1} \\ &= n \text{V}(\hat{\boldsymbol{\beta}}). \end{aligned} \quad (\text{A.100})$$

I stället för att dela med n brukar man dock göra en frihetsgradsjustering och dela med $n - (K + 1)$. Detta påverkar inte konsistensen för finit antal variabler ($K < \infty$), men är i linje med den ”vanliga” skattade kovariansmatrisen i ekvation (A.59) där frihetsgradsjusteringen görs för s_U^2 . Denna ”robusta” kovariansmatris är alltså:

$$\hat{\mathbf{V}}^r(\hat{\boldsymbol{\beta}}) = \frac{n}{n - (K + 1)} (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 \mathbf{x}_i \mathbf{x}_i' \right) (\mathbf{X}'\mathbf{X})^{-1}. \quad (\text{A.101})$$

Denna estimator kallas för robust, ”sandwich”, eller Eicker-Huber-White (EHW) efter (Eicker, 1967; Huber, 1967; White, 1980). Med små stickprov kan denna robusta estimator ge för små variansskattningar. Innebörden av detta är att vi drar för starka slutsatser. Det finns en dock andra estimatorer som föreslagits som har bättre egenskaper vid små stickprov (se MacKinnon 2013 för en genomgång).

Låt $\hat{\mathbf{V}}^r(\hat{\boldsymbol{\beta}})^{jj}$ vara diagonalelement j i den robusta kovariansmatrisen i ekvation (A.101) och definiera det robusta standardfelet för $\hat{\beta}_j$ som

$$s_{\hat{\beta}_j}^r = \sqrt{\hat{\mathbf{V}}^r(\hat{\boldsymbol{\beta}})^{jj}}. \quad (\text{A.102})$$

Den robusta t -statistikan blir

$$T_r^{obs} = \frac{\hat{\beta}_k - \beta_k}{s_{\hat{\beta}_k}^r}, \quad (\text{A.103})$$

och den robusta F -statistikan kan skrivas

$$F_r^{obs} = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})' \left[\mathbf{R}\hat{\mathbf{V}}^r(\hat{\boldsymbol{\beta}})\mathbf{R}' \right]^{-1} (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/Q. \quad (\text{A.104})$$

Asymptotiskt gäller att T_r^{obs} är normalfördelad, medan F_r^{obs} är F -fördelad med Q och ∞ frihetsgrader, vilket innebär att QF_r^{obs} är χ^2 -fördelad med Q frihetsgrader. Vid moderata stickprov (n mindre än ca 100) har det visat sig i Monte Carlo-simuleringar att det är bättre att använda F -fördelning (med Q och $n - (K + 1)$ frihetsgrader) vid inferens och detta även om feltermen inte är normalfördelad.

Notera att F -testet baserat på residuallkvadratsummor eller R^2 i ekvation (A.81) inte lägre är korrekt att använda. Dock är R^2 ett valitt mått för att beskriva hur bra modellen anpassar

till data även under heteroskedasticitet: R^2 är en konsistent estimator av korrelationen mellan $E(Y|\mathbf{x})$ och Y i populationen, vilken kan skrivas $\rho^2 = 1 - \sigma_U^2 / \sigma_Y^2$. Detta då enbart de obetingade varianserna av U och Y förekommer i detta mått.

A.5.2 Deltametoden

För att kunna genomföra statistik inferens måste vi känna fördelningen på en statistika. I regressionsfallet visade vi exempelvis i avsnitt A.4 hur vi kan ta fram fördelningen av $\hat{\beta}$. Ibland är vi intresserade av fördelningen av *funktioner* av variabler med känd fördelning. Säg exempelvis att X är normalfördelad med medelvärde μ_X och varians σ_X^2 . En ny variabel $Y = a + bX$, där a och b är konstanter, är en *linjär transformation* av X . I det fallet känner vi fördelningen direkt då vi vet att Y är normalfördelad med medelvärde $\mu_Y = a + b\mu_X$ och varians $b^2\sigma_X^2$.

Ibland är vi istället intresserade av en *ickelinjär transformation* sådan att $Y = f(X)$, där $f(\cdot)$ är en icke-linjär funktion. I det fallet är det inte lika enkelt att ta fram fördelningen på Y . Det finns metoder, såsom faltning och variabeltransformation, som ibland gör det möjligt att lösa sådana problem. I många fall är det dock tillräckligt att känna den asymptotiska fördelningen hos en funktion av en variabel, i vilket fall *deltametoden* kan användas.

Deltametoden, eller vad som ibland kallas Gauss-approximation, säger att om vi har en variabel X som är asymptotiskt normalfördelad är $Y = f(X)$ approximativt asymptotiskt normalfördelad med medelvärde $Y = f(\mu_X)$ och med varians $\sigma_X^2 \cdot f'(\mu_X)^2$, där $f'(\mu_X)$ är första ordningens derivata av $f(\cdot)$ med avseende på X och utvärderad vid μ_X .

Resultatet kommer från att vi gör en första ordningens Taylorutveckling runt μ_X . En Taylorutveckling är en lokal approximation av en funktion runt en viss punkt. För en funktion $f(X)$ gäller att Taylorutvecklingen runt punkten $X = \mu_X$ är

$$\begin{aligned} f(X) &= f(\mu_X) + \frac{f'(\mu_X)}{1!}(X - \mu_X) + \frac{f''(\mu_X)}{2!}(X - \mu_X)^2 + \frac{f'''(\mu_X)}{3!}(X - \mu_X)^3 + \dots \\ &= \sum_{i=0}^{\infty} \frac{f^{(i)}(\mu_X)}{i!}(X - \mu_X)^i. \end{aligned} \quad (\text{A.105})$$

Poängen med denna utveckling är att funktionen $f(X)$ kan vara svår att hantera, medan Taylorutvecklingen till höger om likhetstecknet är enklare att arbeta med. Det är viktigt att påpeka att Taylorutvecklingen bara ger en bra approximation av funktionen $f(X)$ när X ligger i närheten av μ_X . Denna Taylorutveckling innehåller oändligt många termer, men i praktiken kan man få en bra approximation med några fåtal termer. En första ordningens Taylorutveckling har bara med den första derivata-termen och är då

$$f(X) \approx f(\mu_X) + f'(\mu_X)(X - \mu_X). \quad (\text{A.106})$$

Variansen av $f(X)$ när X är nära μ_X blir då

$$\begin{aligned} E((f(X) - f(\mu_X))^2) &\approx E((f(\mu_X) + f'(\mu_X)(X - \mu_X) - f(\mu_X))^2) \\ &= E((f'(\mu_X)(X - \mu_X))^2) \\ &= (f'(\mu_X))^2 E((X - \mu_X)^2) \\ &= (f'(\mu_X))^2 \sigma_X^2. \end{aligned} \quad (\text{A.107})$$

Som ett exempel, antag att vi har ett obundet slumpmässig urval $\{x_i\}_{i=1}^n$ från den stokastiska variabeln X med medelvärde μ_X och standardavvikelse σ_X . Vi skattar medelvärdet med $\bar{x} = n^{-1} \sum_{i=1}^n x_i$. Vi vet att $\sqrt{n}\bar{x}$ är asymptotiskt normalfördelad med medelvärde $\sqrt{n}\mu_X$ och standardfel σ_X .

För att exemplifiera, låt $\theta = \exp(\mu_X)$ som vi skattar med $\hat{\theta} = \exp(\bar{x})$. Här är således $f(X) = \exp(\bar{x}) = \hat{\theta}$. Eftersom derivatan av $\exp(\mu_X)$ är funktionen själv, $\exp(\mu_X)$, har vi att första ordningens Taylorutveckling runt μ_X är

$$\hat{\theta} \approx \theta + \exp(\mu_X)(\bar{x} - \mu_X), \quad (\text{A.108})$$

Vi har nu att $\hat{\theta} - \theta \approx \exp(\mu_X)(\bar{x} - \mu_X)$. Om vi kvadrerar båda sidor och tar väntevärdet i enlighet med ekvation (A.107) får vi:

$$V(\hat{\theta}) = E\left((\hat{\theta} - \theta)^2\right) \approx (\exp(\mu_X))^2 E((\bar{x} - \mu_X)^2) \quad (\text{A.109})$$

$$= (\exp(\mu_X))^2 \sigma_X^2 / n, \quad (\text{A.110})$$

då $V(\bar{x}) = \sigma_X^2 / n$. Denna varians är bara en bra approximation om \bar{x} är nära μ_X . Här kommer nyckelinsikten: Då \bar{x} konvergerar i sannolikhet till μ_X får vi även en skattare för variansen. Således vid stora stickprov kan vi skatta variansen för $\hat{\theta}$ enligt

$$\hat{V}(\hat{\theta}) = (\exp(\bar{x}))^2 s_X^2 / n. \quad (\text{A.111})$$

Deltametoden är därmed i allmänhet bara användbar för stora stickprov. Det gäller då också att

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}\left(0, (\exp(\bar{x}))^2 s_X^2 / n\right). \quad (\text{A.112})$$

För att, exempelvis, testa $H_0 : \theta = 1$ mot $H_0 : \theta \neq 1$ kan vi skapa en Z -statistika på följande sätt

$$Z^{obs} = \frac{\exp(\bar{x}) - 1}{\exp(\bar{x}) \sqrt{s_X^2 / n}}. \quad (\text{A.113})$$

Vi kan använda deltametoden för regression enligt samma principer. Variansen för en funktion av en regressionskoefficient $h(\hat{\beta}_j)$ i en multipel linjär regressionsmodell är enligt ekvation (A.107):

$$V(h(\hat{\beta}_j)) \approx \left(\frac{\partial h(\beta_j)}{\partial \beta_j} \right)^2 V(\hat{\beta}_j). \quad (\text{A.114})$$

Deltametoden går även att använda för multivariata problem. Säg exempelvis att vi vill testa nollhypotesen $H_0 : \mathbf{h}(\boldsymbol{\beta}) - \mathbf{r} = \mathbf{0}$. Vi utgår från principen för Wald-testet som beskrevs ovan, men där vi ersätter $\mathbf{R}\boldsymbol{\beta} - \mathbf{r}$ med $\mathbf{c}(\boldsymbol{\beta}) = \mathbf{0}$ där $\mathbf{c}(\boldsymbol{\beta})$ är:

$$\mathbf{c}(\boldsymbol{\beta}) = \begin{bmatrix} c_1(\boldsymbol{\beta}) \\ c_2(\boldsymbol{\beta}) \\ \vdots \\ c_Q(\boldsymbol{\beta}) \end{bmatrix} = \begin{bmatrix} h_1(\boldsymbol{\beta}) - r_1 \\ h_2(\boldsymbol{\beta}) - r_2 \\ \vdots \\ h_Q(\boldsymbol{\beta}) - r_Q \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (\text{A.115})$$

Variansen av $\mathbf{c}(\hat{\boldsymbol{\beta}})$ erhålls med den multivariata analogen till ekvation (A.107) så att

$$\mathbf{V}(\mathbf{c}(\hat{\boldsymbol{\beta}})) = \mathbf{C}(\boldsymbol{\beta})' \mathbf{V}(\hat{\boldsymbol{\beta}}) \mathbf{C}(\boldsymbol{\beta}), \quad (\text{A.116})$$

där

$$\mathbf{C}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial c_1(\boldsymbol{\beta})}{\partial \beta} & \frac{\partial c_2(\boldsymbol{\beta})}{\partial \beta} & \cdots & \frac{\partial c_Q(\boldsymbol{\beta})}{\partial \beta} \end{bmatrix} = \begin{bmatrix} \frac{\partial c_1(\boldsymbol{\beta})}{\partial \beta_0} & \frac{\partial c_2(\boldsymbol{\beta})}{\partial \beta_0} & \cdots & \frac{\partial c_Q(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial c_1(\boldsymbol{\beta})}{\partial \beta_1} & \frac{\partial c_2(\boldsymbol{\beta})}{\partial \beta_1} & \cdots & \frac{\partial c_Q(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial c_1(\boldsymbol{\beta})}{\partial \beta_K} & \frac{\partial c_2(\boldsymbol{\beta})}{\partial \beta_K} & \cdots & \frac{\partial c_Q(\boldsymbol{\beta})}{\partial \beta_K} \end{bmatrix}. \quad (\text{A.117})$$

En F -statistika utifrån Wald-testet kan nu skapas genom att ersätta $\mathbf{V}(\mathbf{c}(\hat{\boldsymbol{\beta}}))$ med

$$\hat{\mathbf{V}}(\mathbf{c}(\hat{\boldsymbol{\beta}})) = \mathbf{C}(\hat{\boldsymbol{\beta}})' \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) \mathbf{C}(\hat{\boldsymbol{\beta}}). \quad (\text{A.118})$$

Följande statistika erhålls då:

$$F^{obs} = \mathbf{c}(\hat{\boldsymbol{\beta}})' \hat{\mathbf{V}}(\mathbf{c}(\hat{\boldsymbol{\beta}}))^{-1} \mathbf{c}(\hat{\boldsymbol{\beta}}) / Q. \quad (\text{A.119})$$

Som tidigare gäller att $F^{obs} \stackrel{a}{\sim} F_{Q,\infty}$ och $QF^{obs} \stackrel{a}{\sim} \chi_Q^2$.

Vi kan illustrera hur detta test genomförs med ett enkelt exempel. Antag att vi har följande linjära modell:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \quad (\text{A.120})$$

där vi är intresserade av nollhypotesen att $\beta_1/\beta_2 = 2$. Detta är en icke-linjär hypotes som inte går att skriva på formen $\mathbf{R}\boldsymbol{\beta} - \mathbf{r}$, så det vanliga Wald-testet går inte att genomföra. Vi kan dock använda deltametoden. Vi har att $c(\boldsymbol{\beta}) = \beta_1/\beta_2 - 2 = 0$, vilket ger

$$\mathbf{C}(\boldsymbol{\beta}) = \begin{bmatrix} \frac{\partial c(\boldsymbol{\beta})}{\partial \beta_0} \\ \frac{\partial c(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial c(\boldsymbol{\beta})}{\partial \beta_2} \end{bmatrix} = \begin{bmatrix} 0 \\ \frac{1}{\beta_2} \\ -\frac{\beta_1}{\beta_2^2} \end{bmatrix} \quad (\text{A.121})$$

Vi får nu följande F -statistika:

$$F^{obs} = (\beta_1/\beta_2 - 2)^2 (\mathbf{C}(\hat{\boldsymbol{\beta}})' \hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}) \mathbf{C}(\hat{\boldsymbol{\beta}}))^{-1}, \quad (\text{A.122})$$

där antingen den "vanliga" eller den robusta kovariansmatrisen kan användas för $\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}})$. Vi har bara en restriktion vilket innebär att $F^{obs} \stackrel{a}{\sim} F_{1,\infty}$ och $F^{obs} \stackrel{a}{\sim} \chi_1^2$.

Prediktionsintervall för BNP

I avsnitt 11.4 i boken diskuterade vi hur vi kunde ta fram ett $100 \cdot (1 - \alpha)\%$ -igt prediktionsintervall för BNP när vi hade skattat en modell med logaritmerade differenser med prediktionen vid $T + 1$ lika med:

$$\hat{y}_{T+1} = \exp(\ln y_T + \hat{\beta}_0 + \hat{\beta}_1(\ln y_T - \ln y_{T-1}) + \hat{\beta}_2(\ln y_{T-1} - \ln y_{T-2}) + s_U^2/2). \quad (\text{A.123})$$

För att skatta variansen, notera först att prediktionen kommer från modellen

$$\ln Y_t - \ln Y_{t-1} = \beta_0 + \beta_1(\ln Y_{t-1} - \ln Y_{t-2}) + \beta_2(\ln Y_{t-2} - \ln Y_{t-3}) + U_t. \quad (\text{A.124})$$

och att vi för perioden $T + 1$ kan formulera prognosen en period framåt som

$$\ln Y_{T+1} = \ln Y_T + \beta_0 + \beta_1(\ln Y_T - \ln Y_{T-1}) + \beta_2(\ln Y_{T-1} - \ln Y_{T-2}) + U_{T+1}. \quad (\text{A.125})$$

För att ta fram variansen använder vi först *lagen om total varians* (se kapitel 5). I detta fall har vi

$$V(\hat{y}_{T+1}) = V(E(Y_{T+1}|\hat{\theta})) + E(V(Y_{T+1}|\hat{\theta})). \quad (\text{A.126})$$

där $\hat{\theta} = [\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2 \quad s_U^2]'$. Här är

$$E(Y_{T+1}|Y_T, \hat{\theta}) = \exp(\ln y_T + \hat{\beta}_0 + \hat{\beta}_1(\ln y_T - \ln y_{T-1}) + \hat{\beta}_2(\ln y_{T-1} - \ln y_{T-2}) + s_U^2/2) \quad (\text{A.127})$$

och

$$\begin{aligned} \hat{V}(Y_{T+1}|Y_T, \hat{\theta}) = \\ (\exp(s_U^2) - 1) \exp\left(2(\ln y_T + \hat{\beta}_0 + \hat{\beta}_1(\ln y_T - \ln y_{T-1}) + \hat{\beta}_2(\ln y_{T-1} - \ln y_{T-2})) + s_U^2\right), \end{aligned} \quad (\text{A.128})$$

vilket följer av att $V(\exp(\mu) \exp(U)) = (\exp(\sigma_U^2) - 1) \exp(2\mu + \sigma_U^2)$ om U är normalfördelad.

Eftersom $\hat{\beta}$ och s_U^2 är konsistenta estimatorer för β och σ_U^2 får vi att för ett stort stickprov:

$$E(V(Y_{T+1}|Y_T, \hat{\theta})) \approx \hat{V}(Y_{T+1}|Y_T, \hat{\theta}). \quad (\text{A.129})$$

För att nu ta fram $V(E(Y_{T+1}|Y_T, \hat{\theta}))$ använder vi deltametoden. Vi har att $c(\hat{\theta}) = E(Y_{T+1}|Y_T, \hat{\theta})$ och därmed

$$C(\hat{\theta}) = \begin{bmatrix} \frac{\partial c(\hat{\theta})}{\partial \hat{\beta}_0} \\ \frac{\partial c(\hat{\theta})}{\partial \hat{\beta}_1} \\ \frac{\partial c(\hat{\theta})}{\partial \hat{\beta}_2} \\ \frac{\partial c(\hat{\theta})}{\partial s_U^2} \end{bmatrix} = \begin{bmatrix} \hat{y}_{T+1} \\ \hat{y}_{T+1}(\ln y_T - \ln y_{T-1}) \\ \hat{y}_{T+1}(\ln y_{T-1} - \ln y_{T-2}) \\ \hat{y}_{T+1}/2 \end{bmatrix}. \quad (\text{A.130})$$

Om $U \sim \mathcal{N}(0, \sigma_U^2)$ gäller att en naturlig skattare av $V(\hat{\boldsymbol{\theta}})$ är

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} s^2 (\mathbf{X}'\mathbf{X})^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{2s^4}{T} \end{bmatrix}, \quad (\text{A.131})$$

där $\mathbf{X} = [\mathbf{x}'_3 \ \mathbf{x}'_4 \ \dots \ \mathbf{x}'_T]'$, med $\mathbf{x}_t = [1 \ \ln(Y_t) - \ln(Y_{t-1}) \ \ln(Y_{t-1}) - \ln(Y_{t-2})]'$, för $t \geq 3$. Ekvation (A.116) ger då

$$\hat{V}(E(Y_{T+1}|\hat{\boldsymbol{\theta}})) = \mathbf{C}(\hat{\boldsymbol{\theta}})' \hat{V}(\hat{\boldsymbol{\theta}}) \mathbf{C}(\hat{\boldsymbol{\theta}}). \quad (\text{A.132})$$

Den betingade variansen kan således skattas enligt

$$\begin{aligned} \hat{V}(\hat{Y}_{T+1}|y_T) &= \mathbf{C}(\hat{\boldsymbol{\theta}})' \hat{V}(\hat{\boldsymbol{\theta}}) \mathbf{C}(\hat{\boldsymbol{\theta}}) + \\ &\quad (\exp(s_U^2) - 1) \exp\left(2(\ln y_T + \hat{\beta}_0 + \hat{\beta}_1(\ln y_T - \ln y_{T-1}) + \hat{\beta}_2(\ln y_{T-1} - \ln y_{T-2})) + s_U^2\right). \end{aligned} \quad (\text{A.133})$$

Notera att första delen i denna ekvation inte överensstämmer med den som ges i boken, där det smög sig in några tryckfel och där korrelationen mellan de två termerna $(\ln(Y_t) - \ln(Y_{t-1}))$ och $(\ln(Y_{t-1}) - \ln(Y_{t-2}))$ bortsågs från. Således är förslaget om estimator av den betingade variansen i boken felaktig.

Det bör också noteras att den enkla (första ordningens) deltametod sannolikt inte fungerar speciellt bra i detta fall. Ursprungsserien, $\{Y_t\}_{t=-\infty}^{\infty}$, är inte stationär. En liten förändring från förväntat värde i den logaritmerade tidserien kommer exponentieringen ge en stor förändring i nivån, vilket gör att approximationen med deltametoden fungerar sämre. Sannolikt skulle en andra ordningens approximation fungera bättre. En andra ordningens deltametod är dock komplicerad och inget vi går igenom i denna bilaga.

A.6 Regressionsanatomi

Vi gör här uppdelning mellan en lång och kort regression. Den långa regressionen med K oberoende variabler kan vi för ett stickprov av data skriva som

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^l + \mathbf{u}^l. \quad (\text{A.134})$$

\mathbf{X} är alltså en matris av dimension $n \times (K + 1)$ (då vi har med en kolumn med ettor för att modellen ska ha ett intercept). Den korta regressionen innehåller bara ett subset av X -variablerna, vilka vi samlar i en matris \mathbf{X}_A av dimension $n \times (G + 1)$ matris. Övriga variabler som ingår i \mathbf{X} men ej i \mathbf{X}_A samlar vi i en matris \mathbf{X}_B av dimension $n \times (K - G)$. Vi har alltså:

$$\mathbf{X}_A = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{G1} \\ 1 & x_{12} & x_{22} & \dots & x_{G2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{Gn} \end{bmatrix}, \mathbf{X}_B = \begin{bmatrix} x_{G+1,1} & x_{G+2,1} & \dots & x_{Kn} \\ x_{G+1,2} & x_{G+2,2} & \dots & x_{Kn} \\ \vdots & \vdots & \ddots & \vdots \\ x_{G+1,n} & x_{G+2,n} & \dots & x_{Kn} \end{bmatrix}, \mathbf{X} = [\mathbf{X}_A \ \mathbf{X}_B]. \quad (\text{A.135})$$

Vi kan då skriva om den långa regressionen i ekvation (A.134) som

$$\mathbf{y} = \mathbf{X}_A \boldsymbol{\beta}_A^l + \mathbf{X}_B \boldsymbol{\beta}_B^l + \mathbf{u}^l. \quad (\text{A.136})$$

Den korta regressionen kan skrivas som

$$\mathbf{y} = \mathbf{X}_A \boldsymbol{\beta}_A^k + \mathbf{u}^k. \quad (\text{A.137})$$

Vi kan skriva relationen mellan OLS-estimatorerna $\hat{\boldsymbol{\beta}}_1^k$ och $\hat{\boldsymbol{\beta}}_1^l$ som

$$\hat{\boldsymbol{\beta}}_A^k = \hat{\boldsymbol{\beta}}_A^l + \mathbf{C}, \quad (\text{A.138})$$

där vårt mål är att lista ut hur \mathbf{C} ser ut. Definitionen av residualerna i den långa regressionen är

$$\hat{\mathbf{u}}^l = \mathbf{y} - \mathbf{X}_A \hat{\boldsymbol{\beta}}_A^l - \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l. \quad (\text{A.139})$$

Vi kan multiplicera bägge sidor med \mathbf{X}'_A :

$$\mathbf{X}'_A \hat{\mathbf{u}}^l = \mathbf{X}'_A \mathbf{y} - \mathbf{X}'_A \mathbf{X}_A \hat{\boldsymbol{\beta}}_A^l - \mathbf{X}'_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l. \quad (\text{A.140})$$

Vi vet från ekvation (A.17) att $\mathbf{X}' \hat{\mathbf{u}} = \mathbf{0}$. Detta innebär att $\mathbf{X}'_A \hat{\mathbf{u}}^l = \mathbf{0}$ och att $\mathbf{X}'_B \hat{\mathbf{u}}^l = \mathbf{0}$. Löser vi för $\hat{\boldsymbol{\beta}}_A^l$ får vi då

$$\begin{aligned} \hat{\boldsymbol{\beta}}_A^l &= (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{y} - (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l \\ &= \hat{\boldsymbol{\beta}}_A^k - (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l. \end{aligned} \quad (\text{A.141})$$

Det gäller alltså att $\mathbf{C} = (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l$, så att

$$\hat{\boldsymbol{\beta}}_A^k = \hat{\boldsymbol{\beta}}_A^l + (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l. \quad (\text{A.142})$$

För var och en av kovariaterna som finns i \mathbf{X}_B -matrisen kan följande regressioner skattas med OLS:

$$\begin{aligned} X_j &= \gamma_0^j + \gamma_1^j X_1 + \gamma_2^j X_2 + \dots + \gamma_G^j X_G + V^j \\ &= \mathbf{X}_B \boldsymbol{\gamma}^j + V^j \quad j = G+1, G+2, \dots, K. \end{aligned} \quad (\text{A.143})$$

För alla dessa $K - G$ regressioner kan koefficienterna skrivas i en matris av dimension $(G+1) \times (K - G)$:

$$(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \mathbf{X}_B = \begin{bmatrix} \hat{\gamma}_0^{G+1} & \hat{\gamma}_0^{G+2} & \dots & \hat{\gamma}_0^K \\ \hat{\gamma}_1^{G+1} & \hat{\gamma}_1^{G+2} & \dots & \hat{\gamma}_1^K \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\gamma}_G^{G+1} & \hat{\gamma}_G^{G+2} & \dots & \hat{\gamma}_G^K \end{bmatrix} = [\hat{\gamma}^{G+1} \quad \hat{\gamma}^{G+2} \quad \dots \quad \hat{\gamma}^K]. \quad (\text{A.144})$$

I fallet när \mathbf{X}_B är en kolumnvektor (dvs. innehåller bara en variabel) erhåller vi ekvation (6.24) från boken.

Vi kan också använda ekvation (A.139) för att visa det som kallas för Frisch-Waugh-Lovell-teoremet. I stället för att multiplicera med \mathbf{X}_A så multiplicerar vi med $\mathbf{X}'_B \mathbf{M}_A$, där $\mathbf{M}_A = \mathbf{I}_n - \mathbf{X}_A(\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A$ är förintarmatrisen (se avsnitt A.4) från en regression av \mathbf{y} på \mathbf{X}_A . Vi erhåller

$$\mathbf{X}'_B \mathbf{M}_A \hat{\mathbf{u}}^l = \mathbf{X}'_B \mathbf{M}_A \mathbf{y} - \mathbf{X}'_B \mathbf{M}_A \mathbf{X}_A \hat{\boldsymbol{\beta}}_A^l - \mathbf{X}'_B \mathbf{M}_A \mathbf{X}_B \hat{\boldsymbol{\beta}}_B^l. \quad (\text{A.145})$$

Notera att $\mathbf{X}'_B \mathbf{M}_A \hat{\mathbf{u}}^l = \mathbf{X}'_B \hat{\mathbf{u}}^l - \mathbf{X}_B \mathbf{X}_A (\mathbf{X}'_A \mathbf{X}_A)^{-1} \mathbf{X}'_A \hat{\mathbf{u}}^l$, och således, från villkoren $\mathbf{X}'_A \hat{\mathbf{u}}^l = \mathbf{0}$ och $\mathbf{X}'_B \hat{\mathbf{u}}^l = \mathbf{0}$, följer att vektorn på vänster sida om likhetstecknet i ekvation (A.145) är noll. Detta innebär att vi kan lösa för $\hat{\boldsymbol{\beta}}_B^l$:

$$\hat{\boldsymbol{\beta}}_B^l = (\mathbf{X}'_B \mathbf{M}_A \mathbf{X}_B)^{-1} \mathbf{X}'_B \mathbf{M}_A \mathbf{y}. \quad (\text{A.146})$$

Detta kan skrivas

$$\begin{aligned} \hat{\boldsymbol{\beta}}_B^l &= (\mathbf{X}'_B \mathbf{M}'_A \mathbf{M}_A \mathbf{X}_B)^{-1} \mathbf{X}'_B \mathbf{M}'_A \mathbf{M}_A \mathbf{y} \\ &= ((\mathbf{M}_A \mathbf{X}_B)' (\mathbf{M}_A \mathbf{X}_B))^{-1} (\mathbf{M}_A \mathbf{X}_B)' (\mathbf{M}_A \mathbf{y}) \\ &= (\widetilde{\mathbf{X}}'_B \widetilde{\mathbf{X}}_B)^{-1} \widetilde{\mathbf{X}}'_B \widetilde{\mathbf{y}}, \end{aligned} \quad (\text{A.147})$$

där $\widetilde{\mathbf{X}}_B = \mathbf{M}_A \mathbf{X}_B$ och $\widetilde{\mathbf{y}} = \mathbf{M}_A \mathbf{y}$.⁵ Detta innebär alltså att $\hat{\boldsymbol{\beta}}_B^l$ kan uttryckas som en koefficientvektor från en regression av $\widetilde{\mathbf{y}}$ på $\widetilde{\mathbf{X}}_B$. Det gäller att $\widetilde{\mathbf{y}} = \mathbf{M}_A \mathbf{y} = \hat{\mathbf{u}}^k$, vilket ges av förintarmatrisens egenskap (se ekvation A.52). På samma sätt gäller att

$$\widetilde{\mathbf{X}}_B = \mathbf{M}_A \mathbf{X}_B = [\hat{\mathbf{v}}^{G+1} \quad \hat{\mathbf{v}}^{G+2} \quad \dots \quad \hat{\mathbf{v}}^K], \quad (\text{A.148})$$

där $\hat{\mathbf{v}}^j$ är en vektor av residualer från skattning av ekvation (A.143). Vi kan alltså se på $\widetilde{\mathbf{y}}$ som den delen av \mathbf{y} som inte kan förklaras av \mathbf{X}_A medan $\widetilde{\mathbf{X}}_B$ är den delen av \mathbf{X}_B som inte kan förklaras av \mathbf{X}_A . Det faktum att $\hat{\boldsymbol{\beta}}_B^l$ kan tas fram antingen i den långa regressionen eller i flera steg där \mathbf{X}_B och \mathbf{y} "residualiseras" genom att \mathbf{X}_A "partialiseras ut" är det som brukar kallas för Frisch-Waugh-Lovell-teoremet, efter Frisch och Waugh (1933) och Lovell (1963), vilket också är det som Angrist och Pischke (2008) kallar för *regressionsanatomy*. I boken använder vi det begreppet för att referera till ekvation (A.142), men som vi kunde se ovan så finns en direkt koppling till Frisch-Waugh-Lovell-teoremet.

A.7 Regressionsanalys med beroende felterm

I avsnittet ovan har vi antagit att data, $\{\mathbf{x}_i, y_i\}_{i=1}^n$, är ett slumpmässigt stickprovet från en population. Innebörden av detta är att feltermerna är oberoende. Genom att använda formler för robusta standardfel visade vi att vi, vid inferens, kan tillåta för heteroskedasticitet om vi har stort stickprov, men att antagande om oberoende felterm var essentiell för vår inferens.

I den här delen redogör vi för tre fall där vi tänker på data som insamlat på ett alternativt sätt och då med felterm som sannolikt är korrelerade. Det första fallet är vad som kallas klustade data. Det andra fallet är när vi har *paneldata* eller *longitudinella* data. Slutligen går vi igenom skattning och asymptotisk inferens med både klustrad data och paneldata.

⁵Vi använder oss av räkneregler för transformering av konforma matriser (dvs. att de kan multipliceras). För två konforma matriser \mathbf{A} och \mathbf{B} gäller $(\mathbf{AB})' = \mathbf{B}' \mathbf{A}'$.

A.7.1 Klustrade data

För att få en intuition av vad som menas med klustrade data beskriver vi först ett klusterrandomiserat experiment. I sådana experiment ges alla inom en eller flera grupper en behandling medan ingen får behandling i andra grupper. Denna experimentella design är vanlig förekommande inom utvecklingsekonomi och inom utbildning. Inom utbildning kan vi t.ex. slumpa in elever i små och stora klasser. Behandling är då på klassnivå och här är det ganska tydligt att vi inte kan behandla eleverna inom klasserna (klustren) som oberoende.

Med observationsdata blir det mindre tydligt hur man definierar kluster. Med fokus på orsaksanalys av t.ex. en reform i ett fåtal regioner är det numera standard att behandla de regioner (behandlade och obehandlade) som används i analysen som kluster i en regressionsanalys (läs mer om detta i boken och då specifikt avsnitt 12.5).

För att beskriva hur man kan hantera detta beroende inom kluster måste vi införa lite mer definitioner. För varje grupp eller kluster g , låt $\{y_{gi}, \mathbf{x}_g, \mathbf{z}_{gm} : i = 1, \dots, n_g\}$ vara våra observerbara data, där n_g är (fixt) antalet enheter i kluster g , y_{gi} är en skalär, \mathbf{x}_g är en $K \times 1$ vektor som innehåller förklarande variabler som endast varierar på klusternivå, \mathbf{z}_{gi} är en $L \times 1$ vektor av kovariater som varierar inom (såväl som över) grupper och $n = \sum_{g=1}^G n_g$.

Den viktigaste skillnaden mot tidigare är att urvalet av kluster behandlas som hämtat från en infinit population av kluster. Klustren antas vara oberoende av varandra, men att utfall inom ett kluster bör tillåtas vara korrelerade även betingat på våra kovariater.

Ett exempel är ett slumpmässigt urval av klassrum i fjärde klass och med data på elevutfall. Varje klass är ett kluster och elever inom en klass är de individuella enheterna. Ett annat är ett slumpmässigt urval av industrier och med data på, exempelvis, vinster från företag som utfall. I detta fall är industri ett kluster medan företag är de individuella enheterna. Ett tredje exempel är ett urval av sjukhus och med data på, exempelvis, anställdas löner som utfall. Notera att i alla dessa fall är antalet kluster finita (dvs. det finns en viss mängd skolor, industrier, sjukhus i ett land). Vi betraktar dock dessa kluster som dragna från en oändlig population som vi kallar superpopulation. För de som är intresserade av inferens när man betraktar klustren som fixa, se Abadie m. fl. (2020). För en utökad diskussion om och när man kan klustra, se Abadie m. fl., 2022.

Den linjära modellen med ett additivt fel kan nu skivas

$$y_{gi} = \alpha + \mathbf{x}'_g \boldsymbol{\beta} + \mathbf{z}'_{gi} \boldsymbol{\gamma} + v_{gi}, \quad i = 1, \dots, n_g, g = 1, \dots, G, \quad (\text{A.149})$$

och

$$v_{gi} = c_g + u_{gi}, \quad i = 1, \dots, n_g. \quad (\text{A.150})$$

Feltermen v_{gi} innehåller nu en icke observerbar variabel som är samma för alla individer inom ett kluster, c_g , samt u_{gi} som antas vara helt slumpmässig.

Innebörden av denna stuktur är att det finns ett beroende mellan individer inom ett kluster (via c_g) men att individerna mellan klustren är oberoende, betingat på våra kovariater $(\mathbf{x}_g, \mathbf{z}_{gi})$.⁶

⁶Detta är ett antagande som kan vara felaktigt. Vid rumslig statistik ("spatial statistics") kan det finnas geografiska beroenden. Man kan då tänka på ett väldefinierat geografiskt område, exempelvis kommun, som ett kluster. Men inom klustret finns sannolikt geografiska beroenden så att individer som bor nära varandra inom

Notera att med denna struktur är det i första hand när vi är intresserad av inferens om β som vi måste betrakta data som klustrat för att säkerställa att vi drar rätt inferens. Således, om intrese är av inferens av γ – och z_{gi} inte har beroende inom kluster – kan vi analysera data på samma sätt som tidigare. Det är dock sannolikt bättre att bara använda variationen inom gruppen för analys. Denna analys benämns ofta för ”fixed effect”-analys och beskrivs i detalj nedan.

A.7.1.1 Estimation och inferens

Definiera följande:

$$\boldsymbol{\lambda} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix}_{(1+K+L) \times 1}, \quad \mathbf{W}_g = \begin{bmatrix} 1 & \mathbf{x}'_g & \mathbf{z}'_{g1} \\ 1 & \mathbf{x}'_g & \mathbf{z}'_{g2} \\ \vdots & & \\ 1 & \mathbf{x}'_g & \mathbf{z}'_{gn_g} \end{bmatrix}_{n_g \times (1+K+L)}, \quad \mathbf{y}_g = \begin{bmatrix} y_{g1} \\ y_{g2} \\ \vdots \\ y_{gn_g} \end{bmatrix}_{n_g \times 1}, \quad \mathbf{v}_g = \begin{bmatrix} v_{g1} \\ v_{g2} \\ \vdots \\ v_{gn_g} \end{bmatrix}_{n_g \times 1}, \quad \mathbf{c}_g = \begin{bmatrix} c_g \\ c_g \\ \vdots \\ c_g \end{bmatrix}_{n_g \times 1}, \quad \mathbf{u}_g = \begin{bmatrix} u_{g1} \\ u_{g2} \\ \vdots \\ u_{gn_g} \end{bmatrix}_{n_g \times 1}. \quad (\text{A.151})$$

Vi kan skriva regressionen

$$\mathbf{y}_g = \mathbf{W}_g \boldsymbol{\lambda} + \mathbf{v}_g, \quad g = 1, \dots, G \Leftrightarrow \mathbf{y} = \mathbf{W} \boldsymbol{\lambda} + \mathbf{v} \Leftrightarrow \mathbf{y} = \mathbf{W} \boldsymbol{\lambda} + \mathbf{c} + \mathbf{u}, \quad (\text{A.152})$$

där

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_G \end{bmatrix}_{n \times 1}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_G \end{bmatrix}_{n \times (1+K+L)}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_G \end{bmatrix}_{n \times 1}, \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_G \end{bmatrix}_{n \times 1}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_G \end{bmatrix}_{n \times 1}, \quad (\text{A.153})$$

vilket innebär att OLS-estimatorn kan skrivas

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{y} \\ &= (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'(\mathbf{W}\boldsymbol{\lambda} + \mathbf{v}) \\ &= \boldsymbol{\lambda} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'(\mathbf{c} + \mathbf{u}) \\ &= \boldsymbol{\lambda} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{c} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{u} \\ &= \boldsymbol{\lambda} + (\mathbf{W}'\mathbf{W}/n)^{-1} (\mathbf{W}'\mathbf{c}/n) + (\mathbf{W}'\mathbf{W}/n)^{-1} (\mathbf{W}'\mathbf{u}/n). \end{aligned} \quad (\text{A.154})$$

Denna estimator kallas ibland för *polad OLS*, vilket är vår översättning av engelskans *pooled OLS*. Det engelska verbet *pool* innebär att kombinera olika delar till något gemensamt och

kommunen har starkare beroende än individer som bor långt ifrån varandra inom samma kommun. I så fall är feltermstrukturerna i ekvation (A.150) inte en lämplig beskrivning av beroende.

i detta fall syftar det på att data från de olika klustrena läggs ihop i stora matriser enligt ekvation (A.153) och en enda regression skattas. Det är viktigt att notera att rent tekniskt är polad OLS exakt samma sak som OLS. Begreppet används för att tydliggöra vilken typ av data som används. Normalt brukar begreppet polad OLS användas när man har paneldata (se avsnitt A.7.2), men det går att se på paneldata som ett specialfall av klustrad data, vilket är anledningen till att vi använder begreppet här.

På samma sätt som vi definierade koefficientvektorn i populationen i avsnitt A.2 så har vi här att

$$\boldsymbol{\lambda} = \arg \min_{\boldsymbol{l}} E((Y - \boldsymbol{w}'\boldsymbol{l})^2). \quad (\text{A.155})$$

Detta implicerar att \boldsymbol{w} är okorrelerad med både grupp-termen, c_g , och individ-termen, u_{gi} . Det kan därför vara frestande att tro att de sista två termerna i ekvation (A.154) går till noll när stickprovsstorleken går mot oändligheten. Här måste vi dock vara försiktiga med vad som ska gå mot oändligheten. För att se detta, notera att vi kan skriva ekvation (A.154) som

$$\begin{aligned} \hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda} + & \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} \sum_{i=1}^{n_g} \boldsymbol{w}_{gi} \boldsymbol{w}_{gi}' \right)^{-1} \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} c_g \sum_{i=1}^{n_g} \boldsymbol{w}_{gi} \right) \\ & + \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} \sum_{i=1}^{n_g} \boldsymbol{w}_{gi} \boldsymbol{w}_{gi}' \right)^{-1} \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} \sum_{i=1}^{n_g} \boldsymbol{w}_{gi} u_{gi} \right) \quad (\text{A.156}) \end{aligned}$$

Medan den tredje termen i ekvation (A.154) konvergerar till noll med stickprovsstorleken, $n = \sum_{g=1}^G n_g$, så kommer den andra termen inte att göra det eftersom c_g är samma för alla observationer inom kluster g . Den tredje termen konvergerar i stället med antalet kluster, G under antagandet att kluster dras oberoende från populationen. Det gäller då att $\text{plim}_{G \rightarrow \infty}(\hat{\boldsymbol{\lambda}}) = \boldsymbol{\lambda}$. På samma sätt som i avsnitt A.5.1 kan vi visa asymptotisk normalitet, där variansen av $\hat{\boldsymbol{\lambda}}$ kan skattas som

$$\hat{V}^c(\hat{\boldsymbol{\lambda}}) = \left(\sum_{g=1}^G \boldsymbol{w}_g' \boldsymbol{w}_g \right)^{-1} \left(\sum_{g=1}^G \boldsymbol{w}_g' \hat{\boldsymbol{v}}_g \hat{\boldsymbol{v}}_g' \boldsymbol{w}_g \right) \left(\sum_{g=1}^G \boldsymbol{w}_g' \boldsymbol{w}_g \right)^{-1}, \quad (\text{A.157})$$

där $\hat{\boldsymbol{v}}_g = \boldsymbol{y}_g - \boldsymbol{W}_g \hat{\boldsymbol{\lambda}}$. Denna variansskattare kallas för en *klusterrobust kovariansmatris* och den bygger alltså på att G är stort. Ibland används frihetsgradsjusteringar i kovariansmatrisen. Exempelvis, med funktionen `coefest()` i R som beskrivs i boken så multipliceras formeln ovan med $G/(G-1) \cdot (n-1)/(n-(K+L+1))$. Denna justering får ytterst liten roll när G är stort och det finns olika varianter på sådana justeringar.

Med denna kovariansmatris utnyttjar vi inte ekvation (A.150) där inom-gruppskorrelationen bestäms uteslutande av c_g som är fixt inom gruppen. Denna klusterrobusta kovariansmatris fungerar för vilka andra inom-gruppskorrelationsstrukturer som helst (givet finita varianser och kovarianser). Men om feltermsstrukturen i (A.150) där u_{ig} är helt slumpmässig så går det att få mer effektiva skattare av $\boldsymbol{\lambda}$. Vi återkommer till detta i Bilaga B.

A.7.1.2 Bara inferens om γ

Om vi bara är intresserad av γ kan vi ta hänsyn till c_g genom att enbart använda *inomgruppsvariation*. Med tillräckligt stora n_g i varje kluster kan vi skatta G stycken parametervektorer, dvs $\hat{\gamma}_g$ för $g = 1, \dots, G$. Från dessa enskilda estimat kan vi sedan skapa ett genomsnitt $\hat{\gamma} = \sum_{g=1}^G (n_g/n) \hat{\gamma}_g$.

Med små n_g är det inte säkert att detta fungerar så bra. Istället kan vi använda inomgruppsvariation och skatta en genomsnittsparameter med en polad regression. Estimatoren kallas ofta för en fix-effekt-estimator (FE-estimator), eftersom vi tänker oss att vi tar bort c_g genom att bilda differensen med \bar{y}_g .⁷

Gruppmedelvärdet av ekvation (A.149) kan skrivas som

$$\bar{y}_g = \alpha + \mathbf{x}'_g \boldsymbol{\beta} + \bar{\mathbf{z}}'_g \boldsymbol{\gamma} + \bar{v}_g, \quad i = 1, \dots, n_g, g = 1, \dots, G. \quad (\text{A.158})$$

Inom-gruppsvariationen erhålls genom att dra bort (A.158) från (A.149):

$$\begin{aligned} y_{gi} - \bar{y}_g &= \alpha + \mathbf{x}'_g \boldsymbol{\beta} + \mathbf{z}'_{gi} \boldsymbol{\gamma} + v_{gi} - (\alpha + \mathbf{x}'_g \boldsymbol{\beta} + \bar{\mathbf{z}}'_g \boldsymbol{\gamma} + \bar{v}_g) \\ &= (\mathbf{z}_{gi} - \bar{\mathbf{z}}_g)' \boldsymbol{\gamma} + v_{gi} - \bar{v}_g \\ &= (\mathbf{z}_{gi} - \bar{\mathbf{z}}_g)' \boldsymbol{\gamma} + c_g - c_g + u_{gi} - \bar{u}_g \\ &= (\mathbf{z}_{gi} - \bar{\mathbf{z}}_g)' \boldsymbol{\gamma} + u_{gi} - \bar{u}_g. \end{aligned} \quad (\text{A.159})$$

Låt $\tilde{y}_{gi} = y_{gi} - \bar{y}_g$, $\tilde{\mathbf{z}}_{gi} = \mathbf{z}_{gi} - \bar{\mathbf{z}}_g$ och $\tilde{u}_{gi} = u_{gi} - \bar{u}_g$. FE-estimatoren skrivs då:

$$\hat{\gamma} = (\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}' \tilde{\mathbf{y}} = \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} \sum_{i=1}^{n_g} \tilde{\mathbf{z}}_{gi} \tilde{\mathbf{z}}'_{gi} \right)^{-1} \left(\frac{1}{G} \sum_{g=1}^G \frac{1}{n_g} \sum_{i=1}^{n_g} \tilde{\mathbf{z}}_{gi} \tilde{y}_{gi} \right), \quad (\text{A.160})$$

där $\tilde{\mathbf{Z}}$ och $\tilde{\mathbf{y}}$ är definierade på samma sätt som \mathbf{Z} och \mathbf{y} där kluster-medelvärdena dragits bort. Estimatoren är konsistent och asymptotiskt normalfördelad i stickprovsstorleken.

Låt \mathbf{D} vara en matris av dimension $n \times L$ som antar värdet ett i rad i och kolumn g om observation i tillhör kluster g och noll annars. Det går att visa att $\tilde{\mathbf{Z}} = \mathbf{M}_D \mathbf{Z}$ och $\tilde{\mathbf{y}} = \mathbf{M}_D \mathbf{y}$ där $\mathbf{M}_D = \mathbf{I}_n - \mathbf{D}(\mathbf{D}' \mathbf{D})^{-1} \mathbf{D}'$. Det gäller alltså att ekvation (A.160) kan skrivas, helt ekvivalent, som

$$\hat{\gamma} = (\mathbf{Z}' \mathbf{M}'_D \mathbf{M}_D \tilde{\mathbf{Z}})^{-1} \mathbf{Z}' \mathbf{M}'_D \mathbf{M}_D \mathbf{y} \quad (\text{A.161})$$

Uti från Frish-Waugh-Lovell teoremet i ekvation (A.147) vet vi då att en regression med $\tilde{\mathbf{y}}$ på $\tilde{\mathbf{Z}}$ ger samma koefficientskattningar av $\boldsymbol{\gamma}$ som om vi skattar en regression med \mathbf{y} på $[\mathbf{Z} \quad \mathbf{D}]$. \mathbf{D} är en full uppsättning dummyvariabler för kluster vilket alltså innebär att en lång regression med en full uppsättning dummyvariabler ger samma skattningar av $\boldsymbol{\gamma}$ som en kort regression där vi har dragit bort medelvärdena för \mathbf{y} och \mathbf{Z} . Skillnaden är att med den långa regressionen erhålls även skattningar av c_g . Dessa är dock sällan av intresse och det är betydligt mer beräkningsintensivt

⁷Ibland kallas c_g för fixa effekter. Detta är dock missvisande, då modellen är konsistent med att tänka på c_g som slumpmässiga. Det är proceduren vid skattning som kan ses som en FE-estimator.

att skatta den långa regressionen. Därför kan det av praktiska skäl vara bättre att skatta den korta regressionen där medelvärdena dragits bort från de ingående variablerna.

Givet att u_{gi} är homoskedastisk och oberoende kan vanlig asymptotisk inferens göras med följande skattade kovariansmatris:

$$\hat{V}(\hat{\gamma}) = s_U^2 \left(\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} \right)^{-1} = s_U^2 \left(\sum_{g=1}^G \tilde{\mathbf{Z}}'_g \tilde{\mathbf{Z}}_g \right)^{-1}, \quad (\text{A.162})$$

där

$$s_U^2 = \frac{1}{n - G - L} \sum_{g=1}^G \sum_{i=1}^{n_g} \hat{u}_{gi}^2. \quad (\text{A.163})$$

Notera att, utöver de L parametrarna i γ som skattas så drar vi även bort G i nämnaren därför att vi kan tänka på vår estimator som att den skattar G stycken intercept (c_g).

Generellt finns inga garantier för att u_{gi} är homoskedastisk och dessutom kan det finnas beroende så att $\text{cov}(u_{gi}, u_{gj}) \neq 0$ för $i \neq j$. En klusterrobust kovariansmatris är

$$\hat{V}^c(\hat{\gamma}) = \left(\sum_{g=1}^G \tilde{\mathbf{Z}}'_g \tilde{\mathbf{Z}}_g \right)^{-1} \left(\sum_{g=1}^G \tilde{\mathbf{Z}}'_g \hat{\mathbf{u}}_g \hat{\mathbf{u}}_g' \tilde{\mathbf{Z}}_g \right) \left(\sum_{g=1}^G \tilde{\mathbf{Z}}'_g \tilde{\mathbf{Z}}_g \right)^{-1}. \quad (\text{A.164})$$

Denna kovariansmatris tillåter för heteroskedasticitet och beroende feltermerna av godtycklig struktur inom kluster. Däremot så bygger den på att feltermerna mellan kluster är oberoende. Som tidigare är denna estimator av kovariansmatrisen konsistent om G är stor.

A.7.1.3 Diskussion gällande asymptotik

Vad händer om antagandet om fixt n_g och $G \rightarrow \infty$ inte är realistiskt? Hansen (2007) visar att om både G och n_g går mot oändligheten så är inferensen till β med den klusterrobusta variansestimatoren valid med godtycklig korrelation i feltermerna inom varje grupp. Om vi tex har ett stickprov av $G = 100$ skolor med ca $n_g = 100$ elever i varje skola är alltså den klusterrobusta inferensen korrekt.

Om vi å andra sidan har ett litet antal grupper (G) med stort antal observationer inom varje grupp (n_g) är den klusterrobusta inferensen för β felaktig. Om vi har data från $G = 10$ sjukhus med data på tusentals patienter för varje sjukhus där förklarande variabler endast varierar mellan sjukhus är det frestande att använda polad OLS och att göra klusterrobust inferens. Inferensen kommer dock i sådant fall bli missvisande.

A.7.2 Paneldata

Paneldata innebär att vi har data för stickprov på n enheter från en population där vi observerar oberoende och beroende variabler över tid, $t = 1, \dots, T$. Antalet observationer för individ i är T_i där $T_i \leq T$. Totala antalet observationer betecknar vi med N , vilket således är $N = \sum_{i=1}^n T_i$.

I många fall antar T_i relativt små värden, vilket vi kommer utgå från i detta avsnitt. Om T_i i stället antar stora värden är det mer lämpligt att behandla data som individuella tidserier.

Vad som är stort beror på om de individuella tidserierna är stationära (se avsnitt 11.2 i boken för en definition. Stationäritet innebär exempelvis avsaknad av trend) och med vilken frekvens data observeras. Om data är stationära och vi har data mätt på dagsnivå kan T_i vara mycket stor, exempelvis $T_i > 365$ ska inte vara ett problem. Om vi har årliga data och tidserierna inte kan ses som stationära kan till och med $T_i > 3$ vara ett problem, dvs. vid inferens måste en bedömning av lämpligheten med metoden nedan göras. Vi utgår också från att T_i är fixt och att asymptotiska resultat bygger på att n går mot oändligheten.

Om $T_i = T$, för alla i har vi en balanserad panel, annars har vi en obalanserad panel. För att inferensen ska vara korrekt vid en obalanserad panel måste det vara slumpmässigt att T är olika givet våra kovariater. Detta antagande kallas för *missing at random* (MAR). Innebörden är att residualvariation i T_i efter det att vi justerat för observerade kovariater är helt slumpmässigt.⁸ Vi betecknar det första tidpunkten för vilken vi observerar data för individ i som t_{i1} , den andra som t_{i2} osv. fram till t_{iT_i} . Således observeras data $\{y_{it}, \mathbf{x}_{it}\}_{t=t_{i1}}^{t_{iT_i}}$ för $i = 1, \dots, n$.

Regressionsmodeller med paneldata har samma struktur som vid klustrade data. Man byter bara ut g för individ i och n_g för T_i . Detta innebär alltså att varje individ är att betrakta som ett eget kluster. Den linjära regressionsmodellen med ett additivt fel kan nu skrivas

$$y_{it} = \alpha + \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_{it} \boldsymbol{\gamma} + v_{it}, \quad (\text{A.165})$$

och

$$v_{it} = c_i + f_t + u_{it}. \quad (\text{A.166})$$

Skillnaden mot behandlingen av klustrade data ovan är att feltermen nu innehåller en term f_t . Detta är en variabel som är samma för alla individer vid en viss tidpunkt. Existensen av denna term kan indikera att det kan ske en gemensam shock till utfallet för alla individer, eller att det finns gemensamma tidstrender. Notera att i det vanliga kluster-fallet så har vi inte med en sådan term därför att normalt finns det ingen naturlig ordning på individer inom ett kluster.⁹ Men när vi har paneldata där varje individ är ett kluster så är tid en naturlig ordning inom ett kluster. Det gör också att det ofta är rimligt att göra dekomponeringen av feltermen enligt ekvation (A.166).

A.7.2.1 Estimation och inferens

Definiera

$$\boldsymbol{\lambda} = \begin{bmatrix} \alpha \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}_{(1+K+L) \times 1}, \quad \mathbf{W}_i = \begin{bmatrix} 1 & \mathbf{x}'_i & \mathbf{z}'_{it_{i1}} \\ 1 & \mathbf{x}'_i & \mathbf{z}'_{it_{i2}} \\ \vdots & & \\ 1 & \mathbf{x}'_i & \mathbf{z}'_{it_{iT_i}} \end{bmatrix}_{T_i \times (1+K+L)}, \quad \mathbf{y}_i = \begin{bmatrix} y_{it_{i1}} \\ y_{it_{i2}} \\ \vdots \\ y_{it_{iT_i}} \end{bmatrix}_{T_i \times 1}, \quad \mathbf{v}_i = \begin{bmatrix} v_{it_{i1}} \\ v_{it_{i2}} \\ \vdots \\ v_{it_{iT_i}} \end{bmatrix}_{T_i \times 1}, \quad (\text{A.167})$$

⁸Ett annat, relaterat, antagande som förekommer i litteraturen är *missing completely at random* (MCAR). Innebörden är att variationen i T_i är slumpmässig utan betingning på kovariater.

⁹Ett undantag som också nämndes ovan är när kluster innehåller en geografisk komponent.

och

$$\mathbf{c}_i = \begin{bmatrix} c_i \\ c_i \\ \vdots \\ c_i \end{bmatrix}_{T_i \times 1}, \quad \mathbf{f}_i = \begin{bmatrix} f_{t_{i1}} \\ f_{t_{i2}} \\ \vdots \\ f_{t_{iT_i}} \end{bmatrix}_{T_i \times 1}, \quad \mathbf{u}_i = \begin{bmatrix} u_{it_{i1}} \\ u_{it_{i2}} \\ \vdots \\ u_{it_{iT_i}} \end{bmatrix}_{T_i \times 1}. \quad (\text{A.168})$$

Vi kan skriva regressionen

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\lambda} + \mathbf{v}_i \Leftrightarrow \mathbf{y} = \mathbf{W} \boldsymbol{\lambda} + \mathbf{v}, \quad (\text{A.169})$$

där

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}_{N \times 1}, \quad \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_n \end{bmatrix}_{N \times (1+K+L)}, \quad \mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_n \end{bmatrix}_{N \times 1}, \quad \mathbf{c} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_n \end{bmatrix}_{N \times 1}, \quad \mathbf{f} = \begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \vdots \\ \mathbf{f}_n \end{bmatrix}_{N \times 1}, \quad \mathbf{u} = \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_n \end{bmatrix}_{N \times 1}. \quad (\text{A.170})$$

vilket innebär att den polade OLS-estimatorn kan skrivas

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{y} \\ &= (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'(\mathbf{W}\boldsymbol{\lambda} + \mathbf{v}) \\ &= \boldsymbol{\lambda} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'(\mathbf{c} + \mathbf{f} + \mathbf{u}) \\ &= \boldsymbol{\lambda} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{c} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{f} + (\mathbf{W}'\mathbf{W})^{-1} \mathbf{W}'\mathbf{u}. \end{aligned} \quad (\text{A.171})$$

För enkelhets skull utgår vi från att vi har en balanserad panel så att $T_i = T$ för alla i . I så fall har vi

$$\begin{aligned} \hat{\boldsymbol{\lambda}} &= \boldsymbol{\lambda} + \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \mathbf{w}_{it} \mathbf{w}_{it}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} c_i \sum_{t=1}^T \mathbf{w}_{it} \right) \\ &\quad + \left(\frac{1}{T} \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n \mathbf{w}_{it} \mathbf{w}_{it}' \right)^{-1} \left(\frac{1}{T} \sum_{t=1}^T \frac{1}{n} f_t \sum_{i=1}^n \mathbf{w}_{it} \right) \\ &\quad + \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \mathbf{w}_{it} \mathbf{w}_{it}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \mathbf{w}_{it} u_{it} \right) \end{aligned} \quad (\text{A.172})$$

Till skillnad från fallet med kluster finns här den tredje termen som konvergerar i sannolikhet till noll med T . För att $\hat{\boldsymbol{\lambda}}$ ska vara konsistent krävs alltså att både n och T är stort. Detta är problematiskt då många paneler har relativt små T . Det finns dock en enkel lösning: användning

av tidsfixa effekter. Precis som med de fixa effekterna för kluster som diskuterades ovan skapas nya variabler där medelvärden för respektive tidsperiod dras bort. Alternativt, vilket brukar vara det vanliga när T är litet så används en full uppsättning dummyvariabler för tid.

Om det inte finns några tidschocker så att $f_t = 0$ för $t = 1, \dots, T$ gäller att $\hat{\lambda}$ är konsistent och asymptotiskt (i n) normalfördelad. En klusterrobust kovariansmatris som tillåter för heteroskedasticitet och godtycklig seriekorrelation inom individ över tid är

$$\hat{V}^c(\hat{\lambda}) = \frac{n}{n - (K + L + 1)} \left(\sum_{i=1}^n \mathbf{w}_i' \mathbf{w}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{w}_i' \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \mathbf{w}_i \right) \left(\sum_{i=1}^n \mathbf{w}_i' \mathbf{w}_i \right)^{-1}. \quad (\text{A.173})$$

Om vi vill ta hänsyn till tidseffekter drar vi bort tidsmedelvärden från samtliga variabler i \mathbf{W} och från \mathbf{y} och skattar modellen (alternativt använder en full uppsättning dummyvariabler för tid).

Notera att vi inte har diskuterat den vanliga variansestimatorn. Anledningen är att förekomsten av c_i i feltermen gör att den vanliga variansestimatorn inte är korrekt.

A.7.2.2 Inferens rörande enbart γ

Om vi bara är intresserad av inferens till γ är det som regel bäst att använda en FE-estimator som tar hänsyn till c_i . Om vi inte tror att det finns några tidseffekter att ta hänsyn till, dvs. $f_t = 0$ så har vi

$$y_{it} - \bar{y}_i = (\mathbf{z}_{it} - \bar{\mathbf{z}}_i)' \gamma + u_{it} - \bar{u}_i. \quad (\text{A.174})$$

Vi har exakt samma modell som i ekvation (A.159), med skillnaden att vi ersatt kluster med individ och där vi har flera observationer vid olika tidpunkter i stället för olika individer i samma grupp. Återigen utgår vi för enkelhets skull att panelen är balanserad. FE-estimatorn motsvarande ekvation (A.160) blir då

$$\hat{\gamma} = (\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}})^{-1} \tilde{\mathbf{Z}}' \tilde{\mathbf{y}} = \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{z}}_{it} \tilde{\mathbf{z}}_{it}' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \frac{1}{T} \sum_{t=1}^T \tilde{\mathbf{z}}_{it} \tilde{y}_{it} \right). \quad (\text{A.175})$$

Om feltermen U är homoskedastisk och oberoende kan vi använda följande skattning av kovariansmatrisen:

$$\hat{V}(\hat{\gamma}) = s_U^2 \left(\tilde{\mathbf{Z}}' \tilde{\mathbf{Z}} \right)^{-1} = s_U^2 \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i \right)^{-1}, \quad (\text{A.176})$$

där

$$s_U^2 = \frac{1}{N - n - L} \sum_{i=1}^n \sum_{t=1}^T \hat{u}_{it}^2. \quad (\text{A.177})$$

Normalt sett är det tveksamt om feltermen är homoskedastisk, och framförallt kan vi ha seriekorrelation i feltermen så att feltermerna inte är oberoende. Den kluster-robusta kovariansmatrisen bör då användas i stället:

$$\hat{V}^c(\hat{\gamma}) = \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i \right)^{-1} \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \tilde{\mathbf{Z}}_i \right) \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i \right)^{-1}. \quad (\text{A.178})$$

Om vi också vill bli av med tidseffekterna finns det flera sätt vi kan göra detta på. Låt \mathbf{D}_T vara en matris av dummyvariabler för tid av dimension $N \times T$ medan \mathbf{D}_I är en matris av dummyvariabler för individer av dimension $N \times (n - 1)$ där vi har $n - 1$ kolumner för att vi måste ha en referenskategori. Om vi exempelvis har en balanserad panel med $T = 3$ har vi följande matriser:

$$\mathbf{D}_T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{D}_I = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}, \quad (\text{A.179})$$

där sista tre raderna i \mathbf{D}_I enbart innehåller nollor då sista individen (med data från tre tidsperioder) får agera referensindivid.¹⁰

Om vi låter $\mathbf{D} = [\mathbf{D}_T \quad \mathbf{D}_I]$ och $\mathbf{M}_D = \mathbf{I}_{nT} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ kan vi skatta γ som

$$\hat{\gamma} = (\mathbf{Z}'\mathbf{M}'_D\mathbf{M}_D\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}'_D\mathbf{M}_D\mathbf{y}. \quad (\text{A.180})$$

Alternativt kan vi dra bort medelvärde för individer och skapa $\tilde{\mathbf{Z}}$ och $\tilde{\mathbf{y}}$. Vi kan sedan lägga till en full uppsättning av dummyvariabler för tid så att matrisen av oberoende variabler är $[\tilde{\mathbf{Z}} \quad \mathbf{D}_T]$, vilken har dimension $n \times (L + T)$. Vi kan sedan skatta en linjär regression enligt följande:

$$\hat{\theta} = \left([\tilde{\mathbf{Z}} \quad \mathbf{D}_T]' [\tilde{\mathbf{Z}} \quad \mathbf{D}_T] \right)^{-1} [\tilde{\mathbf{Z}} \quad \mathbf{D}_T]' \tilde{\mathbf{y}}. \quad (\text{A.181})$$

Parameterskattningarn $\hat{\theta}$ är då en vektor med $L + T$ element där de första L raderna är skattningar av γ , vilket enligt Frisch-Waugh-Lovell-teoremet kan skrivas som

$$\begin{aligned} \hat{\gamma} &= (\tilde{\mathbf{Z}}'\mathbf{M}'_{D_T}\mathbf{M}_{D_T}\tilde{\mathbf{Z}})^{-1}\tilde{\mathbf{Z}}'\mathbf{M}'_{D_T}\mathbf{M}_{D_T}\tilde{\mathbf{y}}, \\ &= (\mathbf{Z}'\mathbf{M}'_{D_I}\mathbf{M}'_{D_T}\mathbf{M}_{D_T}\mathbf{M}_{D_I}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}'_{D_I}\mathbf{M}'_{D_T}\mathbf{M}_{D_T}\mathbf{M}_{D_I}\mathbf{y}, \end{aligned} \quad (\text{A.182})$$

där $\mathbf{M}_{D_T} = \mathbf{I}_{nT} - \mathbf{D}_T(\mathbf{D}'_T\mathbf{D}_T)^{-1}\mathbf{D}'_T$ och $\mathbf{M}_{D_I} = \mathbf{I}_{nT} - \mathbf{D}_I(\mathbf{D}'_I\mathbf{D}_I)^{-1}\mathbf{D}'_I$. Om panelen är balanserad går det att visa att $\mathbf{M}'_{D_I}\mathbf{M}'_{D_T}\mathbf{M}_{D_T}\mathbf{M}_{D_I} = \mathbf{M}'_D\mathbf{M}_D$, vilket innebär att estimatorerna i ekvation (A.180) och (A.182) är helt ekvivalenta med varandra. I fall när panelen är obalanserad gäller inte detta.

¹⁰I det här fallet har vi med en T dummyvariabler för tid då modellen skattas utan konstant. En konstant kan dock läggas till i vilket fall vi enbart kan ha med $T - 1$ dummyvariabler för tid. Skattningen av γ blir samma oavsett vilket sätt som väljs, men skattningarna av tidseffekterna kommer skilja sig åt. Med en referenstid i modellen kommer tidseffekterna skattas relativt referenstiden.

Om feltermen U är homoskedastisk och oberoende och vi låter $\tilde{\mathbf{Z}}$ stå för $\mathbf{M}_D \mathbf{Z}$ kan vi skatta kovariansmatrisen enligt ekvation (A.175), där

$$s_U^2 = \frac{1}{N - n - L - T} \sum_{i=1}^n \sum_{t=1}^T \hat{u}_{it}^2. \quad (\text{A.183})$$

Detta gäller alltså där residulerna, \hat{u}_{it} kommer från en modell med både tidsfixa och individfixa effekter.

Men generellt bör vi använda en kluster-robust variansskattning i stället:

$$\hat{V}^c(\hat{\gamma}) = \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i \right)^{-1} \left(\sum_{i=1}^b \tilde{\mathbf{Z}}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \tilde{\mathbf{Z}}_i \right) \left(\sum_{i=1}^n \tilde{\mathbf{Z}}_i' \tilde{\mathbf{Z}}_i \right)^{-1}. \quad (\text{A.184})$$

A.7.2.3 Datagenererande processer och paneldata

Vi har utgått från att koefficientvektorn, λ , definieras utifrån minimerandet av kvadratavvikelse mellan den linjära projektionen och faktiskt utfall i superpopulationen. Denna superpopulation antas bestå av ett oändligt antal individer, med ett fixt, givet, antal tidsperioder per individ.

I många fall när modeller skattas är man dock intresserad av parametrar i en datagenererande process (dvs. en modell, se avsnitt A.3) eller av kausala effekter (se kapitel 12 i boken). I sådana fall introducerar tidsdimensionen i paneldata en extra komplikation som behöver tas hänsyn till. Exempelvis kan det finnas feedback-loopar sådana att värdet på den strukturella feltermen vid en viss tidpunkt påverkar värdet på en oberoende variabel vid följande tidpunkt. Förekomsten av sådana feedback-loopar kan göra att det inte är möjligt att på ett konsistent sätt skatta parametrarna i den datagenererande processen.

Vi diskuterar inte sådana problem i denna bilaga, utan hänvisar den intresserade läsaren till, exempelvis, Wooldridge (2010) för en diskussion om under vilka förutsättningar det går att konsistent skatta parametrarna i den datagenererande processen.

A.7.3 Klustring och paneldata

Ibland har vi både en panel och en klusterdimension i data. Med dessa data finns det således två potentiella typer av korrelationer: över tid och inom kluster. Antag att varje enhet tillhör ett kluster och att enheterna/individerna inte kan byta kluster/grupp mellan tidsperioderna. Exempelvis kan vi ha en årlig panel med data kring, exempelvis, vinster eller löner för företag som tillhör olika industrier. Om vi i detta fall vill analysera vinster för företagen kan det i detta fall vara relevant att se på en industri som ett kluster (gemensamma chocker för respektive industri). Om vi å andra sidan är intressad av att analysera lönesättning är det lämpligt att låta företag vara ett kluster men att kanske göra analyserna separat för varje industri. Således vad som är ett kluster bestäms av vilken frågeställning vi vill besvara och vilken typ av data vi har.

Nu har vi tre index på åtminstone några variabler som vi observerar. Till exempel är utfallet, y_{git} , där g indexerar gruppen eller klustret, i är enheten inom gruppen, och t är tidsindex. Antag

för enkelhet att vi har en balanserad panel med tidsperioderna $t = 1, \dots, T$. Inom klustret g finns n_g enheter, och vi har ett stickprov från G kluster.

Vi kan skriva regressionsmodellen

$$y_{git} = \alpha + \mathbf{r}'_g \boldsymbol{\delta} + \mathbf{x}'_{gi} \boldsymbol{\beta} + \mathbf{z}'_{git} \boldsymbol{\gamma} + v_{git}, \quad (\text{A.185})$$

där feltermen nu antas bestå av den sammansatta feltermen

$$v_{git} = c_g + c_{gi} + f_t + u_{git}. \quad (\text{A.186})$$

Som tidigare är c_g klustereffekter (ex. företag/industri), c_{gi} är inom-klustrereffekter som är konstanta över tid (ex. arbetare inom företag/företag inom industri), f_t är tidschocker gemensamma för samtliga individer i data och u_{git} antas vara idiosynkratiska fel. \mathbf{r}_g är en $K \times 1$ vektor, \mathbf{x}_{gi} är en $H \times 1$ vektor av variabler som varierar över g och i och \mathbf{z}_{git} består av L variabler som kan variera över olika nivåer. Vissa av dessa varierar kanske bara över g och t , och inte mellan enheterna. Vi har således en total mängd av $P = K + H + L$ variabler i modellen. Modellen används ofta vid utvärdering av policyförändringar och benämns i det sammahanget för en skillnad-i-skillnader-ansats (se boken, avsnitt 12.5, för en utförlig diskussion av denna analys eller design). Ofta är det också rimligt att tänka sig att en term i feltermen kan skrivas f_{gt} , vilket innebär chocker som är specifika för kluster och tid. Närvaro av sådana chocker kan komplicera vissa analyser beroende på frågeställning. Vi kommer bortse från sådana chocker här.

För varje individ låt $\mathbf{y}_{gi} = [y_{gi1} \dots y_{giT}]'$ vara en $T \times 1$ vektor för utfallen och låt $\mathbf{W}_{gi} = [\mathbf{1} \quad \mathbf{r}_g \quad \mathbf{x}_{gi} \quad \mathbf{z}_{git}]$ vara en $T \times (P+1)$ matris för de oberoende variablerna för individ i . För varje individ i i en grupp g kan vi nu skriva ekvationen på klusternivå:

$$\mathbf{y}_g = \mathbf{W}_g \boldsymbol{\theta} + \mathbf{v}_g, \quad (\text{A.187})$$

där

$$\mathbf{y}_g = \begin{bmatrix} \mathbf{y}_{g1} \\ \mathbf{y}_{g2} \\ \vdots \\ \mathbf{y}_{gn_g} \end{bmatrix}_{n_g T \times 1}, \quad \mathbf{W}_g = \begin{bmatrix} \mathbf{W}_{g1} \\ \mathbf{W}_{g2} \\ \vdots \\ \mathbf{W}_{gn_g} \end{bmatrix}_{n_g T \times (P+1)}, \quad \boldsymbol{\theta} = \begin{bmatrix} \alpha \\ \boldsymbol{\delta} \\ \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{bmatrix}_{(P+1) \times 1}, \quad \mathbf{v}_g = \begin{bmatrix} \mathbf{v}_{g1} \\ \mathbf{v}_{g2} \\ \vdots \\ \mathbf{v}_{gn_g} \end{bmatrix}_{n_g T \times 1}. \quad (\text{A.188})$$

Hur snabbt den polade OLS-estimatorn konvergerar beror på de ingående delarna i feltermen. Feltermen har en gruppdel (c_g), en individdel (c_{gi}), en tidsdel (f_t) och en del som varierar på både individ och tidsnivå (u_{git}). Som tidigare är det enklaste sättet att hantera f_t att ta med en full uppsättning dummyvariabler för tid. Av de övriga feltermer är c_g den som gör att OLS-estimatorn konvergerar långsammast. Detta då c_{gi} och u_{git} är nestade inom c_g . På samma sätt som ovan gäller då att OLS-estimatorn konvergerar med G (antal kluster).

På samma sätt är den polade OLS-estimatorn normalfördelad asymptotiskt i G . För att inferensen ska bli korrekt i närvaro av c_{gi} och c_g kan vi använda en kluster-robust variansestimator

som tillåter för godtycklig korrelation inom kluster över tid:

$$\hat{V}^c(\hat{\theta}) = \left(\sum_{g=1}^G \mathbf{W}_g' \mathbf{W}_g \right)^{-1} \left(\sum_{g=1}^G \mathbf{W}_g' \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \mathbf{W}_g \right) \left(\sum_{g=1}^G \mathbf{W}_g' \mathbf{W}_g \right)^{-1}. \quad (\text{A.189})$$

där $\hat{\mathbf{v}}_g = \mathbf{y}_g - \mathbf{W}_g \hat{\theta}$.

I den statistiska literaturen kallas modeller med feltermsstruktur enligt ekvation (A.186) ofta för en hierarkisk (*hierarchical*) linjär modell (HLM), blandmodell (*mixed model*), eller flernivåmodell (*multilevel model*). Inom denna litteratur kallas ofta grupp-nivån för den första nivån, medan individ-nivån för den andra nivån. Inom denna litteratur estimeras modellen med generaliserad minsta kvadratmetod (GLS), *maximum likelihood* eller Bayesiansk metod. När det gäller GLS eller maximum likelihood görs detta genom att explicit inkorporera korrelationsstruktur som ges av modellen i ekvation (A.186). Detta kan ge mer effektiva skattningar av θ , men kräver att korrelationsstrukturen är korrekt specificerad. GLS diskuteras vidare i Bilaga B.

A.7.3.1 Inferens rörande enbart γ

Om vi enbart är intresserad av inferens för γ kan vi ta bort $c_g + c_{gi}$ genom att bilda inom-individ-differenser. Vi kan ta fram medelvärdet av utfallet för individ i i grupp g över tid som

$$\bar{y}_{gi} = \alpha + \mathbf{r}_g' \boldsymbol{\delta} + \mathbf{x}_{gi}' \boldsymbol{\beta} + \bar{\mathbf{z}}_{gi}' \boldsymbol{\gamma} + c_g + c_{gi} + \bar{f} + \bar{u}_{gi}, \quad (\text{A.190})$$

där $\bar{y}_{gi} = \frac{1}{T} \sum_{t=1}^T y_{git}$, $\bar{\mathbf{z}}_{gi} = \frac{1}{T} \sum_{t=1}^T \mathbf{z}_{git}$, $\bar{f} = \frac{1}{T} \sum_{t=1}^T f_t$ och $\bar{u}_{gi} = \frac{1}{T} \sum_{t=1}^T u_{git}$.

Genom att dra ifrån (A.190) från (A.185) erhålls

$$y_{git} - \bar{y}_{gi} = \boldsymbol{\gamma}'(\mathbf{z}_{git} - \bar{\mathbf{z}}_{gi}) + f_t - \bar{f} + u_{git} - \bar{u}_{gi}. \quad (\text{A.191})$$

Vi kan nu skatta γ på analogt sätt till det som angavs i ekvation (A.180) eller (A.182). Dvs. antingen skattas en modell med dummyvariabler för både individ och tid eller så skapas nya variabler där medelvärde för individ dragits bort och så skattas en modell med dummyvariabler för tid. Med en balanserad panel är dessa estimatorer ekvivalenta. Vi kan nu skatta variansen med den vanliga kovariansmatrisen, eller med den robusta EHW-matrisen om vi misstänker heteroskedasticitet.

Om vi tror att det finns seriekorrelation i feltermen över tid inom individ så kan vi använda den klusterrobusta kovariansmatrisen där varje individ är ett kluster, vilken är asymptotiskt korrekt i n . Om vi dessutom tror att det finns korrelationer inom grupp g för både de oberoende variablerna och feltermen bör vi i stället använda den klusterrobusta kovariansmatrisen där varje grupp g är ett kluster. Denna senare kovariansmatris är den som är mest tillåtande (dvs. kräver minst antaganden), men den är enbart valid asymptotiskt i G och kan ha dåliga egenskaper om G är litet.

Bilaga B: Generaliserad minsta kvadratmetoden (GLS)

Kapitlet börjar med att introducera GLS som en metod för analys av tvärsnittdata. Det är dock framförallt med klustrade data eller paneldata som GLS-estimatorn är användbar, något som diskuteras i avsnitt B.2.

B.1 GLS och tvärsnittsdata

Vi har som tidigare en tanke en populationsregression:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \dots + \beta_K X_K + U \\ &= \mathbf{x}'\boldsymbol{\beta} + U. \end{aligned} \tag{B.1}$$

För ett slumpmässigt stickprov om n individer definierades OLS-estimatorn i ekvation (A.5) som

$$\hat{\boldsymbol{\beta}} = \arg \min_{\mathbf{b}} \text{RSS}(\mathbf{b}), \tag{B.2}$$

där

$$\text{RSS}(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}). \tag{B.3}$$

Notera att vi kan skriva om denna residualkvadratsumma som

$$\text{RSS}(\mathbf{b}) = (\mathbf{y} - \mathbf{X}\mathbf{b})'\mathbf{I}_n(\mathbf{y} - \mathbf{X}\mathbf{b}), \tag{B.4}$$

där vi lägger till \mathbf{I}_n utan att förändra uttrycket. Vi kan ersätta identitetsmatrisen med en annan matris, \mathbf{V} , i vilket fall vi erhåller en viktad residualkvadratsumma. Den *generaliserade minsta kvadratmetoden*, GLS (eng. *generalized least squares*) ger GLS-estimatorn

$$\hat{\boldsymbol{\beta}}_{GLS} = \arg \min_{\mathbf{b}} Q(\mathbf{b}), \tag{B.5}$$

där

$$Q(\mathbf{b}) = (\mathbf{Y} - \mathbf{X}\mathbf{b})'\mathbf{V}(\mathbf{Y} - \mathbf{X}\mathbf{b}). \tag{B.6}$$

\mathbf{V} är en $n \times n$ symmetrisk matris av vikter som, för tillfället, är fixa. Om \mathbf{V} är en diagonal matris med olika vikter är GLS en viktad minsta kvadratmetod som kallas WLS (eng. *weighted least squares*).

Precis som med OLS kan vi hitta det \mathbf{b} som minimerar $Q(\mathbf{b})$ genom att derivera $Q(\mathbf{b})$ med avseende på \mathbf{b} och sätta derivatan till noll:

$$-2\mathbf{X}'\mathbf{V}(\mathbf{Y} - \mathbf{X}\mathbf{b}) = 0. \quad (\text{B.7})$$

Denna ekvation har en explicit lösning förutsatt att $\mathbf{X}'\mathbf{V}\mathbf{X}$ kan inverteras:

$$\hat{\beta}_{GLS} = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{y}. \quad (\text{B.8})$$

Om vi ersätter \mathbf{y} med $\mathbf{X}\beta + \mathbf{u}$, kan detta skrivas:

$$\hat{\beta}_{GLS} = \beta + (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{u}. \quad (\text{B.9})$$

Vi har att

$$\text{plim}(\hat{\beta}_{GLS}) = \beta + (\text{plim}(\mathbf{X}'\mathbf{V}\mathbf{X}/n))^{-1} \text{plim}(\mathbf{X}'\mathbf{V}\mathbf{u}/n). \quad (\text{B.10})$$

Estimatoren är alltså konsistent om $\text{plim}(\mathbf{X}'\mathbf{V}\mathbf{u}/n) = 0$. Frågan är när detta gäller? Ett tillräckligt villkor är om $E(U|\mathbf{x}) = 0$, dvs. sambandet mellan Y och \mathbf{x} är linjärt. Detta är dock ett starkare villkor än vad som behövs. För ett slumpmässigt stickprov räcker det att $E(\mathbf{X}'\mathbf{V}\mathbf{u}/n) = \mathbf{0}$.¹¹

Givet att $E(\mathbf{X}'\mathbf{V}\mathbf{X}/n)$ existerar och är inverterbar och $E(\mathbf{X}'\mathbf{V}\mathbf{u}/n) = \mathbf{0}$ gäller att:

$$\mathbf{X}'\mathbf{V}\mathbf{u}/\sqrt{n} \xrightarrow{d} \mathcal{N}(\mathbf{0}, E(\mathbf{X}'\mathbf{V}\mathbf{X}/n)), \quad (\text{B.11})$$

vilket innebär att

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (E(\mathbf{X}'\mathbf{V}\mathbf{X}/n))^{-1} E(\mathbf{X}'\mathbf{V}\mathbf{u}\mathbf{u}'\mathbf{V}'\mathbf{X}/n)(E(\mathbf{X}'\mathbf{V}\mathbf{X}/n))^{-1}). \quad (\text{B.12})$$

Vi definierar matrisen Ω som en matris av dimension $n \times n$ på följande sätt:

$$\Omega = E(\mathbf{u}\mathbf{u}'|\mathbf{X}). \quad (\text{B.13})$$

Eftersom \mathbf{V} är fix kan ekvation (B.12) kan nu skrivas

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (E(\mathbf{X}'\mathbf{V}\mathbf{X}/n))^{-1} E(\mathbf{X}'\mathbf{V}\Omega\mathbf{V}'\mathbf{X}/n)(E(\mathbf{X}'\mathbf{V}\mathbf{X}/n))^{-1}), \quad (\text{B.14})$$

där $E(\mathbf{X}'\mathbf{V}\Omega\mathbf{V}'\mathbf{X}/n) = E(E(\mathbf{X}'\mathbf{V}\mathbf{u}\mathbf{u}'\mathbf{V}'\mathbf{X}/n)|\mathbf{X}) = E(\mathbf{X}'\mathbf{V}E(\mathbf{u}\mathbf{u}'|\mathbf{X})\mathbf{V}'\mathbf{X}/n)$. Notera att om \mathbf{X} också är fix kan detta skrivas

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (\mathbf{X}'\mathbf{V}\mathbf{X}/n)^{-1}(\mathbf{X}'\mathbf{V}\Omega\mathbf{V}'\mathbf{X}/n)(\mathbf{X}'\mathbf{V}\mathbf{X}/n)^{-1}). \quad (\text{B.15})$$

Vi har hittills inte sagt något om varför vi vill använda en matris \mathbf{V} i stället för \mathbf{I}_n (dvs. OLS). Ett svar är att vi kan få en mer effektiv estimator. Normalt brukar man med GLS låta $\mathbf{V} = \Omega^{-1}$. I så fall så förenklas ekvation (B.14) till

$$\sqrt{n}(\hat{\beta}_{GLS} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (E(\mathbf{X}'\Omega^{-1}\mathbf{X}/n))^{-1}). \quad (\text{B.16})$$

¹¹Detta gäller om vi vill göra inferens till β så som den definierats utifrån kriteriet i ekvation (A.26). I bland är det dock mer naturligt att göra inferens till en parameter som definieras som populationsanalogen till ekvation (B.5). Så länge inget annat anges kommer vi dock här utgå från att β definieras enligt ekvation (A.26).

B.1.1 Maximum Likelihood

Ett alternativ till skattningar med GLS är att skatta med det som kallas *Maximum Likelihood* (ML). Med ML modelleras fördelningen på feltermen (och därmed Y betingat på \mathbf{x}). Vi går inte igenom hur detta görs i denna bilaga, men vi kan konstatera att under antagandet om att U är normalfördelad går det att ta fram ML-estimatorn, $\hat{\beta}_{ML}$. Genom att låta $\mathbf{V} = \mathbf{\Omega}^{-1}$, så är $\hat{\beta}_{GLS} = \hat{\beta}_{ML}$, och då har ML-estimatorn egenskaper som ges i ekvation (B.16). Denna egenskap gäller oavsett om U är normalfördelad eller ej.¹² Detta resultat är grunden för det som kallas den *generaliserade linjära modellen* (GLM, eng. *generalized linear model*). För samtliga GLS-estimatorer i denna bilaga går det också att ta fram ML-estimatorer under samma feltermsspecifikation, $\mathbf{\Omega}$, om U antas vara normalfördelad. Detta antagande är det mest vanligt förekommande vid ML-estimation, men vid utfall som är binära, frekvenser, etc. är andra feltermsantaganden mer lämpliga (exempelvis Bernoullifördelning för binära utfall och Poissonfördelning för frekvenser). När man vill ta hänsyn till korrelerade observationer med dessa typer av utfall säger man som regel att man skattar "GLM mixed models". Detta är inget som vi diskuterar mer i denna bilaga.

B.1.2 Viktning

GLS-estimatorn kan skrivas om som en vanlig OLS-estimator men med transformerade variabler där dessa bestäms från $\mathbf{\Omega}$. Detta ger en intuition till hur estimatorn fungerar. Dessutom används språkbruket med vikter i termer av dessa transformationer i litteraturen, varför det är motiverat att diskutera dessa här.

Antag först att $\mathbf{\Omega}$ är känd. Förutsatt att $\mathbf{\Omega}$ är inverterbar kan vi skriva följande regression

$$\mathbf{\Omega}^{-1/2}\mathbf{y} = \mathbf{\Omega}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{\Omega}^{-1/2}\mathbf{u}, \quad (\text{B.17})$$

Här är $\mathbf{\Omega}^{-1/2}$ en symmetrisk, inverterbar matris sådan att $\mathbf{\Omega}^{-1/2}\mathbf{\Omega}^{-1/2} = \mathbf{\Omega}^{-1}$ och $\mathbf{\Omega}^{-1/2}\mathbf{\Omega}\mathbf{\Omega}^{-1/2} = \mathbf{I}_n$. Låt $\tilde{\mathbf{u}} = \mathbf{\Omega}^{-1/2}\mathbf{u}$, vilket ger

$$\text{E}(\tilde{\mathbf{u}}\tilde{\mathbf{u}}'|\mathbf{X}) = \mathbf{\Omega}^{-1/2}\text{E}(\mathbf{u}\mathbf{u}'|\mathbf{X})\mathbf{\Omega}^{-1/2} = \mathbf{I}_n. \quad (\text{B.18})$$

Låt $\tilde{\mathbf{X}} = \mathbf{\Omega}^{-1/2}\mathbf{X}$ och $\tilde{\mathbf{y}} = \mathbf{\Omega}^{-1/2}\mathbf{y}$. Regressionen kan nu skrivas som

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} + \tilde{\mathbf{u}}. \quad (\text{B.19})$$

Detta innebär att det är ekvivalent att skatta $\boldsymbol{\beta}$ med GLS enligt ekvation (B.8) som att skatta $\boldsymbol{\beta}$ med OLS med transformerade data $\tilde{\mathbf{y}}$ och $\tilde{\mathbf{X}}$.

Om $\text{E}(U|\mathbf{x}) = 0$ och $\mathbf{V} = \mathbf{\Omega}^{-1}$ så är GLS-estimatorn den bästa linjära väntevärdesriktiga estimatorn. Bevis följer direkt från det tidigare beviset för Gauss-Markov-teoremet (se ekvation A.51) men med $\tilde{\mathbf{y}}$ och $\tilde{\mathbf{X}}$ i stället för \mathbf{y} och \mathbf{X} .

I de flesta fall är $\mathbf{\Omega}$ dock okänd och omöjlig att skatta ickeparametriskt. Således om vi vill använda oss av GLS måste detta byggas på antaganden om hur kovariansstrukturen ser ut.

¹²ML-estimatorer som är konsistenta under felspecifikation av fördelning på feltermen kallas för *pseudo maximum likelihood* (PML)-estimatorer. Om Y antas komma från den generaliserade linjära familjen och det betingade väntvärdet är korrekt specificerad – t.ex. $\text{E}(Y|\mathbf{x}) = g(\mathbf{x}'\boldsymbol{\beta})$, där $g(\cdot)$ är känd – är PML-estimatorn konsistent för $\boldsymbol{\beta}$.

Innebörden är att vi behöver en modell för Ω eller att vi på lägger restriktioner som gör att Ω kan skattas från data.

B.1.3 Möjlig GLS

Vid möjlig GLS (eng. *feasible GLS*, förkortad FGLS) används en skattare av V , \hat{V} , i stället för Ω^{-1} . FGLS-estimatoren är då

$$\hat{\beta}_{FGLS} = (X' \hat{V} X)^{-1} X' \hat{V} y. \quad (B.20)$$

Om $\text{plim}(\hat{V}) = \Omega^{-1}$ så är FGLS-estimatoren lika effektiv som GLS-estimatoren asymptotiskt (med egenskaper som ges i ekvation B.16). En naturlig variansskattare skulle då vara $\hat{V}(\hat{\beta}_{FGLS}) = (X' \hat{V} X)^{-1}$. Ibland har man dock en situation när \hat{V} inte konvergerar till Ω^{-1} , men där det finns ett proportionerligt samband sådant att för en konstant $c > 0$ gäller att $\text{plim}(c\hat{V}) = \Omega^{-1}$ (exempel på detta ges nedan). I sådant fall är FGLS-estimatoren i ekvation (B.20) fortfarande valid eftersom $(X' \hat{V} X)^{-1} X' \hat{V} y = (X' c\hat{V} X)^{-1} X' c\hat{V} y$. Däremot så blir $\hat{V}(\hat{\beta}_{FGLS}) \neq (X' \hat{V} X)^{-1}$. Det är dock enkelt att ta fram en korrekt kovariansmatris.

För att se hur detta görs, notera från föregående avsnitt att vi kan skriva GLS-estimatoren i termer av transformerade variabler. Samma sak gäller för FGLS-estimatoren där vi låter $\tilde{y} = \hat{V}^{1/2} y$, $\tilde{X} = \hat{V}^{1/2} X$ och $\tilde{u} = \hat{V}^{1/2} u$. Från ekvation (B.18) vet vi att om $\hat{V} = \Omega^{-1}$ är variansen för \tilde{u} ett. Om $\text{plim}(c\hat{V}) = \Omega^{-1}$ kommer i stället \tilde{u} asymptotiskt vara en oberoende felterm men med varians $\sigma_U^2 = 1/c$. Detta innebär att för en konsistent skattare av σ_U^2 , \hat{s}_U^2 , har vi $\text{plim}(\hat{V}/\hat{s}_U^2) = \Omega^{-1}$.

Låt $\hat{u}_i = \tilde{y}_i - \tilde{x}_i' \hat{\beta}_{FGLS}$. Den naturliga skattaren av den transformerade feltermens varians är

$$\hat{s}_U^2 = \frac{1}{n - (K + 1)} \sum_{i=1}^n \hat{u}_i^2. \quad (B.21)$$

Vi har då

$$\hat{V}(\hat{\beta}_{FGLS}) = \left(X' \left(\hat{V} / \hat{s}_U^2 \right) X \right)^{-1} = \hat{s}_U^2 (X' \hat{V} X)^{-1} = \hat{s}_U^2 (\tilde{X}' \tilde{X})^{-1}, \quad (B.22)$$

där det sista likhetstecknet kommer av att $X' \hat{V} X = X' (\hat{V}^{1/2})' \hat{V}^{1/2} X = \tilde{X}' \tilde{X}$. Detta är alltså den vanliga kovariansmatrisen fast för de transformerade variablerna, vilket är en korrekt kovariansmatris om \hat{V} asymptotiskt är proportionerlig mot Ω^{-1} .

Det är dock inte alltid säkert att V är proportionerlig mot Ω . Om det är fallet att $E(X' V u) = 0$, kan dock korrekt inferens göras så länge $\text{plim}(\hat{V}) = V$ och de transformerade feltermerna, \tilde{u} , är oberoende även om de inte är homoskedastiska. Vi kan använda en robust kovariansmatris, vilken som vanligt är valid asymptotiskt:

$$\hat{V}^r(\hat{\beta}_{FGLS}) = \frac{n}{n - (K + 1)} (\tilde{X}' \tilde{X})^{-1} \left(\sum_{i=1}^n \hat{u}_i^2 \tilde{x}_i \tilde{x}_i' \right) (\tilde{X}' \tilde{X})^{-1}. \quad (B.23)$$

Om \mathbf{V} är nära $\mathbf{\Omega}^{-1}$ kommer denna variansestimator asymptotiskt vara nära den asymptotiskt effektiva GLS-estimatoren, men även om den är långt ifrån så kommer inferensen bli korrekt asymptotiskt så länge $E(\mathbf{X}'\mathbf{V}\mathbf{u}) = \mathbf{0}$. Precis som med OLS har denna robusta variansestimator inte bra egenskaper vid små stickprov, men vid större stickprov är den väl värd att användas.

FGLS kan användas för att hantera heteroskedasticitet om det går att specificera en modell för hur heteroskedasticiteten ser ut. För att illustrera hur FGLS kan användas använder vi en parametrisk modell där vi antar att $E(U|\mathbf{x}) = 0$ och $E(U^2|\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\gamma})$,¹³ vilket innebär att vi, för ett slumpmässigt draget stickprov, har

$$\mathbf{\Omega} = E(\mathbf{u}\mathbf{u}'|\mathbf{X}) = \begin{bmatrix} \exp(\mathbf{x}'_1\boldsymbol{\gamma}) & 0 & \dots & 0 \\ 0 & \exp(\mathbf{x}'_2\boldsymbol{\gamma}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \exp(\mathbf{x}'_n\boldsymbol{\gamma}) \end{bmatrix}. \quad (\text{B.24})$$

Om vi kände $\boldsymbol{\gamma}$ hade vi kunnat skatta $\boldsymbol{\beta}$ med GLS. Men då denna normalt är okänd måste vi först skatta $\boldsymbol{\gamma}$. Sättet att göra detta på är att först skatta den vanliga regressionsmodellen med \mathbf{y} på \mathbf{X} med OLS och ta fram residualerna, $\hat{\mathbf{u}}$. Därefter kan $\boldsymbol{\gamma}$ skattas i följande regression med OLS:

$$\log(\hat{u}_i^2) = \mathbf{x}'_i\boldsymbol{\gamma} + v_i. \quad (\text{B.25})$$

Från dessa OLS-skattningar, $\hat{\boldsymbol{\gamma}}$, så bildar vi $\tilde{y}_i = y_i/\sqrt{\exp(\mathbf{x}'_i\hat{\boldsymbol{\gamma}})}$ och $\tilde{\mathbf{x}}_i = \mathbf{x}_i/\sqrt{\exp(\mathbf{x}'_i\hat{\boldsymbol{\gamma}})}$.¹⁴ Vi skattar sedan

$$\tilde{y}_i = \beta_0 + \tilde{\mathbf{x}}'_i\boldsymbol{\beta} + \tilde{u}_i, \quad (\text{B.27})$$

med OLS. För inferens använder vi sedan antingen den vanliga eller den robusta kovariansmatrisen. Detta är liktydigt med FGLS enligt ekvation (B.20) och ekvation (B.22) eller ekvation (B.23) för att skatta variansen av estimatoren.

Ett skäl att använda en viktningsmatris är när vi har aggregerade data. Antag att population av individer kan delas in i G olika kluster som är ömsesidigt utslutande. Låt y_{gi}, \mathbf{x}_{gi} vara observerad data för individ i i kluster g , där data kommer från ett obundet slumpmässigt urval från populationen inom respektive kluster. Sedan tidigare vet vi att OLS är en konsistent estimator av $\boldsymbol{\beta}$.

¹³Se Harvey (1976) för en diskussion och mer generell variant av denna modell.

¹⁴Notera att OLS-estimatoren är dock inte konsistent för koefficienten framför konstanttermen i \mathbf{x} , γ_0 . För att se detta, notera att vi har

$$\begin{aligned} E(U_i|\mathbf{x}_i) &= \exp(\mathbf{x}'_i\boldsymbol{\gamma}) \Leftrightarrow \\ \log(E(U_i^2|\mathbf{x}_i)) &= \mathbf{x}'_i\boldsymbol{\gamma} \Leftrightarrow \\ \log(E(U_i^2|\mathbf{x}_i)) + \log(\hat{u}_i^2) &= \mathbf{x}'_i\boldsymbol{\gamma} + \log(\hat{u}_i^2) \Leftrightarrow \\ \log(\hat{u}_i^2) &= \mathbf{x}'_i\boldsymbol{\gamma} + \log(\hat{u}_i^2) - \log(E(U_i^2|\mathbf{x}_i)) \Leftrightarrow \\ \log(\hat{u}_i^2) &= \mathbf{x}'_i\boldsymbol{\gamma} + \log(\hat{u}_i^2/E(U_i^2|\mathbf{x}_i)). \end{aligned} \quad (\text{B.26})$$

Vi har alltså att $v_i = \log(\hat{u}_i^2/E(U_i^2|\mathbf{x}_i))$. Om u_i är normalfördelad går det att visa att v_i asymptotiskt följer en fördelning som en logaritmerad χ^2 -fördelad variabel med en frihetsgrad. En sådan har ett väntevärde som skiljer sig mot noll vilket innebär att $\text{plim}(\hat{\gamma}_0) \neq \gamma_0$. Övriga parametrar i $\boldsymbol{\gamma}$ går dock att skatta konsistent. Detta innebär i sin tur att $\text{plim}(c\hat{\mathbf{V}}) = \mathbf{\Omega}^{-1}$ för någon konstant c .

I stället för dessa individdata har vi nu tillgång till genomsnittet för varje grupp enligt: $\bar{y}_g = \sum_{i=1}^{n_g} y_{gi}$ och $\bar{x}_g = \sum_{i=1}^{n_g} x_{gi}$, där n_g är antalet individer i kluster g i stickprovet. Om vi nu skattar

$$\bar{y}_g = \bar{x}_g \beta + \bar{u}_g, \quad (\text{B.28})$$

där $\bar{u}_g = \sum_{i=1}^{n_g} u_{gi}$, så är OLS-estimatoren med aggregerade data fortsatt valid för inferens till population förutsatt att antal kluster G är stort. Om vi utgår från att feltermen är homoskedastisk och oberoende på individnivå (sfäriska feltermen) med varians σ_U^2 , så kommer $V(\bar{u}_g | n_g) = \sigma_U^2 / n_g$. Detta innebär att om klustren är olika stora ($n_g \neq n_{g'}$ för $g \neq g'$) har vi heteroskedasticitet i de aggregerade data. Men under antagandet om sfäriska feltermen vet vi formen på Ω (en matris av dimension $G \times G$ då vi använder aggregerade data), vilken alltså är

$$\Omega = \begin{bmatrix} \sigma_U^2/n_1 & 0 & \dots & 0 \\ 0 & \sigma_U^2/n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_U^2/n_G \end{bmatrix} = \sigma_U^2 \begin{bmatrix} 1/n_1 & 0 & \dots & 0 \\ 0 & 1/n_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/n_G \end{bmatrix}. \quad (\text{B.29})$$

Ω är okänd eftersom vi inte känner σ_U^2 . Men då denna är en konstant så känner vi en matris som är proportionerlig mot Ω , vilket innebär att vi kan använda GLS (vilket är WLS i detta fall eftersom Ω är en diagonal matris; FGLS behöver alltså inte användas). Som tidigare kan vi transformera variablerna genom att dela med roten av diagonalelementen av viktningssmatrisen. De transformerade variablerna blir då $\tilde{y}_g = y_g / (1/\sqrt{n_g}) = y_g \sqrt{n_g}$ och $\tilde{x}_g = x_g / (1/\sqrt{n_g}) = x_g \sqrt{n_g}$. Därefter kan en regression skattas med dessa transformerade variabler med OLS.¹⁵ När denna modell skattas säger man som regel att man viktar med stickprovsstorleken inom respektive grupp. Det som avses då är alltså att varje observation multipliceras med $\sqrt{n_g}$.¹⁶ Notera att om variansen på individnivå inte är homoskedastisk så ska den robusta variansestimatören användas vid inferens.

Notera att detta resultat bygger på att stickprovsmedelvärdena \bar{y}_g och \bar{x}_g är konsistenta skattningar av populationsmedelvärde i grupperna. Att så är fallet här kommer av att vi har antagit att data är obundet slumpmässigt draget från populationen inom respektive kluster. Om så inte är fallet kommer WLS generellt *inte* vara en konsistent estimator av β .

Många gånger är det inte uppenbart när eller hur man ska vikta så för att illustrera problemet låt oss ta ett exempel. Antag att vi har ett slumpmässigt urval av skolor i Sverige och att vi har observerat genomsnittligt betyg för dessa \bar{y}_g . Vi observerar om skolorna är privata $x_g = 1$ eller

¹⁵Variansskattningen ska då göras enligt principerna som diskuterades ovan, se ekvation (B.22) eller ekvation (B.23)

¹⁶Det faktum att man säger att varje observation vikts med stickprovsstorleken avspeglas också i syntaxen för WLS i R. Om vi har aggregerad data i en dataram `df` med utfallsvariabel `y`, oberoende variabler `x1` och `x2`, samt en variabel `group_size` som innehåller antalet individer per kluster, så är syntaxen för att skatta den aggregerade modellen med WLS i R: `lm(y ~ x1 + x2, data=df, weights=group_size)`. Detta ger alltså samma resultat som om vi multiplicerar samtliga ingående variabler (inklusive konstanttermen) med `sqrt(df$group_size)` och sedan skattar en vanlig, oviktad, regression med de transformerade variablerna.

kommunala $x_g = 0$ och vi skattar följande modell med OLS:

$$\bar{y}_g = \beta_0 + \beta_1 x_g + u_g. \quad (\text{B.30})$$

Låt säga att vi finner att $\hat{\beta}_1 < 0$ och att vi har en precis skattning. Från denna analys kan vi då säga att i förväntan är betygen lägre i privata skolor än i kommunala.

Antag nu att vi också har data på antal elever för varje skola. Vi tänker då kanske att om vi skattar modellen med WLS så kan vi göra inferens till populationen av elever. Om resultatet skulle kvarstå så skulle vi då säga: ”elever som går i privata skolor har i förväntan lägre betyg än de som går i kommunala”. Detta är dock inte nödvändigtvis en korrekt slutsats. Skälet är att vi inte har ett slumpmässigt stickprov av elever utan endast av skolor. Detta innebär att vi inte kan göra inferens till eleverna. Med ett slumpmässigt urval av skolor kommer vi att få med elever från mindre skolor i mycket högre utsträckning än om vi hade ett slumpmässigt stickprov av elever. Men om vi känner till storleken på skolorna kan vi använda en viktningsestimator för att göra inferens till elevpopulationen. Hur detta kan gå till diskuteras i avsnitt B.3.

B.2 Paneldata och klustrade data

Vi såg tidigare i fallet med paneldata och klustrade data att det kan vara naturligt att tro att det finns beroende mellan observationer. Om det går att specificera hur detta beroende ser ut kan det i många fall vara möjligt att skatta Ω vilket möjliggör skattning med FGLS med dessa typer av data.

B.2.1 Paneldata och modeller med slumpmässiga intercept

Vi utgår från samma modell som introducerades i avsnitt A.7.2:

$$y_{it} = \alpha + \mathbf{x}'_i \boldsymbol{\beta} + \mathbf{z}'_{it} \boldsymbol{\gamma} + v_{it}. \quad (\text{B.31})$$

Denna modell kunde också skrivas i vektorform som

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\lambda} + \mathbf{v}_i. \quad (\text{B.32})$$

där $\mathbf{W}_i = [\mathbf{1} \quad \mathbf{X}_i \quad \mathbf{Z}_i]$. Feltermen v_{it} bestod av tre delar (se ekvation A.166): c_i var en term som var konstant för en individ över tid men som varierade mellan individer, f_t var en term som var konstant för alla individer men som varierade över tid, medan u_{it} var en term som var helt slumpmässigt (och därmed varierade både inom individ över tid och mellan individer). För enkelhets skull kommer vi ignorera termen f_t , exempelvis för att vi tar med en full uppsättning dummyvariabler över tid. Feltermen blir då $v_{it} = c_i + u_{it}$. Vi kommer utgå från linjäritet så att $E(C|\mathbf{x}) = E(U|\mathbf{x}) = 0$.

För att kunna använda FGLS måste Ω specificeras, dvs. strukturen för v_{it} . Låt $\Omega_{ii} = E(\mathbf{u}_i \mathbf{u}'_i | \mathbf{W})$, vilken alltså är en matris av dimension $T \times T$. I den vanligaste modellen – en modell med

slumpmässiga intercept (eng. *random effects*) – har denna matris följande utseende:

$$\mathbf{\Omega}_{ii} = \begin{bmatrix} \sigma_C^2 + \sigma_U^2 & \sigma_C^2 & \dots & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 + \sigma_U^2 & \dots & \sigma_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_C^2 & \sigma_C^2 & \dots & \sigma_C^2 + \sigma_U^2 \end{bmatrix}. \quad (\text{B.33})$$

Denna matris skrivs ibland – helt ekvivalent – som

$$\mathbf{\Omega}_{ii} = (\sigma_C^2 + \sigma_U^2) \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}, \quad (\text{B.34})$$

där $\rho = \sigma_C^2 / (\sigma_C^2 + \sigma_U^2)$. ρ är alltså en parameter som beskriver hur stor andel av feltermens variation som kan tillskrivas det som är konstant över tid för en individ.

Det faktum att $\mathbf{\Omega}_{ii}$ har detta utseende innehåller flera antaganden. För det första säger den att U_{it} har samma varians vid alla tidpunkter, σ_U^2 , och dessutom att det inte finns något beroende mellan U_{it} och U_{is} på det sätt att $E(U_{it}U_{is}|\mathbf{W}_i) = 0$ för $t \neq s$. Slutligen är beroendet inom en individ mellan olika tidpunkter samma, nämligen $E(V_{it}V_{is}|\mathbf{W}_i) = E(C_iC_i|\mathbf{W}_i) = \sigma_C^2$ för $t \neq s$. Det är viktigt att notera att dessa antaganden om $\mathbf{\Omega}_{ii}$ är starka. Exempelvis så utesluter den att feltermerna som ligger nära varandra i tid har starkare beroende än feltermer som ligger längre ifrån varandra.

Vidare, om vi antar att observationer för olika individer är oberoende,¹⁷ vilket innebär att $\mathbf{\Omega}_{ij} = E(\mathbf{u}_i\mathbf{u}_j'|\mathbf{W}) = \mathbf{0}$ så innebär att vi har en blockdiagonal matris $\mathbf{\Omega} = E(\mathbf{u}\mathbf{u}'|\mathbf{W})$ med dimension $nT \times nT$ på följande sätt:

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Omega}_{nn} \end{bmatrix}. \quad (\text{B.35})$$

Då $\mathbf{\Omega}_{11} = \mathbf{\Omega}_{22} = \dots = \mathbf{\Omega}_{nn}$ innehåller $\mathbf{\Omega}$ bara två parametrar som ska skattas: σ_C^2 och σ_U^2 .

Dessa parametrar skattas på liknande sätt som beskrevs i föregående avsnitt. Dvs. först skattas λ i ekvation (B.32) med OLS. Därifrån erhålls residualerna \hat{v}_{it} . Därifrån är det rättframt att

¹⁷Vi trycker återigen på att det ofta finns tidseffekter, f_t , som gör att detta oberoende inte finns. Men det är enkelt att ta hand om dessa genom att ta med dummyvariabler för tid, varför vi bortser från dessa här.

skatta σ_V^2 som

$$\hat{\sigma}_V^2 = \frac{\sum_{i=1}^n \sum_{t=1}^T \hat{v}_{it}^2}{nT - (K + L + 1)} = \frac{\text{RSS}}{nT - (K + L + 1)}, \quad (\text{B.36})$$

där \hat{v}_{it} är residualer från den polade OLS skattningen. Estimatern är konsistent i n för fix T . σ_c^2 skattas med

$$\hat{\sigma}_C^2 = \frac{\sum_{i=1}^n \sum_{t=1}^{T-1} \sum_{s=t+1}^T \hat{v}_{it} \hat{v}_{is}}{nT(T-1)/2 - (K + L + 1)}. \quad (\text{B.37})$$

Även denna estimator är konsistent i n för fix T . Vi ersätter nu σ_U^2 och σ_C^2 med $\hat{\sigma}_U^2$ och $\hat{\sigma}_C^2$ i ekvation (B.33):

$$\hat{\Omega}_{ii} = \begin{bmatrix} \hat{\sigma}_C^2 + \hat{\sigma}_U^2 & \hat{\sigma}_C^2 & \dots & \hat{\sigma}_C^2 \\ \hat{\sigma}_C^2 & \hat{\sigma}_C^2 + \hat{\sigma}_U^2 & \dots & \hat{\sigma}_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_C^2 & \hat{\sigma}_C^2 & \dots & \hat{\sigma}_C^2 + \hat{\sigma}_U^2 \end{bmatrix}, \quad (\text{B.38})$$

och

$$\hat{\Omega} = \begin{bmatrix} \hat{\Omega}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \hat{\Omega}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \hat{\Omega}_{nn} \end{bmatrix}, \quad (\text{B.39})$$

där $\hat{\Omega}_{11} = \hat{\Omega}_{22} = \dots = \hat{\Omega}_{nn}$. FGLS-estimatern, som här brukar kallas för *random effects-estimatern* (RE), är då

$$\hat{\lambda}_{RE} = (\mathbf{W}'\hat{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\Omega}^{-1}\mathbf{y}. \quad (\text{B.40})$$

Ibland kan denna procedur itereras. Innebörden är att nya residualer, $\hat{\mathbf{v}} = \mathbf{y} - \mathbf{W}\hat{\lambda}_{RE}$, tas för att estimeras Ω på nytt. Denna process kan sedan pågå fram till dess att skattningen av λ konvergerar. Detta kan möjligtvis förbättra egenskaper för RE-estimatern i små stickprov, men man kan visa teoretiskt (dvs. för stora stickprov) att det inte finns någon effektivitetsvinst med denna iterativa procedur.

Kovariansmatrisen för $\hat{\lambda}_{RE}$ kan skattas som

$$\hat{\mathbf{V}}(\hat{\lambda}_{RE}) = \frac{nT}{nT - (K + L + 1)} (\mathbf{W}'\hat{\Omega}^{-1}\mathbf{W})^{-1}. \quad (\text{B.41})$$

Det finns dock flera skäl till att göra robust inferens även i detta fallet. Anledningen är att Ω kanske inte har den struktur som antas. Detta kan antingen bero på förekomsten av autokorrelation inom individer över tid så att Ω_{ii} inte har nollor utanför diagonalen eller på att det finns heteroskedasticitet så att elementen i diagonalen av Ω inte är konstanta. Den klusterrobusta kovariansmatrisen för paneldata som introducerades i avsnitt A.7.3 – fast där

den anpassats för random effects-estimatoren – kan nu användas. Denna tillåter för godtycklig autokorrelation inom individer och heteroskedasticitet och är

$$\hat{V}^c(\hat{\lambda}_{RE}) = \left(\sum_{i=1}^n \mathbf{W}_i' \hat{\Omega}^{-1} \mathbf{W}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{W}_i' \hat{\Omega}^{-1} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \hat{\Omega}^{-1} \mathbf{W}_i \right) \left(\sum_{i=1}^n \mathbf{W}_i' \hat{\Omega}^{-1} \mathbf{W}_i \right)^{-1}, \quad (\text{B.42})$$

där $\hat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{W}_i \hat{\lambda}_{RE}$ är residualerna från RE-estimatoren (och alltså inte de residualer från OLS-skattningen där $\hat{\Omega}$ erhålls).

B.2.1.1 Test och alternativ GLS-estimator

Normalt kommer $\hat{\sigma}_C^2 > 0$. Till skillnad från en ”vanlig” variansskattning finns dock inga garantier att så är fallet, utan det kan vara fallet att $\hat{\sigma}_C^2 < 0$. Detta kan bero på att $\sigma_C^2 = 0$ (och att $\hat{\sigma}_C^2 < 0$ av slumpen), men det kan också bero på att den antagna strukturen av Ω är felaktig. Det finns också test utvecklade för nollhypotesen att $\sigma_C^2 = 0$ mot alternativet att $\sigma_C^2 > 0$. Ett sådant test är Breusch-Pagan-testet som härleds under antagandet är feltermerna är normalfördelade.

Random effects-modellen är en tämligen restriktiv modell på det sättet att kovariansmatrisen bestäms av enbart två parametrar, σ_C^2 och σ_U^2 . Vi kan dock använda FGLS för att skatta en mer flexibel modell. Vad som behövs är en konsistent estimator av Ω , vilken vi fortsatt antar har den blockdiagonala strukturen i ekvation (B.35). Under antagandet att $\Omega_{11} = \Omega_{22} = \dots = \Omega_{nn}$ går detta att göra utan att vi behöver anta någon särskild struktur för Ω_{ii} (annat än att denna existerar och att Ω är inverterbar). Om vi låter $\hat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{W}_i \hat{\lambda}_{OLS}$ vara residualerna från en OLS-skattning kan vi sedan skatta Ω som

$$\hat{\Omega}_{ii} = \frac{1}{n} \sum_{i=1}^n \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i'. \quad (\text{B.43})$$

Det gäller att $\hat{\Omega}$ är en konsistent estimator av Ω i n . För att se detta, notera att vi kan skriva OLS-residualerna för individ i som

$$\hat{\mathbf{v}}_i = \mathbf{y}_i - \mathbf{W}_i \boldsymbol{\lambda} - \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}) = \mathbf{v}_i - \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}), \quad (\text{B.44})$$

vilket innebär att

$$\hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' = \mathbf{v}_i \mathbf{v}_i' - \mathbf{v}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda})' \mathbf{W}_i' - \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}) \mathbf{v}_i' + \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}) (\hat{\lambda}_{OLS} - \boldsymbol{\lambda})' \mathbf{W}_i'. \quad (\text{B.45})$$

Tillsammans med ekvation (B.43) har vi då

$$\begin{aligned} \hat{\Omega}_{ii} = \frac{1}{n} \sum_{i=1}^n \mathbf{v}_i \mathbf{v}_i' - \frac{1}{n} \sum_{i=1}^n \left(\mathbf{v}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda})' \mathbf{W}_i' - \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}) \mathbf{v}_i' \right) + \\ \frac{1}{n} \sum_{i=1}^n \mathbf{W}_i (\hat{\lambda}_{OLS} - \boldsymbol{\lambda}) (\hat{\lambda}_{OLS} - \boldsymbol{\lambda})' \mathbf{W}_i'. \end{aligned} \quad (\text{B.46})$$

Den första termen konvergerar i sannolikhet till Ω_{ii} när n ökar, medan de andra två termerna konvergerar i sannolikhet till noll eftersom $\hat{\lambda}_{OLS}$ är en konsistent estimator av λ . Därmed har vi att $\hat{\Omega}_{ii}$ är konsistent och det gäller då även att $\hat{\Omega}$ enligt ekvation (B.39) (där $\hat{\Omega}_{ii}$ från ekvation B.46 är blockdiagonalelementen) är konsistent för Ω .

Givet att denna FGLS-estimator är mer flexibel än RE-estimatoren kan man fråga sig varför man någonsin skulle vilja använda den senare estimatoren? Teoretiskt finns ingen alledning att göra detta för stora stickprov. Bägge estimatorer är lika effektiva asymptotiskt om feltermerna följer en RE-struktur, medan om feltermerna inte följer denna är FGLS-estimatoren (dvs. när $\hat{\Omega}_{ii}$ skattas enligt ekvation B.43) asymptotiskt mer effektiv. Sannolikt har det att göra med att man historisk modellerat paneldata med RE-struktur och att det i sig också kan vara intressant att undersöka hur stor $\rho = \sigma_C^2 / (\sigma_C^2 + \sigma_U^2)$ är.

B.2.2 Klustrade data

Som tidigare diskuterats så är strukturen för paneldata och klusterdata densamma. I stället för *individer* har vi G kluster med n_g enheter (individer) i varje kluster $g = 1, \dots, G$. På samma sätt som i föregående avsnitt diskuteras inferens till en population när vi explicit tar hänsyn till den icke observerbara heterogeniteten, c_g , som kommer från dragningar från en supopulation. Strukturen för Ω är samma som med paneldata, men modellen brukar inte kallas för en RE-modell utan för en *flernivåmodell*.

För tydlighets skull, låt oss först repetera ramverket som gavs i avsnitt A.7.1. För ett visst kluster g har vi data $\{(y_{gi}, \mathbf{x}_g, \mathbf{z}_{gm}) : i = 1, \dots, n_g\}$ och totalt har vi G kluster med en total stickprovsstorlek $n = \sum_{g=1}^G n_g$. Vi har att \mathbf{x}_g är en $K \times 1$ vektor av kovariater som enbart varierar mellan grupper, medan \mathbf{z}_{gi} är en $L \times 1$ vektor av kovariater som varierar inom (såväl som över) grupper. Den linjära modellen skrevs

$$y_{gi} = \alpha + \mathbf{x}'_g \boldsymbol{\beta} + \mathbf{z}'_{gi} \boldsymbol{\gamma} + v_{gi}, \quad (\text{B.47})$$

och

$$v_{gi} = c_g + u_{gi}. \quad (\text{B.48})$$

Vi definierade sedan $\mathbf{w}_{gi} = [1 \quad \mathbf{x}'_g \quad \mathbf{z}'_{gi}]'$, $\boldsymbol{\lambda} = [\alpha \quad \boldsymbol{\beta}' \quad \boldsymbol{\gamma}']'$, $\mathbf{y}_g = [y_{g1} \quad y_{g2} \quad \dots \quad y_{gn_g}]'$, $\mathbf{W}_g = [\mathbf{w}'_{g1} \quad \mathbf{w}'_{g2} \quad \dots \quad \mathbf{w}'_{gn_g}]'$ och $\mathbf{v}_g = [v_{g1} \quad v_{g2} \quad \dots \quad v_{gn_g}]'$. Med denna struktur antas det finnas ett beroende mellan individer inom ett kluster (via c_g) men att individerna mellan klustren är oberoende, betingat på kovariaterna \mathbf{x}_g och \mathbf{z}_{gi} .

Vi antar att observationerna i de enskilda klustren $\{y_{gi}, \mathbf{w}_{gi}\}_{i=1}^{n_g}$ är slumpmässigt dragna från en oändlig population inom respektive kluster. Precis som var fallet med paneldata ovan så utgår vi också från ett linjärt samband sådant att $E(C|\mathbf{x}) = E(U|\mathbf{x}) = 0$. Feltermens struktur

inom ett kluster ser identisk ut som feltermens struktur inom individ i fallet med paneldata:

$$\mathbf{\Omega}_{gg} = \begin{bmatrix} \sigma_C^2 + \sigma_U^2 & \sigma_C^2 & \dots & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 + \sigma_U^2 & \dots & \sigma_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_C^2 & \sigma_C^2 & \dots & \sigma_C^2 + \sigma_U^2 \end{bmatrix} = (\sigma_C^2 + \sigma_U^2) \begin{bmatrix} 1 & \rho & \dots & \rho \\ \rho & 1 & \dots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \dots & 1 \end{bmatrix}, \quad (\text{B.49})$$

där $\rho = \sigma_C^2 / (\sigma_C^2 + \sigma_U^2)$. Vi har också

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Omega}_{GG} \end{bmatrix}. \quad (\text{B.50})$$

Först skattar vi $\boldsymbol{\lambda}$ med OLS: $\hat{\boldsymbol{\lambda}}_{OLS} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{y}$, där $\mathbf{W} = [\mathbf{W}'_1 \ \mathbf{W}'_2 \ \dots \ \mathbf{W}'_G]'$ och $\mathbf{y} = [\mathbf{y}'_1 \ \mathbf{y}'_2 \ \dots \ \mathbf{y}'_G]'$. Därifrån erhålls residualerna, \hat{v}_{it} . $\sigma_V^2 = \sigma_C^2 + \sigma_U^2$ skattas sedan som

$$\hat{\sigma}_V^2 = \frac{\sum_{g=1}^G \sum_{i=1}^{n_g} \hat{v}_{gi}^2}{n - (K + L + 1)} = \frac{\text{RSS}}{n - (K + L + 1)}. \quad (\text{B.51})$$

σ_C^2 skattas med

$$\hat{\sigma}_C^2 = \frac{\sum_{g=1}^G \sum_{i=1}^{n_g-1} \sum_{j=i+1}^{n_g} \hat{v}_{gi} \hat{v}_{gj}}{\sum_{g=1}^G (n_g(n_g - 1)/2) - (K + L + 1)}. \quad (\text{B.52})$$

Vi får skattning av $\mathbf{\Omega}$ enligt

$$\hat{\mathbf{\Omega}} = \begin{bmatrix} \hat{\mathbf{\Omega}}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{\Omega}}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \hat{\mathbf{\Omega}}_{GG} \end{bmatrix}, \quad (\text{B.53})$$

där

$$\hat{\mathbf{\Omega}}_{gg} = \begin{bmatrix} \hat{\sigma}_C^2 + \hat{\sigma}_U^2 & \hat{\sigma}_C^2 & \dots & \hat{\sigma}_C^2 \\ \hat{\sigma}_C^2 & \hat{\sigma}_C^2 + \hat{\sigma}_U^2 & \dots & \hat{\sigma}_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}_C^2 & \hat{\sigma}_C^2 & \dots & \hat{\sigma}_C^2 + \hat{\sigma}_U^2 \end{bmatrix}. \quad (\text{B.54})$$

λ kan sedan skattas som

$$\hat{\lambda}_{FNM} = (\mathbf{W}'\hat{\Omega}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\Omega}^{-1}\mathbf{y}, \quad (\text{B.55})$$

där vi använder FNM som förkortning för *flernivåmodell*. Precis som var fallet med paneldata så kan denna procedur itereras. Detta kan möjligtvis förbättra egenskaper för estimatoren i små stickprov men man kan visa teroretiskt (dvs. för stora stickprov) att det inte finns någon effektivitetsvinst med detta.

Precis som var fallet med paneldata finns flera skäl till att göra robust inferens då Ω kan vara felspecificerad. Detta kan bero på att det finns autokorrelation inom grupperna som inte är konstant med σ_C^2 . I fallet med paneldata är det ofta sannolikt att sådan autokorrelation finns, medan det med klusterdata inte är lika uppenbart. Ett exempel när sådan autokorrelation kan finnas med klusterdata är när det finns en geografisk komponent så att vissa individer inom ett kluster ligger närmare varandra än andra individer i samma kluster. Dessutom kan vi ha heteroskedasticitet så att diagonalelementen i Ω inte är konstanta. Den klusterrobusta kovariansmatrisen kan då användas:

$$\hat{V}^c(\hat{\lambda}_{FNM}) = \left(\sum_{i=1}^n \mathbf{w}_i' \hat{\Omega}^{-1} \mathbf{w}_i \right)^{-1} \left(\sum_{i=1}^n \mathbf{w}_i' \hat{\Omega}^{-1} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i' \hat{\Omega}^{-1} \mathbf{w}_i \right) \left(\sum_{i=1}^n \mathbf{w}_i' \hat{\Omega}^{-1} \mathbf{w}_i \right)^{-1}, \quad (\text{B.56})$$

där $\hat{\mathbf{v}}_g = \mathbf{y}_g - \mathbf{W}_g \hat{\lambda}_{FNM}$ är residualerna från FGLS-estimatoren. Summa summarum så är flernivåmodellen som diskuterats i detta avsnitt jämförbar med RE-modellen i föregående avsnitt i nästan alla avseenden. På samma sätt är diskussionen kring test och alternativ FGLS-estimator från avsnitt B.2.1.1 också tillämplig här. Det bör dock tilläggas att flernivåmodeller inte behöver skattas med FGLS, utan dessa modeller kan även skattas med *maximum likelihood* eller med Bayesianska metoder. Det finns också en lång rad olika typer av flernivåmodeller. Vi kommer dock inte att gå in på detta i denna bilaga.

B.2.3 Både klustring och paneldata

I fallet med både klustring och paneldata var vår modell i avsnitt A.7.3

$$y_{git} = \alpha + \mathbf{r}_g' \boldsymbol{\delta} + \mathbf{x}_{gi}' \boldsymbol{\beta} + \mathbf{z}_{git}' \boldsymbol{\gamma} + v_{git}, \quad (\text{B.57})$$

med feltermen

$$v_{git} = c_g + c_{gi} + f_t + u_{git}. \quad (\text{B.58})$$

Det är användbart att definiera följande varianser: $V(C_g) = \sigma_G^2$, $V(C_{gi}) = \sigma_C^2$ och $V(U_{git}) = \sigma_U^2$. Precis som var fallet med paneldata ovan så utgår vi från att f_t inte finns i feltermen, eller att vi tagit om hand om denna med en full uppsättning dummyvariabler för tid. Utfallet för en viss individ i i kluster g definierades som en $T \times 1$ vektor $\mathbf{y}_{gi} = [y_{gi1} \ \dots \ y_{giT}]'$, medan vi hade end $T \times (P+1)$ matris av kovariater $\mathbf{W}_{gi} = [\mathbf{1} \ \mathbf{r}_g \ \mathbf{x}_{gi} \ \mathbf{z}_{git}]$. Vi kunde sedan skriva vår modell på klusternivå (se avsnitt A.7.3 för definitioner av de ingående variablerna) som

$$\mathbf{y}_g = \mathbf{W}_g \boldsymbol{\theta} + \mathbf{v}_g. \quad (\text{B.59})$$

Efter det att vi först estimerat $\boldsymbol{\theta}$ med OLS, $\hat{\boldsymbol{\theta}}_{OLS}$, estimeras Ω och sedan appliceras FGLS.

Beroende på vilka antaganden vi gör kan vi få lite olika strukturer på $\mathbf{\Omega}$. Normalt kommer feltermerna mellan grupper antas vara oberoende så att

$$\mathbf{\Omega} = \begin{bmatrix} \mathbf{\Omega}_{11} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Omega}_{GG} \end{bmatrix}. \quad (\text{B.60})$$

Däremot är frågan hur $\mathbf{\Omega}_{gg}$ modelleras? En möjlighet är att anta en RE-struktur för individer och oberoende mellan individer så att

$$\mathbf{\Omega}_{gg} = \begin{bmatrix} \mathbf{\Omega}_{11}^g & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{\Omega}_{22}^g & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{\Omega}_{n_g n_g}^g \end{bmatrix}, \text{ där } \mathbf{\Omega}_{ii}^g = \begin{bmatrix} \sigma_C^2 + \sigma_U^2 & \sigma_C^2 & \dots & \sigma_C^2 \\ \sigma_C^2 & \sigma_C^2 + \sigma_U^2 & \dots & \sigma_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_C^2 & \sigma_C^2 & \dots & \sigma_C^2 + \sigma_U^2 \end{bmatrix}. \quad (\text{B.61})$$

Denna korrelationsstruktur bygger på att $c_g = 0$ för alla g (dvs. inget beroende mellan individer inom gruppen betingat på kovariaterna). I stället kan vi anta att c_g existerar, men att c_{gi} och c_{gj} för alla $i \neq j$ och alla g är okorrelerade och att u_{git} är en helt idiosynkratisk felterm (dvs. inget beroende finns i dessa vare sig inom eller mellan individer och kluster). Slutligen finns inget beroende mellan c_g , c_{gi} och u_{git} . Vi får då följande struktur på $\mathbf{\Omega}_{gg}$:

$$\mathbf{\Omega}_{gg} = \begin{bmatrix} \mathbf{\Omega}_{11}^g & \mathbf{\Omega}_{12}^g & \dots & \mathbf{\Omega}_{1G}^g \\ \mathbf{\Omega}_{21}^g & \mathbf{\Omega}_{22}^g & \dots & \mathbf{\Omega}_{G2}^g \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{\Omega}_{n_g 1}^g & \mathbf{\Omega}_{n_g 2}^g & \dots & \mathbf{\Omega}_{n_g n_g}^g \end{bmatrix}, \quad (\text{B.62})$$

där

$$\mathbf{\Omega}_{ii}^g = \begin{bmatrix} \sigma_G^2 + \sigma_C^2 + \sigma_U^2 & \sigma_G^2 + \sigma_C^2 & \dots & \sigma_G^2 + \sigma_C^2 \\ \sigma_G^2 + \sigma_C^2 & \sigma_G^2 + \sigma_C^2 + \sigma_U^2 & \dots & \sigma_G^2 + \sigma_C^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_G^2 + \sigma_C^2 & \sigma_G^2 + \sigma_C^2 & \dots & \sigma_G^2 + \sigma_C^2 + \sigma_U^2 \end{bmatrix}, \quad (\text{B.63})$$

och

$$\mathbf{\Omega}_{ij}^g = \begin{bmatrix} \sigma_G^2 & \sigma_G^2 & \dots & \sigma_G^2 \\ \sigma_G^2 & \sigma_G^2 & \dots & \sigma_G^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_G^2 & \sigma_G^2 & \dots & \sigma_G^2 \end{bmatrix}, \text{ för } i \neq j. \quad (\text{B.64})$$

Som tidigare skattas först modellen med OLS och residualerna \hat{v}_{git} erhålls. Därefter kan σ_G^2 , $\sigma_G^2 + \sigma_C^2$ samt $\sigma_G^2 + \sigma_C^2 + \sigma_U^2$ skattas. Slutligen kan σ_C^2 och σ_U^2 backas ut från dessa skattningar. Givet att $\mathbf{\Omega}$ är korrekt specificerad kan kovariansmatrisen skattas som

$$\hat{\mathbf{V}}(\hat{\boldsymbol{\theta}}_{FGLS}) = \left(\mathbf{W}' \hat{\mathbf{\Omega}}_g^{-1} \mathbf{W} \right)^{-1}. \quad (\text{B.65})$$

Den klusterrobusta kovariansmatrisen där vi tillåter för godtyckligt beroende i feltermerna inom kluster kan användas om vi inte är säkra på att $\mathbf{\Omega}$ är korrekt specificerad.

$$\hat{\mathbf{V}}^c(\hat{\boldsymbol{\theta}}_{FGLS}) = \left(\sum_{g=1}^G \mathbf{W}_g' \hat{\mathbf{\Omega}}^{-1} \mathbf{W}_g \right)^{-1} \left(\sum_{g=1}^G \mathbf{W}_g' \hat{\mathbf{\Omega}}^{-1} \hat{\mathbf{v}}_g \hat{\mathbf{v}}_g' \hat{\mathbf{\Omega}}^{-1} \mathbf{W}_g \right) \left(\sum_{g=1}^G \mathbf{W}_g' \hat{\mathbf{\Omega}}^{-1} \mathbf{W}_g \right)^{-1}. \quad (\text{B.66})$$

Det är också möjligt att skatta variansen som ett "mellanting" genom att använda den klusterrobusta kovariansmatrisen som tillåter för godtyckligt beroende inom individ över tid, men inte mellan individer inom kluster.

B.2.3.1 Inferens rörande γ med aggregerade data

Om målet enbart är att skatta γ visade vi i avsnitt A.7.3.1 att vi kan göra detta genom att dra bort medelvärdet för varje individ för samtliga variabler för att bli av med c_g och c_{gi} . Ett alternativt sätt är att ta medelvärdet av data för varje kluster och tidpunkt. I så fall erhålls

$$\bar{y}_{gt} = \alpha + \mathbf{r}_g' \boldsymbol{\delta} + \bar{\mathbf{x}}_g' \boldsymbol{\beta} + \bar{\mathbf{z}}_g' \boldsymbol{\gamma} + c_g^* + f_t + \bar{u}_{gt}, \quad (\text{B.67})$$

där

$$c_g^* = c_g + \frac{1}{n_g} \sum_{i=1}^{n_g} c_{gi}. \quad (\text{B.68})$$

Vi kan nu även aggregera upp data till klusternivå (så att vi tar medelvärdet för alla $T \cdot n_g$ observationer för kluster g för respektive variabel). Vi erhåller då

$$\bar{y}_g = \alpha + \mathbf{r}_g' \boldsymbol{\delta} + \bar{\mathbf{x}}_g' \boldsymbol{\beta} + \bar{\mathbf{z}}_g' \boldsymbol{\gamma} + c_g^* + \bar{f} + \bar{u}_g. \quad (\text{B.69})$$

Om vi drar bort (B.69) från (B.67) erhålls

$$\bar{y}_{gt} - \bar{y}_g = (\bar{\mathbf{z}}_{gt} - \bar{\mathbf{z}}_g)' \boldsymbol{\gamma} + f_t - \bar{f} + \bar{u}_{gt} - \bar{u}_g. \quad (\text{B.70})$$

Precis som vi gjort tidigare, för enkelhets skull, så utgår vi från att $f_t = 0$, alternativt att vi har med en full uppsättning dummyvariabler för tid i modellen. Låt $\tilde{y}_{gt} = \bar{y}_{gt} - \bar{y}_g$, $\tilde{z}_{gt} = \bar{z}_{gt} - \bar{z}_g$ och $\tilde{u}_{gt} = \bar{u}_{gt} - \bar{u}_g$. Vi får då

$$\tilde{y}_{gt} = \tilde{z}_{gt}'\gamma + \tilde{u}_{gt}. \quad (\text{B.71})$$

Vi kan skatta γ med OLS med en regression med \tilde{y}_{gt} som beroende variabel och \tilde{z}_{gt} som oberoende variabel. Om u_{git} är homoskedastisk med varians σ_U^2 så kommer $V(\tilde{U}_{gt})$ vara heteroskedastisk med varians som är proportionerlig mot inversen av n_{gt} . Vi kan då använda WLS för att transformera variablerna genom att multiplicera \tilde{y}_{gt} och \tilde{z}_{gt} med $1/\sqrt{n_{gt}}$ och sedan skatta modellen med dessa transformerade variabler med OLS. För mer diskussion om detta sätt att skatta modellen, se avsnitt B.1.3. Precis som tidigare finns det fortsatt anledning av använda någon form av robust kovariansmatris då det inte finns några garantier för att u_{git} verkligen är oberoende och homoskedastisk.

En fördel med ansatsen här jämfört med den i avsnitt A.7.3.1 är att vi kan använda denna ansats även i fall när vi inte har tillgång till data på individnivå utan enbart på grupp-nivå.

B.3 Viktning vid olika sannolikhetsurval

Vi utgår här från att vi har en finit, men stor, population av N individer som kan delas in i G olika kluster där antalet individer i kluster g är N_g . Vi är intresserade av $Y = \mathbf{x}'\beta + U$. För att få bra skattning av β måste vi ta hänsyn till hur stickprovet dragits. Vi låter p_{gi} vara sannolikheten att individ i i kluster g väljs ut och vi kommer utgå från att $p_{gi} = p_g$ så att alla individer inom ett kluster har samma sannolikhet att bli valda. Låt $\pi_g = N_g/N$ vara sannolikheten att en slumpmässigt vald individ tillhör grupp g . Vid ett obundet slumpmässigt urval av n individer från populationen vet vi att vi kan göra korrekt inferens till β . I det fallet gäller att $p_g = \pi_g$. Frågan är nu vad som gäller om $p_g \neq \pi_g$?

I det fallet är OLS-estimatorn generellt inte en konsistent estimator av β . Detta gäller även om vi har linjäritet i population, $E(U|\mathbf{x}) = 0$. Med OSU innebär linjäritet att $E(\mathbf{u}|\mathbf{X}) = 0$, men detta gäller inte nödvändigtvis för andra samplingsmetoder. Detta är fallet om de kluster där $p_g > \pi_g$ skiljer sig systematiskt från de kluster där $p_g < \pi_g$. Men om vi känner π_g kan vi genom att använda den viktade minstakvadratmetoden (WLS) få korrekt inferens.

Vi utgår först från fallet när n_g (det antal observationer som väljs ut inom respektive grupp) är bestämt på förhand och där $n_g > 0$ för alla $g = 1, \dots, G$. Eftersom $\sum_{g=1}^G n_g = n$ är stickprovsandelen $h_g = n_g/n$ i kluster g känd. Om vi skattar regressionsmodellen med OLS kommer observationerna i grupp g få vikten h_g . För att få korrekt inferens måste vi vikta om observationerna så att observationerna i grupp g får vikten π_g . Låt v_g vara denna vikt; det gäller då att $h_g v_g = \pi_g \Rightarrow v_g = \pi_g/h_g$. Det gäller då att viktningssmatrisen \mathbf{V} från ekvation (B.6) är en diagonal matris med värden π_g/h_g på diagonalen.

En alternativ urvalsmetod är så kallade *klusterurval*. Med en sådan urvalsmetod väljs slumpmässigt ett på förhand bestämt antal kluster från populationen och sen samplas individer från de utvalda klustren. Vi utgår här från att samtliga individer inom ett utvalt kluster samplas, vilket var fallet i exemplet med skolorna som diskuterades i slutet av avsnitt B.1.3. Låt antalet kluster som samplas vara n_c . I det fallet får vi att sannolikheten att en individ i i kluster g

väljs ut är n_c/G , vilket inte beror på g . Hade vi i stället valt ut stickprovet med OSU hade sannolikheten att en slumpmässigt utvald individ tillhört kluster g varit $\pi_g = N_g/N$. För att få korrekt inferens måste vi därför vikta observationerna med $v_g n_c/G = \pi_g \Rightarrow v_g = \pi_g G/n_c$. Från dessa vikter kan vi sedan skapa den diagonala viktningsmatrisen \mathbf{V} med värden $\pi_g G/n_c$ på diagonalen.

För att skatta variansen kan antingen den vanliga, den heteroskedasticitetsrobusta eller den klusterrobusta kovariansmatrisen användas beroende på vad man tror om feltermsstrukturen. Det finns också specifika variansestimatorer för den specifika samplingmetoden som ibland kan ge mer effektiva variansskattningar.

Bilaga C:

Instrumentvariabelmetoder (IV)

C.1 IV med en oberoende variabel

Vi börjar med den enkla regressionsmodellen i ekvation (12.31) i boken

$$Y = \beta_0 + \beta_1 D + \beta_2 X + U, \quad (\text{C.1})$$

där D är variabeln av intresse och målet är att skatta parametern β_1 . Som i boken kan D vara en binär variabel, men det går lika bra att använda modellen om den är en kontinuerlig variabel. Problemet är att X är en oobserverad variabel, vilket gör att en enkel linjär regression med Y på D inte kommer skatta β_1 . Från formlerna kring regressionsanatomi vet vi i stället att den parameter som skattas är $\beta_1^k = \beta_1 + \beta_2 \text{cov}(D, X)/\sigma_D^2$, där β_1^k är riktningskoefficienten från regressionen av Y på D och $\text{cov}(D, X)/\sigma_D^2$ är riktningskoefficienten från regressionen av X på D . Om X går att observera är det bara att inkludera denna variabel i modellen och skatta med OLS. Om den inte är observerad – men vi i stället observerar en *instrumentvariabel* – kan β_1 i stället skattas med en instrumentalvariabelestimator.

För att underlätta notation framöver så definierar vi $E = \beta_2 X + U$. En instrumentvariabel, Z ska ha två egenskaper:

$$\text{cov}(Z, E) = 0, \quad (\text{C.2})$$

$$\text{cov}(Z, D) \neq 0. \quad (\text{C.3})$$

Den första egenskapen, som brukar kallas *exogenitet*, går inte att undersöka direkt med data (eftersom E inte observeras), utan det måste kunna motiveras utifrån kunskaper om de processer som genererade data i verkligheten. Däremot kan vi undersöka den andra förutsättningen, som brukar kallas *relevans*.

Kovariansen mellan Z och Y är

$$\text{cov}(Z, Y) = \beta_1 \text{cov}(Z, D) + \text{cov}(Z, E). \quad (\text{C.4})$$

Tillsammans med ekvation (C.2) erhåller vi då

$$\beta_1 = \frac{\text{cov}(Z, Y)}{\text{cov}(Z, D)}. \quad (\text{C.5})$$

Eftersom Z , Y och D är observerbara variabler innebär detta att vi kan skatta β_1 trots att X är oobserverad. Med ett slumpmässigt urval av data, $\{y_i, z_i, d_i\}_{i=1}^n$, är det nu bara att ersätta populationskovarianserna med stickprovskovarianserna. Detta ger oss den så kallade instrumentvariabel-estimatoren för β_1 :

$$\hat{\beta}_{1,IV} = \frac{\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\frac{1}{n-1} \sum_{i=1}^n (z_i - \bar{z})(d_i - \bar{d})} = \frac{\sum_{i=1}^n (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^n (z_i - \bar{z})(d_i - \bar{d})}. \quad (\text{C.6})$$

Genom att substituera y_i mot $\beta_0 + \beta_1 d_i + e_i$ erhålls

$$\begin{aligned} \hat{\beta}_{1,IV} &= \frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_0 + \beta_1 d_i + e_i - (\beta_0 + \beta_1 \bar{d} + \bar{e}))}{\sum_{i=1}^n (z_i - \bar{z})(d_i - \bar{d})} \\ &= \frac{\sum_{i=1}^n (z_i - \bar{z})(\beta_1(d_i - \bar{d}) + e_i - \bar{e})}{\sum_{i=1}^n (z_i - \bar{z})(d_i - \bar{d})} \\ &= \beta_1 + \frac{\sum_{i=1}^n (z_i - \bar{z})(e_i - \bar{e})}{\sum_{i=1}^n (z_i - \bar{z})(d_i - \bar{d})} \end{aligned} \quad (\text{C.7})$$

Eftersom d är i nämnaren går det inte att visa att IV-estimatoren är väntevärdesriktig. Däremot konvergerar den i sannolikhet till

$$\text{plim}(\hat{\beta}_{1,IV}) = \beta_1 + \frac{\text{cov}(Z, E)}{\text{cov}(Z, D)}. \quad (\text{C.8})$$

Estimatoren är alltså konsistent om exogenitet gäller, $\text{cov}(Z, E) = 0$. Ibland har man ett instrument som "nästan" är exogent så att $\text{cov}(Z, E) \approx 0$. Om relevansen hos ett sådant instrument är låg så kan instrumentet fungera dåligt även vid stora stickprov. För att se varför, kan det vara bra att kontrastera med OLS-estimatoren. Från formler för regressionsanatomi har vi att

$$\text{plim}(\hat{\beta}_{1,OLS}) = \beta_1 + \frac{\text{cov}(D, E)}{\sigma_D^2}. \quad (\text{C.9})$$

Då $\text{corr}(D, E) = \rho_{DE} = \text{cov}(D, E)/(\sigma_D \sigma_E)$ kan detta skrivas som

$$\text{plim}(\hat{\beta}_{1,OLS}) = \beta_1 + \frac{\sigma_E}{\sigma_D} \cdot \rho_{DE}. \quad (\text{C.10})$$

Låt $\text{corr}(Z, E) = \rho_{ZE} = \text{cov}(Z, E)/(\sigma_Z \sigma_E)$ och $\text{corr}(Z, D) = \rho_{ZD} = \text{cov}(Z, D)/(\sigma_Z \sigma_D)$. Nu kan ekvation (C.8), på samma sätt, skrivas som

$$\text{plim}(\hat{\beta}_{1,IV}) = \beta_1 + \frac{\rho_{ZE} \sigma_Z \sigma_E}{\rho_{ZD} \sigma_Z \sigma_D} = \beta_1 + \frac{\sigma_E}{\sigma_D} \cdot \frac{\rho_{ZE}}{\rho_{ZD}}. \quad (\text{C.11})$$

Den asymptotiska biasen för IV-estimatoren är alltså större än den för OLS i absoluta tal om

$$\left| \frac{\rho_{ZE}}{\rho_{ZD}} \right| > |\rho_{DE}| \quad (\text{C.12})$$

Detta innebär att om instrumentet har låg relevans (vi säger att instrumentet är *svagt*) så att ρ_{ZD} är liten så kan biasen bli betydligt större med IV än med OLS då endogeniteten som kommer av att $\rho_{ZE} \neq 0$ inflateras med en faktor $1/\rho_{ZD}$.

Under antagande om homoskedasticitet (förtydligas nedan) så har vi

$$V\left(\sqrt{n}(\hat{\beta}_{1,IV} - \beta_1)\right) = \frac{\sigma_E^2}{\sigma_D^2 \rho_{ZD}^2}. \quad (\text{C.13})$$

Således är den asymptotiska standardfelet av $\sqrt{n}(\hat{\beta}_{1,IV} - \beta_1)$:

$$\frac{\sigma_E}{\sigma_D} \cdot \frac{1}{|\rho_{ZD}|}. \quad (\text{C.14})$$

När $|\rho_{ZD}|$ är litet (svagt instrument) blir därmed den asymptotiska variansen stor. Dessutom har estimatoren andra dåliga egenskaper när instrumentet är svagt, se avsnitt C.2.2 nedan.

C.2 Det allmänna fallet

Låt oss igen specificera en populationsregressionsmodell:

$$Y = \mathbf{x}'\boldsymbol{\beta} + \mathbf{f}'\boldsymbol{\delta} + U = \mathbf{x}'\boldsymbol{\beta} + E, \quad (\text{C.15})$$

där \mathbf{x} är en vektor av observerade kovariater (inklusive konstant), medan \mathbf{f} är en vektor av oobserverade variabler. Vektorn $\mathbf{x} = [\mathbf{c}' \quad \mathbf{d}']'$ delar vi i sin tur i variabler som vi definierar som exogena, $\mathbf{c} = [1 \quad C_1 \quad \dots \quad C_K]'$, och endogena, $\mathbf{d} = [D_1 \quad \dots \quad D_H]'$, variabler. Precis som tidigare låter vi $E = \mathbf{f}'\boldsymbol{\delta} + U$.

De exogena variablerna, \mathbf{c} , är variabler som kan antas vara okorrelerade med \mathbf{f} , så att $\text{cov}(C_j, E) = 0$ för $j = 0, \dots, K$, medan för de endogena, \mathbf{d} , så finns det skäl att tro att $\text{cov}(D_j, E) \neq 0$ för $j = 1, \dots, H$. För enkelhets skull antar vi att $E(E) = 0$. Detta behöver inte vara fallet, men en avvikelse från detta kommer bara påverka skattningen av interceptet, β_0 . Detta är sällan av intresse, så därför är detta antagande inte begränsande. Vi har då att $E(C_j \cdot E) = 0$ för $j = 0, \dots, K$, medan $E(D_j \cdot E) \neq 0$ för $j = 1, \dots, H$.

Vi har också en vektor \mathbf{z} av L kovariater som antas okorrelerade med E och vi sammanför samtliga exogena kovariater i en vektor $\mathbf{w} = [\mathbf{c}' \quad \mathbf{z}']' = [1 \quad C_1 \quad \dots \quad C_K \quad Z_1 \quad \dots \quad Z_L]'$. Variablerna i \mathbf{z} kallas för *instrument*. Ibland kallas samtliga variabler \mathbf{w} för instrument där \mathbf{c} kallas *inkluderade instrument* medan \mathbf{z} kallas *exkluderade instrument*. När vi enbart skriver *instrument* syftar vi på \mathbf{z} . Vi har

$$E(\mathbf{w}E) = \mathbf{0}. \quad (\text{C.16})$$

Hur hjälper dessa antaganden oss att identifiera $\boldsymbol{\beta}$? Antag att $L = H$, vilket innebär att vi kan multiplicera ekvation (C.15) med \mathbf{w} . Tar vi sedan väntevärdet får vi

$$\begin{aligned} E(\mathbf{w}Y) &= E(\mathbf{w}\mathbf{x}')\boldsymbol{\beta} + E(\mathbf{w}E) \\ &= E(\mathbf{w}\mathbf{x}')\boldsymbol{\beta} \end{aligned} \quad (\text{C.17})$$

Givet att $E(\mathbf{w}\mathbf{x}')$ är inverterbar kan vi identifiera, eller lösa för, β

$$\beta = (E(\mathbf{w}\mathbf{x}'))^{-1} E(\mathbf{w}Y) \quad (\text{C.18})$$

Detta är en generalisering av momentvillkoren för OLS. Givet att vi nu har ett slumpmässigt stickprov kan vi skatta β som

$$\begin{aligned} \hat{\beta}_{IV} &= (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{y} \\ &= (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'(\mathbf{X}\beta + \mathbf{e}) \\ &= \beta + (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{e}. \end{aligned} \quad (\text{C.19})$$

Det gäller nu att

$$\begin{aligned} \text{plim}(\hat{\beta}_{IV}) &= \text{plim}(\beta + (\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{e}) \\ &= \beta + (\text{plim}(\mathbf{W}'\mathbf{X}))^{-1} \text{plim}(\mathbf{W}'\mathbf{e}) \\ &= \beta + (\text{plim}(\mathbf{W}'\mathbf{X}/n))^{-1} \text{plim}(\mathbf{W}'\mathbf{e}/n), \\ &= \beta + (E(\mathbf{w}\mathbf{x}'))^{-1} E(\mathbf{w}E), \\ &= \beta. \end{aligned} \quad (\text{C.20})$$

$\hat{\beta}_{IV}$ är alltså en konsistent estimator av β .

Eftersom E inte observeras är det omöjligt att testa om $E(\mathbf{w}E) = \mathbf{0}$ såvida vi inte har tillgång till annan information. Däremot kan villkoret att $E(\mathbf{w}\mathbf{x}')$ är inverterbar testas.

För att se detta, antag för tillfället att vi endast har en endogen variabel, D_1 , och ett instrument, Z_1 (dvs. $H = L = 1$). Vi har sedan en regressionsmodell, som brukar kallas *förstastegsregressionen* (en förklaring till detta namn kommer i nästa avsnitt), som är

$$D_1 = \gamma_0 + \gamma_1 C_1 + \dots + \gamma_K C_K + \gamma_{K+1} Z_1 + V, \quad (\text{C.21})$$

där V är en felterm. Det gäller att $E(\mathbf{w}\mathbf{x}')$ är inverterbar om och endast om

$$\gamma_{K+1} \neq 0. \quad (\text{C.22})$$

Genom att skatta förstastegsregressionen är det således möjligt att testa detta villkor. Vid ett sådant test bör ofta den robusta EHW-estimatorn användas. Då IV-estimatorn med svaga instrument har dåliga egenskaper räcker det inte med att testet är signifikant på fem procents signifikansnivå, något vi återkommer till.

Notera att vi inte bryr oss om tolkning av paramterarna i ekvation (C.21). Skälet att ha med C_1, \dots, C_K är att det är endast den unika (exogena) variationen i Z_1 som kan användas vid skattning av en effekt av D_1 på Y . Av den anledningen måste vi betinga bort variation som kommer från C_1, \dots, C_K . Vi säger att D_1 är *partiellt* korrelerad med Z_1 .

C.2.1 2SLS

I fallet ovan hade vi lika många instrument som det fanns endogena variabler (dvs. $L = H$). När så inte är fallet kan inte IV-estimatorn i ekvation (C.19) användas. Detta eftersom \mathbf{W} har

dimension $n \times (K + L + 1)$, medan \mathbf{X} har dimension $n \times (K + H + 1)$. Däremed går det inte att ta fram $\mathbf{W}'\mathbf{X}$. När så är fallet är modellen *överidentifierad*. Vi kan såklart bara slänga bort instrument så att $L = H$, men det är inte effektivt. Ett bättre sätt är att vikta ihop instrumenten.

Detta kan vi åstadkomma genom följande linjära regressioner, för $j = 0, \dots, K + H$:

$$X_j = \gamma_0^j + \gamma_1^j C_1 + \dots + \gamma_K^j C_K + \gamma_{K+1}^j Z_1 + \dots + \gamma_{K+L}^j Z_{K+L} + V_j. \quad (\text{C.23})$$

Från dessa regressioner är linjärprojektionerna för respektive variabel:

$$X_j^* = \gamma_0^j + \gamma_1^j C_1 + \dots + \gamma_K^j C_K + \gamma_{K+1}^j Z_1 + \dots + \gamma_{K+L}^j Z_{K+L}. \quad (\text{C.24})$$

För $j = 0, \dots, K$ finns X_j även på högersidan vilket innebär att $X_j^* = X_j$. Detta är dock inte fallet för $j = K + 1, \dots, K + H$. Låt $\mathbf{\Gamma}_j = (\mathbf{E}(\mathbf{w}\mathbf{w}'))^{-1} \mathbf{E}(\mathbf{w}X_j)$ och $\mathbf{\Gamma} = [\mathbf{\Gamma}_0 \quad \mathbf{\Gamma}_1 \quad \dots \quad \mathbf{\Gamma}_{K+H}]$. Vi kan nu skriva

$$\mathbf{x}^* = \begin{bmatrix} X_0^* \\ X_1^* \\ \vdots \\ X_{K+H}^* \end{bmatrix} = \mathbf{\Gamma}'\mathbf{w} = \begin{bmatrix} \gamma_0^0 & \gamma_1^0 & \dots & \gamma_{K+L}^0 \\ \gamma_0^1 & \gamma_1^1 & \dots & \gamma_{K+L}^1 \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_0^{K+H} & \gamma_1^{K+H} & \dots & \gamma_{K+L}^{K+H} \end{bmatrix} \begin{bmatrix} W_0 \\ W_1 \\ \vdots \\ W_{K+L} \end{bmatrix}. \quad (\text{C.25})$$

Om vi nu multiplicerar ekvation (C.15) med \mathbf{x}^* och tar väntevärdet får vi

$$\mathbf{E}(\mathbf{x}^*Y) = \mathbf{E}(\mathbf{x}^*\mathbf{x}')\boldsymbol{\beta} + \mathbf{E}(\mathbf{x}^*E). \quad (\text{C.26})$$

Vi har att X_j^* är en linjärkombination av exogena variabler i vektorn \mathbf{w} . Samtliga dessa är okorrelerade med E , vilket innebär att $\mathbf{E}(X_j^* \cdot E) = 0$. Detta gäller för alla $j = 0, 1, \dots, K + H$. Därmed har vi att $\mathbf{E}(\mathbf{x}^*E) = \mathbf{0}$. Om $\mathbf{E}(\mathbf{x}^*\mathbf{x}')$ är inverterbar har vi då

$$\boldsymbol{\beta} = (\mathbf{E}(\mathbf{x}^*\mathbf{x}'))^{-1} \mathbf{E}(\mathbf{x}^*Y). \quad (\text{C.27})$$

Hur gör vi nu för att skatta $\boldsymbol{\beta}$? Vi utgår från att data är slumpmässigt dragna från populationen. Detta innebär att vi antar att $\{y_i, \mathbf{c}_i, \mathbf{z}_i\}_{i=1}^n$ är ett OSU. $\boldsymbol{\beta}$ kan nu skattas i två steg med det som kallas för tvåstegsminstakvadrat-estimatorn (2SLS, eng. *two stage least squares*).

1. Först skattas $\mathbf{\Gamma}$ som $\hat{\mathbf{\Gamma}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}$.¹⁸ Från denna får vi predicerade värden

$$\hat{\mathbf{X}}^* = \mathbf{W}\hat{\mathbf{\Gamma}}. \quad (\text{C.28})$$

Proceduren kan beskrivas med ord som att vi skattar $K + H + 1$ linjära regressioner med OLS för varje element i \mathbf{x} som beroende och \mathbf{z} som oberoende och sedan tar vi fram

¹⁸Detta är den multivariata extensionen till ekvation (C.21) med mer än en endogen variabeln och mer än ett instrument. För att inferens ska vara korrekt så får dessa *förstastegs estimationer* inte vara svaga. Vi diskuterade detta ovan i relation till ekvation (C.22) och återkommer till en mer allmän diskussion om detta i slutet av denna bilaga.

predicerade värden för respektive regressionen. De predicerade värdena för konstanten och alla exogena C -variabler kommer vara identiska med ursprungsvariablerna. Därför brukar man bara säga att man skattar H stycken *förstastegsregressioner* (dvs. där de endogena variablerna är beroende variabler).

2. Använd $\widehat{\mathbf{X}}^*$ som instrument för \mathbf{X} enligt ekvation (C.28):

$$\widehat{\beta}_{2SLS} = (\widehat{\mathbf{X}}^{*'} \mathbf{X})^{-1} \widehat{\mathbf{X}}^{*'} \mathbf{y}. \quad (\text{C.29})$$

Vi kan visa att under homoskedasticitet (se nedan för förtydligande) är $\widehat{\mathbf{X}}^*$ de bästa instrumenten för \mathbf{X} och 2SLS-estimatoren är då den mest effektiva av alla IV-estimatorer.

Frågan nu är hur man ska göra inferens? För att svara på frågan kan vi först göra en omväg genom att beskriva estimatoren på det sätt som man historisk använde den, vilket också gav estimatoren dess namn.

Låt $\widehat{\mathbf{V}}$ vara en matris av dimension $n \times (K + H + 1)$ av residualer från de $K + H + 1$ OLS-skattningarna av ekvation (C.23). Vi har då att $\mathbf{X} = \widehat{\mathbf{X}}^* + \widehat{\mathbf{V}}$, vilket också innebär att för en viss variabel \mathbf{x}_j gäller att $\mathbf{x}_j = \widehat{\mathbf{x}}_j^* + \widehat{\mathbf{v}}_j$. Från egenskapen hos OLS, (se ekvation A.17 från avsnitt A.1) vet vi att att $\mathbf{W}'\widehat{\mathbf{v}}_j = \mathbf{0}$, från vilket det följer att $\widehat{\mathbf{x}}_j^{*'}\widehat{\mathbf{v}}_j = 0$. Detta innebär att $\widehat{\mathbf{X}}^{*'}\widehat{\mathbf{V}} = \mathbf{0}$. Slutligen innebär detta att $\widehat{\mathbf{X}}^{*'}\mathbf{X} = \widehat{\mathbf{X}}^{*'}(\widehat{\mathbf{X}}^* + \widehat{\mathbf{V}}) = \widehat{\mathbf{X}}^{*'}\widehat{\mathbf{X}}^*$. 2SLS-estimatoren i ekvation (C.29) kan nu skrivas som

$$\widehat{\beta}_{2SLS} = (\widehat{\mathbf{X}}^{*'} \widehat{\mathbf{X}}^*)^{-1} \widehat{\mathbf{X}}^{*'} \mathbf{y}. \quad (\text{C.30})$$

Men detta är en vanlig OLS-regression med de predicerade variablerna för förstastegsregressionerna, $\widehat{\mathbf{X}}^*$, som oberoende variabler. Man skulle då kunna tänka sig att använda residualerna från denna regression som grund för inferens. Detta är dock *inte* korrekt. De korrekta residualerna är

$$\widehat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\widehat{\beta}_{2SLS}. \quad (\text{C.31})$$

Dvs. \mathbf{X} ska användas och inte $\widehat{\mathbf{X}}^{*'}$ för att skapa residualerna.

För ett slumpmässigt stickprov och instrument som uppfyller villkoren för relevans och exogenitet gäller att

$$\sqrt{n}(\widehat{\beta}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mathbf{E}(\mathbf{x}^* \mathbf{x}^{*'}))^{-1} \mathbf{E}(E^2 \mathbf{x}^* \mathbf{x}^{*'}) (\mathbf{E}(\mathbf{x}^* \mathbf{x}^{*'}))^{-1}\right). \quad (\text{C.32})$$

Vi kan nu göra robust inferens på sedvanligt sätt och skatta den asymptotiska kovariansmatrisen för $\widehat{\beta}_{2SLS}$ som:

$$\widehat{\mathbf{V}}^r(\widehat{\beta}_{2SLS}) = (\widehat{\mathbf{X}}^{*'} \widehat{\mathbf{X}}^*)^{-1} \left(\sum_{i=1}^n \widehat{\mathbf{e}}_i^2 \widehat{\mathbf{x}}_i^* \widehat{\mathbf{x}}_i^{*'} \right) (\widehat{\mathbf{X}}^{*'} \widehat{\mathbf{X}}^*)^{-1}. \quad (\text{C.33})$$

Med homoskedasticitet som nämndes ovan har vi att

$$\mathbf{E}(E^2 \mathbf{x}^* \mathbf{x}^{*'}) = \mathbf{E}(E^2) \mathbf{E}(\mathbf{x}^* \mathbf{x}^{*'}) = \sigma_E^2 \mathbf{E}(\mathbf{x}^* \mathbf{x}^{*'}). \quad (\text{C.34})$$

Detta gör att ekvation (C.32) förenklas till

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_E^2 \mathbb{E}(\mathbf{x}^* \mathbf{x}^{*'})^{-1}). \quad (\text{C.35})$$

Givet homoskedasticitet kan variansen för $\hat{\beta}_{2SLS}$ skattas enligt

$$\hat{V}(\hat{\beta}_{2SLS}) = s_E^2 (\hat{\mathbf{X}}^{*'} \hat{\mathbf{X}}^*)^{-1}, \quad (\text{C.36})$$

där

$$s_E^2 = \frac{1}{n - (K + H + 1)} \sum_{i=1}^n \hat{e}_i^2. \quad (\text{C.37})$$

Givet homoskedasticitet kan man visa att 2SLS-estimatorn har den minsta variansen av alla IV-estimatorer som använder linjära funktioner av instrumenten. Notera att detta resultat endast är meningsfullt för situationen med $L > H$.

C.2.2 Svaga instrument

Om den partiella korrelationen (dvs. det första steget) inte är tillräckligt starkt har vi en situation med *svaga instrument*. Detta är mycket problematiskt. Problemet är att IV-estimatorerna (inkluderande 2SLS) inte är väntevärdesriktiga och att inferensen blir felaktig (estimatorn är inte längre normalfördelad). Detta problem löses inte av att stickprovet är stort, utan dessa problem kvarstår även då om instrumenten är svaga. Se exempelvis Bound m. fl. (1995), Nelson och Startz (1990a, 1990b) och Stock m. fl. (2002) för mer information om detta, samt Imbens och Wooldridge (2009) för en fördjupande diskussion.

En lösning är att använda en annan strategi som föreslagits av Anderson och Rubin (1949), se I. Andrews m. fl. (2019) för en genomgång. Andra sätt är att använda betingade "likelihood ratio-test" (se D. W. K. Andrews m. fl., 2006; Blundell m. fl., 2005; Moreira, 2003). Dessa metoder är svåra att tillämpa och ger som regel resultat som inte är informativa om instrumentet är svagt.

För att testa om instrumenten är svaga brukar F -statistikan för nollhypotesen att instrumenten i förstastegsregressionerna tillsammans inte kan förklara något av variationen i den endogena variabeln användas. Staiger och Stock (1997) föreslog en F -statistika över 10 för att instrumenten inte ska anses vara svaga utan relevanta. Även när denna tumregel är uppfylld har det visat sig att instrumenten är för svaga, se exempelvis Stock m. fl. (2002) och I. Andrews m. fl. (2019). För en uppdaterad genomgång av olika test för svaga instrument i överidentifierade system se Windmeijer (2023).

Referenser

- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2020). Sampling-Based versus Design-Based Uncertainty in Regression Analysis. *Econometrica*, 88(1), 265–296.
- Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. M. (2022). When Should You Adjust Standard Errors for Clustering? *The Quarterly Journal of Economics*, 138(1), 1–35.
- Anderson, T., & Rubin, H. (1949). estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20, 46–63.
- Andrews, D. W. K., Moreira, M. J., & Stock, J. H. (2006). Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression. *Econometrica*, 74(3), 715–752.
- Andrews, I., Stock, J. H., & Sun, L. (2019). Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1), 727–753.
- Angrist, J. D., & Pischke, J. S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.
- Blundell, R., Newey, W., & Persson, T. (Ed.). (2005). *Inference with Weak Instruments* [Prepared for the 2005 World Congress of the Econometric Society]. Cambridge University Press.
- Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak. *Journal of the American Statistical Association*, 90(430), 443–450.
- Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 59–82.
- Frisch, R., & Waugh, F. V. (1933). Partial time regressions as compared with individual trends. *Econometrica*, 1(4), 387–401.
- Greene, W. H. (2012). *Econometric Analysis* (7. utg.). Pearson Education Limited.
- Hansen, C. B. (2007). Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics*, 141(2), 597–620.
- Harvey, A. C. (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, 44(3), 461–465.
- Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 221–233.
- Imbens, G. W., & Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), 5–86.

- Lovell, M. C. (1963). Seasonal adjustment of economic time series and multiple regression analysis. *Journal of the American Statistical Association*, 58(304), 993–1010.
- MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. I X. Chen & N. R. Swanson (Red.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis: Essays in Honor of Halbert L. White Jr.* (s. 437–461). Springer, New York.
- Moreira, M. J. (2003). A Conditional Likelihood Ratio Test for Structural Models. *Econometrica*, 71(4), 1027–1048.
- Nelson, C. R., & Startz, R. (1990a). The distribution of the instrumental variable estimator and its t ratio when the instrument is a poor one. *Journal of Business*, 63(63), 124–140.
- Nelson, C. R., & Startz, R. (1990b). Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator. *Econometrica*, 58(4), 967–976.
- Staiger, D., & Stock, J. (1997). Instrumental variables regression with weak Instruments. *Econometrica*, 65(3), 557–586.
- Stock, J. H., Wright, J. H., & Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business and Economic Statistics*, 20(4), 518–529.
- Student. (1908). The Probable Error of a Mean. *Biometrika*, 6, 1–25.
- White, H. (1980). Using least squares to approximate unknown regression functions. *International Economic Review*, 21(1), 149–170.
- Windmeijer, F. (2023). The Robust F-Statistic as a Test for Weak Instruments.
- Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.