Den multiple lineære regressionsmodel (MLR)

Økonometri A

Bertel Schjerning

,,

Program

Motivation (W3.1)

MLR på matrix form (W.E2)

Egenskaber ved MLR (W3.2)

- Frisch-Waughs Teorem

Fordelingen af OLS estimaterne (W3.3-W3.4)

- Middelrethed
- Irrelevante og udeladte variable
- Varians og standard fejl for OLS

OLS er BLUE (W3.5, W3.A6)

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Motivation

Motivation

Indtil videre har vi undersøgt sammenhænge mellem et outcome (y) og en enkelt forklarende variabel (x).

Nu udvider vi dette til flere forklarende variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Notation:

- y - den afhængige variabel
- $x_1, \ldots, x_k$ - de k forklarende variable
- β_0 - konstantled
- $\beta_1, \ldots, \beta_k$ - hældningskoefficienter
- u - fejlleddet
- k - antallet af forklarende variable, $k + 1$ parametre i alt

Motivation: Hvorfor MLR?

- Ofte er der flere faktorer som påvirker y
  - MLR kan eksplicit kontrollere for flere faktorer
  - Disse faktorer er ikke længere inkluderet i u
  - Ofte lettere at få opfyldt antagelsen om $E(u|x) = 0$

- Det gør det lettere at lave ceteris paribus fortolkninger
  - Vi kan nu eksplicit holde faktorerne $x_2, .., x_k$ faste, når vi bestemmer effekten af at ændre x_1 med en enhed.

- Mulighed for mere fleksibel modellering af funktionelle former

Motivation: Hvorfor MLR?

Hvis vi holder alle øvrige forklarende variable konstante undtagen x_j:

$$\Delta \hat{y} = \hat{\beta}_j \Delta x_j$$

MLR muliggør "ceteris paribus" fortolkninger, selvom vi i data ikke kan holde alle faktorer konstante.

Vi siger, at vi "kontrollerer for de øvrige effekter".

Motivation: Hvorfor MLR?

Mulighed for mere fleksibel modellering af funktionelle former

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + u$$
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \cdot x_2 + u$$

Med x_1^2 eller $x_1 \cdot x_2$ bliver sammenhængen mellem y og x'erne ikke-lineær:

$$dE(y|X)/dx_1 = \beta_1 + 2\beta_2 x_1$$
$$dE(y|X)/dx_1 = \beta_1 + \beta_3 x_2$$

(hvis $dE(u|X)/dx_1 = 0$)

OBS: Modellerne er stadig **lineær i parameterne**

Let's Code!

- **Jupyter Notebook:** `03_mlr_examples.ipynb`
- **Python Module:** `mymlr.py`
- **Part 1:** Timeløn, uddannelse og erfaring (OLS Estimation)

Let's implement the empirical example!

MLR på matrix form

Den multiple lineære regressionsmodel

Som ved SLR kan vi udlede OLS som en momentestimator (MM).

Vi har momenterne:

$$\begin{aligned} E(u|X) &= 0 \Rightarrow \\ E(u) &= 0 \\ E(ux_j) &= 0 \quad \text{for alle } j \end{aligned}$$

Vi har altså $k+1$ momentbetingelser og samme antal parametre.

MLR på matrix form (Se appendiks E1.)

Alle observationer:

$$y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \cdots + \beta_k x_{k1} + u_1$$
$$\vdots$$
$$y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \cdots + \beta_k x_{kn} + u_n$$

På matrix form:

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} \cdots x_{k1} \\ \vdots & \vdots \\ 1 & x_{1n} \cdots x_{kn} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix}$$

Modellen kan nu skrives som

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

Udledningen af OLS estimatoren

Det teoretiske moment ($k+1$ momenter)

$$E(\mathbf{X}'\mathbf{u}) = \mathbf{0}$$

Vi finder OLS ved at erstattet dette med det empiriske moment

$$\tfrac{1}{n}\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$$

hvor $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Dette er $k+1$ ligninger. Udledning

$$\begin{aligned}
\tfrac{1}{n}\mathbf{X}'\hat{\mathbf{u}} &= \mathbf{0} \Leftrightarrow \\
\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \mathbf{0} \Leftrightarrow \\
\mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}
\end{aligned}$$

Hvis $\mathbf{X}$ har fuld rang, så eksisterer $(\mathbf{X}'\mathbf{X})^{-1}$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

OLS som minimizer af kvadrerede residualer (1/2)

OLS estimatoren findes ved at minimere de kvadrerede residualer:

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \hat{u}_i^2 = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \beta)^2$$

På matrixform:

$$\min_{\boldsymbol{\beta}} \hat{\mathbf{u}}' \hat{\mathbf{u}} = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)$$

Dette er summen af kvadrerede residualer, som vi ønsker at minimere med hensyn til $\boldsymbol{\beta}$.

OLS som minimizer af kvadrerede residualer (2/2)

Førsteordensbetingelser (FOC):

$$\frac{\partial}{\partial \boldsymbol{\beta}} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right] = \mathbf{0}$$

Udregning af gradienten giver:

$$\begin{aligned} -2\mathbf{X}'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) &= \mathbf{0} \\ \mathbf{X}'\mathbf{y} &= \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \end{aligned}$$

Dette fører til samme OLS estimator:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Bemærk, at FOC leder til de samme empiriske momentbetingelser som momentmetoden:

$$\frac{1}{n}\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}, \quad \text{hvor } \hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Udledningen af OLS estimatoren

Hvornår har en matrix fuld rang?

Søjlerne i $\mathbf{X}$ skal være lineært uafhængige.

Eksempler på matricer, der ikke har fuld rang:

- Hvis der er en variabel, som er konstant:
  Fx inkluderer $x =$ år i en cross-section regression.
- Hvis der findes en lineær relation mellem to eller flere variable:
  Fx inkluderer alder og fødselsår i en cross-section regression.
  $$\text{fødselsår} + \text{alder} = \text{året (som er konstant)}.$$
- I begge tilfælde antager vi, at der er et konstantled i modellen.

Dvs. hvis vi observerer, at ældre mennesker har en højere løn i et cross-section datasæt, kan vi ikke sige om det skyldes alder eller en fødselsårseffekt.

QR Decomposition og lineær afhængighed

QR Decomposition: En matrix $\mathbf{X}$ kan opdeles som:

$$\mathbf{X} = \mathbf{QR}$$

- $\mathbf{Q}$: Ortogonal matrix med ortonormale kolonner ($\mathbf{Q}'\mathbf{Q} = \mathbf{I}$).
- $\mathbf{R}$: Øvre triangulær matrix (alle elementer under diagonalen er 0).

Lineær afhængighed:

- Diagonalelementer i $\mathbf{R}$ tæt på 0 betyder, at variablen er lineært afhængig.
- Afhængige variable fjernes for at sikre, at $\mathbf{X}$ har fuld rang.

Anvendelse i regression:

- QR decomposition bruges til at sikre, at designmatricen har fuld rang, hvilket er nødvendigt for at estimere OLS koefficienterne.

Let's Code!

- **Jupyter Notebook:** `03_mlr_examples.ipynb`
- **Part 2:** Timeløn, uddannelse og erfaring
  (OLS estimator og rangbetingelsen)

Egenskaber ved MLR

Egenskaberne ved MLR

Mange af egenskaberne fra SLR gælder også for MLR:

Prædikterede værdier:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \quad \text{(elementform)}$$

$$\hat{y} = \mathbf{X}\hat{\beta} \quad \text{(matrixform)}$$

For MLR med konstantled gælder:

- $\sum_{i=1}^{n} \hat{u}_i = 0$ eller $\mathbf{1}'\hat{\mathbf{u}} = 0$
- $\sum_{i=1}^{n} \hat{u}_i x_{ji} = 0$ eller $\mathbf{X}'\hat{\mathbf{u}} = \mathbf{0}$
- Regressionslinjen går gennem $(\bar{y}, \bar{x}_1, \ldots, \bar{x}_k)$:

$$\bar{y} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j \bar{x}_j \quad \text{eller} \quad \bar{\mathbf{y}} = \bar{\mathbf{X}}\hat{\beta}$$

Egenskaberne ved MLR

Vi kan udregne goodness of fit:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

hvor:

$SST \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2 = (\mathbf{y} - \bar{y})'(\mathbf{y} - \bar{y})$
$SSE \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = (\hat{\mathbf{y}} - \bar{y})'(\hat{\mathbf{y}} - \bar{y})$
$SSR \equiv \sum_{i=1}^{n} \hat{u}_i^2 = \hat{\mathbf{u}}'\hat{\mathbf{u}}$

Om $R^2 \in [0, 1]$:

- Jo højere R^2, jo mere af variationen er forklaret.
- Flere variable vil altid få R^2 til at stige (eller være uændret).
- R^2 alene kan ikke afgøre antallet af variable i modellen.

Justeret goodness-of-fit

Netop fordi R^2 altid vil stige ved flere variable i modellen, beregner Stata også et justeret R^2:

$$\bar{R}^2 = 1 - \frac{SSR/(n-k-1)}{SST/(n-1)} = 1 - \frac{SSR}{SST} \cdot \frac{n-1}{n-k-1}$$

- $\bar{R}^2$ "straffer" for mange forklarende variable.
- $\bar{R}^2$ stiger kun, hvis de nye variable forklarer tilstrækkeligt meget af variationen i data.
- $\bar{R}^2$ kan være negativ.

Mere om det i Wooldridge kapitel 6.3.

Egenskaberne ved MLR: Frisch-Waughs Teorem

Estimatoren af den multiple regressionsmodel kan foretages ved to regressionsestimationer.

Model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

$\hat{\beta}_j$ kan bestemmes ved to simple lineære regressioner:

Procedure:

1. Regresser x_j på alle de andre x'er (inkl. konstantled), og gem residualerne som $\hat{r}_j$.
2. Regresser y på $\hat{r}_j$ for at få $\hat{\beta}_j$.

Resultat:

- Estimatet $\hat{\beta}_j$ fra trin 2 er det samme som koefficienten til x_j i den fulde model.

Egenskaberne ved MLR: Frisch-Waugh Teorem (Matrix form)

Model:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

$\hat{\boldsymbol{\beta}}_1$ kan bestemmes ved hjælp af to regressioner:

Procedure:

1. Regresser $\mathbf{X}_1$ på $\mathbf{X}_2$, og gem residualerne som $\mathbf{r}_1 = \mathbf{X}_1 - \mathbf{X}_2\hat{\gamma}$:

$$\hat{\gamma} = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1$$

2. Regresser $\mathbf{y}$ på $\mathbf{r}_1$ for at få $\hat{\boldsymbol{\beta}}_1$:

$$\hat{\boldsymbol{\beta}}_1 = (\mathbf{r}_1'\mathbf{r}_1)^{-1}\mathbf{r}_1'\mathbf{y}$$

Resultat: Estimatet $\hat{\boldsymbol{\beta}}_1$ fra trin 2 er det samme som koefficienten til $\mathbf{X}_1$ i den fulde model.

Egenskaberne ved MLR: Frisch-Waughs Teorem

Trin 1:

Regresser x_j på alle de andre x'er og gem residualerne som $\hat{r}_j$:

$$x_j = \gamma_0 + \gamma_1 x_1 + \gamma_2 x_2 + \cdots + \gamma_{j-1} x_{j-1} + \gamma_{j+1} x_{j+1} + \cdots + \gamma_n x_n + r_j$$

Residualerne er $\hat{r}_j = x_j - \hat{x}_j$.

Trin 2:

Regresser y på $\hat{r}_j$:

$$y = \delta_0 + \beta_j \hat{r}_j + e$$

OLS estimatoren for $\hat{\beta}_j$ fås som:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} y_i \hat{r}_{ji}}{\sum_{i=1}^{n} \hat{r}_{ji}^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(\hat{r}_{ji} - \bar{\hat{r}}_{ji})}{\sum_{i=1}^{n} (\hat{r}_{ji} - \bar{\hat{r}}_{ji})^2}$$

I skal vise dette i den første obligatoriske opgave.

Frisch-Waugh Teorem: Python Eksempel

Let's implement the Frisch-Waugh theorem in Python!

- Use Jupyter Notebook: `03_mlr_examples.ipynb`
- Part 3: Timeløn, uddannelse og erfaring (Frisch-Waugh)

Frisch-Waugh Teorem: Intuition

I SLR estimerer vi β_j uden at kontrollere for andre variable:

$$\hat{\beta}_j^{SLR} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_{ji} - \bar{x}_j)}{\sum_{i=1}^{n}(x_{ji} - \bar{x}_j)^2}$$

I MLR estimeres β_j ved at fjerne variationen i x_j, som er korreleret med de øvrige variable:

$$\hat{\beta}_j^{MLR} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(\hat{r}_{ji} - \bar{\hat{r}}_j)}{\sum_{i=1}^{n}(\hat{r}_{ji} - \bar{\hat{r}}_j)^2}$$

MLR kontrollerer altså for andre x-variable ved at bruge residualer $\hat{r}_{ji}$ fra en regression af x_j på de andre variable.

Frisch-Waugh Teorem: Intuition

Hvorfor giver det mening?

Den forventede værdi af SLR-estimatet:

$$E(\hat{\beta}_j|X) = \beta_j + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(u|X)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

Estimatet er ikke middelret, hvis x_j er korreleret med u, som indeholder effekten af de øvrige x'er.

Ved at bruge residualerne fra regressionen af x_j på de andre x'er, fjerner vi denne korrelation.

$\rightarrow$ Forsøg på at "rense data for dårlig variation".

Frish-Waughs Teorem: Anvendelse

Frish-Waughs Teorem kan blandt andet bruges til at visualisere data selvom modellen indeholder mange forklarende variable.

- I MLR er korrelationen mellem x_j og y ikke (altid) informative om den partielle effekt af x_j og y.
- Noget af korrelationen skyldes i stedet andre x'er, som vi gerne vil kontrollere for.

Derfor:

- I stedet for at plotte x_j og y, bør vi plotte $\hat{r}_j$ og y.
- På den måde kan vi fx validere den funktionelle form.

Derudover skal vi se at intuitionen fra Frish-Waughs går igen i mange af vores senere udledninger.

Partial Linear Models: Robinson Estimator (IKKE PENSUM)

Samme idé kan bruges for mere *partiel lineære modeller*, som bygger videre på intuitionen fra Frisch-Waugh Teoremet.

$$y = \mathbf{X}\boldsymbol{\beta} + g(Z) + u$$

Robinson Estimatoren bruges til at estimere $\boldsymbol{\beta}$ i en model med en lineær og en ikke-lineær komponent.

Trin:

1. Estimér $E(y|Z)$ og $E(\mathbf{X}|Z)$ ikke-parametrisk (fx kernel regression eller splines).
2. Få residualerne $y^* = y - E(y|Z)$ og $\mathbf{X}^* = \mathbf{X} - E(\mathbf{X}|Z)$.
3. Regresser y^* på $\mathbf{X}^*$ for at estimere $\boldsymbol{\beta}$.

Fordelen ved Robinson Estimatoren er, at vi kan estimere $\boldsymbol{\beta}$ uden samtidigt at skulle estimere en specifik form af $g(Z)$.

Fordelingen af OLS estimaterne

Centrale antagelser for den multiple lineære regressionsmodel

MLR.1 Populationsmodellen er lineær i parametrene:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$

På matrixform: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$

MLR.2 Tilfældig udvælgelse:

Vi har tilfældigt udvalgte og uafhængige observationer af variablerne $(x_i, y_i) : i = 1, ..., n$ fra en population.

MLR.3 Ingen perfekt multikollinearitet mellem de forklarende variable:

I datasættet skal $\mathbf{X}$ have fuld rang.

MLR.4 Den betingede middelværdi af fejlleddet skal være 0:

$E(u|x_1, x_2, x_2, ..., x_k) = 0.$

På matrixform: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$

Hvad betyder MLR.4?

Hvis MLR.4 er opfyldt, gælder:

$$E(ux_j) = 0 \quad \text{for} \quad j = 1, .., k$$

Altså er fejlleddet ukorreleret med alle de forklarende variable.

Eksempler på at **MLR.4** ikke er opfyldt:

- Udeladte variable (der vil være indeholdt i u) som er korreleret med de forklarende variable.
- Forkert funktionel form (se kap. 9).
- Målefejl i de forklarende variable (se kap. 9).

Når MLR.4 er opfyldt, er de forklarende variable **eksogene**. Hvis x_j derimod er korreleret med u, kaldes x_j for **endogen**.

Middelrethed af OLS estimatoren

Teorem 3.1: Middelrethed af OLS estimatoren

Under antagelse af MLR.1–MLR.4 er OLS estimatoren **middelret**:

$$E(\hat{\beta}_j) = \beta_j \quad \text{for } j = 0, 1, 2, ..., k$$

for alle værdier af β_j.

Hvad betyder det, at OLS er middelret?

- OLS estimatoren er **i gennemsnit** korrekt, hvis vi gentager estimationen på uendeligt mange stikprøver.
- Den forventede værdi af estimaterne vil være lig de sande parametre.
- På ethvert givet stikprøvedatasæt kan der være små afvigelser fra den sande værdi.

Middelrethed af OLS estimatoren: Bevis

Ingredienser til bevis:

1. **Regneregler** ($\mathbf{A}$ matrix, $\mathbf{x}$ stok. vektor, $\mathbf{b}$ vektor):
   1.1 $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ og $\mathbf{A}^{-\mathbf{1}}\mathbf{A} = \mathbf{I}$
   1.2 $E(\mathbf{b} + \mathbf{A}\mathbf{x}) = \mathbf{b} + \mathbf{A}\mathbf{E}(\mathbf{x})$
2. **Antagelser**:
   2.1 MLR.1 (Model): $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
   2.2 MLR.2: Tilfældig udvalgt datasæt med n observationer
   2.3 MLR.3: $\mathbf{X}$ har fuld rang
   2.4 MLR.4: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$
3. **OLS Estimator**:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$$

Middelrethed af OLS estimatoren: Bevis (1/2)

1. Start med OLS estimatoren:

Fra OLS estimatoren ved vi, at:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

2. Erstat y med modellen (MLR.1):

Fra MLR.1: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$, så vi kan skrive:

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

Dette udtryk er kun gyldigt, hvis $\mathbf{X}$ har fuld rang (MLR.3).

Middelrethed af OLS estimatoren: Bevis (2/2)

3. Tag forventning betinget på X (MLR.4):

Fra MLR.4 ved vi, at $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$.

$$E(\hat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\mathbf{u}|\mathbf{X}) = \boldsymbol{\beta}$$

4. Brug loven om itereret forventning (MLR.2):

Med tilfældigt udvalgte observationer og uafhængighed kan vi bruge loven om itereret forventning til at fjerne betingningen på **X**:

$$E(\hat{\boldsymbol{\beta}}) = E\left[E(\hat{\boldsymbol{\beta}}|\mathbf{X})\right] = E[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

Konklusion:

OLS estimatoren er middelret, dvs. $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Inklusion af irrelevante variable

Hvad sker der, hvis man inkluderer en irrelevant variabel?

Antag at den "sande model" opfylder MLR.1-MLR.4 og er:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Men vi estimerer en model, som også inkluderer x_3, så:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{\beta}_2 x_2 + \tilde{\beta}_3 x_3 + \tilde{u}$$

Spørgsmål: Hvordan påvirker inkluderingen af den irrelevante variabel x_3 vores estimater?

Inklusion af irrelevante variable

Hvad sker der, hvis man inkluderer en irrelevant variabel?

OLS er stadig middelret for alle værdier af β:

$$E(\hat{\beta}_j) = \beta_j$$

Det betyder, at:

$$E(\hat{\beta}_3) = \beta_3 = 0$$

Er det ligegyldigt at inkludere irrelevante variable?

Nej! Det påvirker variansen af estimaterne. (Mere om dette nedenfor.)

Udeladte variable

Hvad sker der, hvis man udelader en relevant variabel?

Antag at den "sande model" opfylder MLR.1-MLR.4 og er givet ved:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Men vi estimerer i stedet en simpel lineær regressionsmodel:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{u}$$

Underspecificeret model:

- Vi udelader x_2, selvom det kan være korreleret med x_1, hvilket kan resultere i bias.
- Hvis x_1 og x_2 ikke er korreleret, opfyldes MLR.4 stadig, og vi kan udelade x_2 uden at introducere bias.

Udeladte variable

OLS estimatoren i den “underspecificerede model”:

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

$$= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i)(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 + \frac{\sum_{i=1}^n (\beta_2 x_{i2} + u_i)(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 + \beta_2 \frac{\sum_{i=1}^n x_{i2}(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2} + \frac{\sum_{i=1}^n u_i(x_{i1} - \bar{x}_1)}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2}$$

Udeladte variable

Middelværdien af estimatoren:

$$E(\tilde{\beta}_1|X) = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n} E[x_{i2}(x_{i1} - \bar{x}_1)|X]}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} + \frac{\sum_{i=1}^{n} E[u_i(x_{i1} - \bar{x}_1)|X]}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

$$= \beta_1 + \beta_2 \frac{cov(x_1, x_2)}{var(x_1)}$$

Dvs. OLS estimatoren i en "underspecificeret" model er kun middelret, hvis:

- $\beta_2 = 0$ (modellen er ikke "underspecificeret").
- $cov(x_1, x_2) = 0$ (dvs. at x_1 og x_2 er ukorrelerede).

I alle andre tilfælde er estimatoren **ikke middelret** (også kaldet biased eller skæv).

Udeladte variable

I mange tilfælde er det ikke muligt at inkludere alle relevante variable i estimationen pga. manglende data.

Dog kan vi typisk sige noget om retningen af bias.

TABLE 3.2 Summary of Bias in $\tilde{\beta}_1$ when x_2 Is Omitted in Estimating Eqution (3.40)

	$\mathrm{Corr}(x_1, x_2) > 0$	$\mathrm{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Bias i modeller med flere variable kan være komplekst (se W3A.4).

- En måde at forstå bias i multivariate modeller er ved hjælp af Frisch-Waugh-teoremet. Det ser vi på nu.

Bias ved udeladelse af variable: Multivariate case

Antag en model med flere forklarende variable:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u$$

Udelades x_k, kan estimaterne af $\beta_1, \ldots, \beta_{k-1}$ blive biased.

Bias i estimatet af β_1:

$$E(\tilde{\beta}_1 | X) = \beta_1 + \beta_k \frac{\text{cov}(\hat{r}_1, x_k)}{\text{var}(\hat{r}_1)}$$

hvor $\hat{r}_1$ er residualerne fra regressionen af x_1 på $x_2, \ldots, x_{k-1}$ (brug FWL teoremet).

Fortegn på bias:

- Hvis $\text{cov}(\hat{r}_1, x_k) > 0$, har bias samme fortegn som β_k.
- Hvis $\text{cov}(\hat{r}_1, x_k) < 0$, har bias modsat fortegn af β_k.

Udeladte variable

Quiz Antag at den sande model opfylder MLR.1-MLR.4

$$\log(\textit{timeløn}) = \beta_0 + \beta_1 \textit{uddannelse} + \beta_2 \textit{evner} + u$$

Vi estimerer nu følgende model

$$\log(\textit{timeløn}) = \tilde{\beta}_0 + \tilde{\beta}_1 \textit{uddannelse} + \tilde{u}$$

Ræk hånden op, hvis du forventer følgende:

1. $E(\tilde{\beta}_1) > \beta_1$ (positiv bias)
2. $E(\tilde{\beta}_1) = \beta_1$ (ingen bias)
3. $E(\tilde{\beta}_1) < \beta_1$ (negativ bias)

Udeladte variable: Python Eksempel

Lad os genbesøge løn ligningen i Python!

- Use Jupyter Notebook: `03_mlr_examples.ipynb`
- Part 4: Udeladte variable

Variansen af OLS estimatoren

For lettere at kunne beregne variansen af OLS estimatoren bruger vi fortsat antagelsen om homoskedasticitet.

I MLR modeller er det:

MLR.5 Variansen af fejlleddet er konstant

$$Var(u|x_1, x_2, .., x_k) = \sigma^2$$

Variansen af OLS estimatoren

Med matrixnotation kan antagelserne MLR.5 og MLR.2 skrives som

$$
\begin{aligned}
Var(\mathbf{u}|\mathbf{X}) &= E(\mathbf{u}\mathbf{u}'|\mathbf{X}) \\
&= \begin{pmatrix} Var(u_1|\mathbf{X}) & cov(u_1,u_2|\mathbf{X}) & \cdots & cov(u_1,u_n|\mathbf{X}) \\ cov(u_2,u_1|\mathbf{X}) & Var(u_2|\mathbf{X}) & \cdots & cov(u_2,u_n|\mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(u_n,u_1|\mathbf{X}) & cov(u_n,u_2|\mathbf{X}) & \cdots & Var(u_n|\mathbf{X}) \end{pmatrix} \\
&= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{I}.
\end{aligned}
$$

Hvor vi også har benyttet MLR.4: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$)

Variansen af OLS estimatoren: bevis

Som i SLR antager vi, at variansen **ikke** afhænger af x'erne, og at kovariansen mellem u_i og u_j er nul for alle $i \neq j$.

Antagelserne MLR.1-MLR.5 kaldes **Gauss-Markov** antagelserne.

Teorem E2 Variansen af OLS estimatoren

Under antagelse af MLR.1-MLR.5 er variansen af OLS estimatoren

$$Var(\hat{\beta}|\mathbf{X}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

Variansen af OLS estimatoren: bevis

Ingredienser til bevis

1. Regneregler ($\mathbf{A}$ og $\mathbf{B}$ matricer, $\mathbf{x}$ stok. vektor, c skalar)
   1.1 $c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$
   1.2 $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ og $(\mathbf{A}')' = \mathbf{A}$
   1.3 $\mathbf{AA}^{-\mathbf{1}} = \mathbf{I}$ og $\mathbf{A}^{-\mathbf{1}}\mathbf{A} = \mathbf{I}$
   1.4 $\mathbf{A}$ symmetrisk ($\mathbf{A} = \mathbf{A}'$) så er $\mathbf{A}^{-\mathbf{1}}$ (hvis den eksisterer) det også.
   1.5 $Var(\mathbf{A}x) = \mathbf{A}Var(x)\mathbf{A}'$
2. Antagelser
   2.1 MLR.1: Model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
   2.2 MLR.3: $\mathbf{X}$ fuld rang
   2.3 MLR.4 og MLR.2: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$
   2.4 MLR.5 og MLR.2: $Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}$
3. Estimator: $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-\mathbf{1}}\mathbf{X}'\mathbf{y}$

Variansen af OLS estimatoren: bevis

Variansen af OLS estimatoren: bevis

1. **Indsæt sandt** y **i** $\hat{\beta}$:

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

2. **Tag variansen af** $\hat{\beta}$ **givet** $\mathbf{X}$:

$$\begin{aligned}\mathrm{var}(\hat{\beta}|\mathbf{X}) &= \mathrm{var}\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \mid \mathbf{X}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\,\mathrm{var}(u|\mathbf{X})\,\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

$\mathrm{var}(\hat{\beta}|\mathbf{X})$ er proportional med σ^2 og $(\mathbf{X}'\mathbf{X})^{-1}$.

Variansen af OLS estimatoren

Variansen for hvert $\hat{\beta}_j$ kan udregnes som:

$$\text{Var}(\hat{\beta}_j|X) = \frac{\sigma^2}{SST_j(1-R_j^2)} \quad \text{for } j = 1, \dots, k$$

hvor

- $SST_j = \sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$
- R_j^2 er R^2 for en regression af x_j på de øvrige x'er.

Vi kan omskrive udtrykket til

$$\text{Var}(\hat{\beta}_j|X) = \frac{\sigma^2}{SST_j \cdot (SSR_j/SST_j)} = \frac{\sigma^2}{SSR_j}$$

hvor $SSR_j = \sum_{i=1}^{n}(\hat{r}_{ij} - \bar{r}_j)^2$. Dette følger af Frisch-Waughs Teorem.

Variansen af OLS estimatoren

Husk:

$$\hat{\beta}_j^{SLR} = \frac{cov(x_j, y)}{var(x_j)} \qquad \rightarrow \qquad var(\hat{\beta}_j^{SLR}|X) = \frac{\sigma^2}{\sum(x_j - \bar{x}_j)^2}$$

$$\hat{\beta}_j^{MLR} = \frac{cov(\hat{r}_j, y)}{var(\hat{r}_j)} \qquad \rightarrow \qquad var(\hat{\beta}_j^{MLR}|X) = \frac{\sigma^2}{\sum(\hat{r}_j - \bar{\hat{r}}_j)^2}$$

hvor $\hat{r}_j$ er residualerne fra en regression af x_j på alle de andre $x'er$.

Variansen af OLS estimatoren

Quiz: Antag følgende regressionsmodel:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

hvor variansen på β_1 er givet ved:

$$\text{Var}(\hat{\beta}_1 | X) = \frac{\sigma^2}{SST_1(1 - R_2^1)}$$

Spørgsmål: Ræk hånden op, hvis variansen af $\hat{\beta}_1$ er *mindst* når:

1. x_2 er positivt korreleret med x_1
2. x_2 er ukorreleret med x_1
3. x_2 er negativt korreleret med x_1

Estimation af variansen af OLS

Vi mangler nu at estimere $\hat{\sigma}^2$.

Estimatoren for variansen af fejlleddet estimeres ved residualerne

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\hat{u}_i)^2}{n-k-1} = \frac{SSR}{n-k-1}. \tag{2}$$

Variansen består af

- Tælleren: Summen af de kvadrerede residualer (SSR)
- Nævneren: Antal obs - antal parametre $= n-(k+1) = n-k-1$

Nævneren sikrer, at estimatoren er middelret (frihedsgradskorrektion).

Teorem 3.3: Middelret estimator af variansen af OLS

Hvis Gauss-Markov antagelserne (MLR.1-MLR.5) holder, så er (2) en middelret estimator for fejlledsvariansen.

Variansen af OLS estimatoren: Udeladte og irrelevante variable

Hvordan påvirker udeladelsen af en variabel variansen på de øvrige paramemeter estimater?

Betragt igen modellen

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

Lad $\hat{\beta}_1$ være estimatet fra en regression af y på både x_1 og x_2, og $\tilde{\beta}_1$ estimatet fra en regression af y på kun x_1.

Umiddelbart har vi at

$$var(\hat{\beta}_1|X) = \frac{\sigma^2}{SST_1(1 - R_1^2)} > \frac{\sigma^2}{SST_1} = var(\tilde{\beta}_1|X)$$

Hvis x_1 og x_2 er korrelerede.

Variansen af OLS estimatoren: Udeladte og irrelevante variable

Vi står med andre ord overfor et "**bias-variance**" trade-off.

Hvis vi inkluderer ekstra variable mindsker vi udeladt variabel bias, men øger umiddelbart variansen.

- Det kan være problematisk i små datasæt.
- Derfor kan det nogle gange betale sig at smide variable (med $\hat{\beta}_j \approx 0$) ud af regressionen for at reducere variansen af de øvirge estimater.

I praksis er det dog ikke helt så simpelt.

- σ^2 er variansen af fejlleddet.
- I den simple model uden x_2, er x_2 indeholdt i fejlleddet.
- Derfor vil σ^2 ofte stige, når vi undlader x_2.
- Nogle gange stiger σ^2 nok til at øge variansen af $\hat{\beta}_j$.

Variansen af OLS estimatoren: Udeladte og irrelevante variable

```
. estimates table reg_SLR reg_MLR1 reg_MLR2, b(%7.0g) se(%7.0g)

--------------------------------------------
    Variable | reg_SLR   reg_M~1   reg_M~2
-------------+------------------------------
        educ |  .02821    .02813    .02743
             |  .00283    .00266    .00265
  experience |            .01246    .02536
             |            .00103    .00395
 experience2 |                     -.00041
             |                      .00012
       _cons |  4.5604    4.3791    4.3155
             |  .03374    .03501     .0396
--------------------------------------------
                                legend: b/se
```

OLS er BLUE

Gauss-Markov Teoremet

Teorem 3.4: Gauss-Markov Teoremet

Under antagelse af MLR.1–MLR.5 er OLS estimatoren den bedst lineære middelrette estimator (**BLUE**).

- **Best:** Mindste varians blandt lineære, middelrette estimatore
- **Linear:** Estimatoren er lineær i y, dvs. $\tilde{\beta} = A\mathbf{y}$
- **Unbiased:** $E(\tilde{\beta}) = \beta$
- **Estimator:** Bruger data til at estimere parametre

Gauss-Markov teoremet

Hvad vil det sige at en estimator er lineær?

OLS estimatoren i den simple lineære regressionsmodel

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n y_i w_i, \tag{3}$$

OLS estimatoren kan således skrives som en vægtet sum af y'erne, hvor $w_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$.

Det er netop definitionen på en lineær estimator.

Om vægtene gælder:

$$\sum_{i=1}^n w_i = 0, \quad \sum_{i=1}^n w_i x_i = 1 \tag{4}$$

Gauss-Markov Teoremet

Hvad vil det sige, at en estimator er lineær?

OLS estimatoren i den simple lineære regressionsmodel:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n y_i w_i, \tag{5}$$

OLS estimatoren kan skrives som en vægtet sum af y'erne, hvor $w_i = \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$. Dette er definitionen på en lineær estimator.

Om vægtene gælder:

$$\sum_{i=1}^n w_i = 0, \quad \sum_{i=1}^n w_i x_i = 1 \tag{6}$$

Disse egenskaber sikrer, at estimatoren er uafhængig af konstantleddet og korrekt skaleret i forhold til x_i.

Gauss-Markov Teoremet

Vi kan konstruere en alternativ lineær estimator ved at vælge andre vægte:

$$\tilde{\beta}_1 = \sum_{i=1}^{n} y_i \tilde{w}_i \tag{7}$$

Under MLR.1-MLR.4 vil den alternative estimator kun være middelret, hvis vægtene opfylder følgende betingelser:

$$\sum_{i=1}^{n} \tilde{w}_i = 0, \quad \sum_{i=1}^{n} \tilde{w}_i x_i = 1 \tag{8}$$

Spørgsmål: Hvorfor skal disse betingelser opfyldes for at estimatoren er middelret? (Næste slide)

Lineær estimator:

$$\tilde{\beta}_1 = \sum y_i \tilde{w}_i$$

1) Indsæt sandt y:

$$\tilde{\beta}_1 = \sum (\beta_0 + \beta_1 x_i + u_i)\tilde{w}_i = \beta_0 \sum \tilde{w}_i + \beta_1 \sum x_i \tilde{w}_i + \sum u_i \tilde{w}_i$$

2) Tag $E(\tilde{\beta}_1|X)$:

$$E(\tilde{\beta}_1|X) = \beta_0 \sum \tilde{w}_i + \beta_1 \sum x_i \tilde{w}_i + \sum E(u_i|X)\tilde{w}_i$$

Under MLR.4, $E(u_i|X) = 0$, så for at estimatoren skal være middelret:

$$E(\tilde{\beta}_1|X) = \beta_1 \quad \text{hvis} \quad \sum \tilde{w}_i = 0 \quad \text{og} \quad \sum x_i \tilde{w}_i = 1$$

Gauss-Markov teoremet

En alternativ lineær middelret estimator (i SLR model):

$$\begin{aligned}\tilde{w}_1 &= \frac{-1}{x_1 - x_n} \\ \tilde{w}_j &= 0 \quad j = 2, .., n-1 \\ \tilde{w}_n &= \frac{1}{x_1 - x_n}\end{aligned}$$

Den alternative estimator

$$\hat{\beta}_1 = \frac{y_n - y_1}{x_n - x_1}.$$

Estimatoren er middelret

- ...men vi foretrækker OLS, fordi den har mindre varians.
- Der findes alternative "ikke-åndsvage" estimatorer, fx matching.

Gauss-Markov teoremet: Bevis

Ingredienser til bevis

1. Regneregler ($\mathbf{A}$ og $\mathbf{B}$ matricer, $\mathbf{x}$ stok. vektor, c skalar)
   1.1 $c\mathbf{AB} = \mathbf{A}c\mathbf{B} = \mathbf{AB}c$
   1.2 $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ og $(\mathbf{A}')' = \mathbf{A}$
   1.3 $\mathbf{AA}^{-1} = \mathbf{I}$ og $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
   1.4 $Var(\mathbf{A}x) = \mathbf{A}Var(x)\mathbf{A}'$
   1.5 En kvadratisk, symmnetrisk matrix ($\mathbf{A} = \mathbf{A}'$) er positiv semi definit.
   1.6 En idempotent matrix ($\mathbf{A} = \mathbf{AA}$) er kvadratisk.
   1.7 Hvis $\mathbf{A}$ er positiv semi definit og $\mathbf{q}'\mathbf{Aq}$ en skalar, gælder $\mathbf{q}'\mathbf{Aq} > 0$.
2. Antagelser
   2.1 MLR.1: Model $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
   2.2 MLR.3: $\mathbf{X}$ fuld rang
   2.3 MLR.4 og MLR.2: $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$
   2.4 MLR.5 og MLR.2: $Var(\mathbf{u}|\mathbf{X}) = \sigma^2\mathbf{I}$
3. Variansen af OLS estimator: $Var(\hat{\beta}|\mathbf{X}) = \sigma^2\mathbf{X}'\mathbf{X}^{-1}$

Gauss-Markov teoremet: Bevis

Vi starter med at betragte alle lineære middelrette estimatorer (LUE).

Alle estimatorer i denne kategori kan skrives som

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'\mathbf{y} \quad \text{hvor} \quad \mathbf{A}'\mathbf{X} = \mathbf{I}$$

Hvorfor det?

$$\tilde{\boldsymbol{\beta}} = \mathbf{A}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'\mathbf{u}$$

$$E(\tilde{\boldsymbol{\beta}}|X) = \mathbf{A}'\mathbf{X}\boldsymbol{\beta} + \mathbf{A}'E(\mathbf{u}|\mathbf{X}) = \boldsymbol{\beta}$$

Under MLR.4: $E(\mathbf{u}|\mathbf{X}) = 0$ og derfor $\mathbf{A}'\mathbf{X} = I$ for at bevare middelrethed.

Gauss-Markov teoremet: Bevis

Nu skal vi kigge på "Best" delen.

Til det skal vi bruge variansen på vores (alternative) LUE

$$Var(\tilde{\beta}|X) = Var(\mathbf{A}'\mathbf{y}|\mathbf{X}) = Var(\mathbf{A}'\mathbf{X}\beta|X) + Var(\mathbf{A}'\mathbf{u}|\mathbf{X})$$

Første led er 0 (fordi $Var(\mathbf{A}'\mathbf{X}\beta|X) = 0$).

Andet led:

$$Var(\mathbf{A}'\mathbf{u}|\mathbf{X}) = \mathbf{A}'Var(\mathbf{u}|\mathbf{X})\mathbf{A} = \sigma^2\mathbf{A}'\mathbf{A} \quad \text{(MLR.5)}$$

Gauss-Markov teoremet: Bevis

Dvs. forskellen i variansen på vores alternative LUE og OLS er

$$Var(\tilde{\beta}|X) - Var(\hat{\beta}|X) = \sigma^2 \mathbf{A}'\mathbf{A} - \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

Brug $\mathbf{A}'\mathbf{X} = I$ til at gange fra venstre og højre:

$$= \sigma^2 \mathbf{A}' \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \right] \mathbf{A}$$

Hvor $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, der er symmetrisk og idempotent ($\mathbf{MM} = \mathbf{M}$).

$\mathbf{M}$ er positiv semidefinit, så $\mathbf{A}'\mathbf{M}\mathbf{A} \geq 0$.

Ovenstående beviser Gauss-Markov Teoremet:

$$\mathbf{q}' \left[Var(\tilde{\boldsymbol{\beta}}|X) - Var(\hat{\boldsymbol{\beta}}|X) \right] \mathbf{q} \geq 0 \quad \text{for alle } \mathbf{q}.$$

Hvorfor er Gauss-Markov Teoremet vigtigt?

Teoremet siger, at hvis $\tilde{\boldsymbol{\beta}}$ er en anden lineær middelret estimator, så vil OLS estimatoren $\hat{\boldsymbol{\beta}}$ altid have mindre varians:

$$Var(\hat{\beta}_j|X) \leq Var(\tilde{\beta}_j) \quad \text{for } j = 0, \ldots, k$$

Desuden gælder det, at for enhver kombination af parameterestimatorerne er variansen af OLS mindre, fx

$$Var(\hat{\beta}_i - \hat{\beta}_j|X) \leq Var(\tilde{\beta}_i - \tilde{\beta}_j|X)$$

Gauss-Markov teoremet: Relevance

Hvorfor er det relevant at beregne

$$\mathbf{q}' Var(\hat{\beta}|\mathbf{X})\mathbf{q}$$

Eksempler:

$$\begin{aligned} var(\hat{\beta}_1|\mathbf{X}) \\ var(\hat{\beta}_1 - \hat{\beta}_2|\mathbf{X}) \\ var(a\hat{\beta}_1 + b\hat{\beta}_2|\mathbf{X}) &= a^2 var(\hat{\beta}_1|\mathbf{X}) + b^2 var(\hat{\beta}_2|\mathbf{X}) + 2abcov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) \end{aligned}$$

Gauss-Markov teoremet: Relevance

Lad: $\mathbf{q} = \begin{bmatrix} a \\ b \end{bmatrix}$ (vektor af vægte)

$$Var(\hat{\beta}|\mathbf{X}) = \begin{pmatrix} var(\hat{\beta}_1|\mathbf{X}) & cov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) \\ cov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) & var(\hat{\beta}_2|\mathbf{X}) \end{pmatrix}$$

Så gælder:

$$\begin{aligned} \mathbf{q}' Var(\hat{\beta}|\mathbf{X})\mathbf{q} &= (\mathbf{a}\ \mathbf{b}) \begin{pmatrix} var(\hat{\beta}_1|\mathbf{X}) & cov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) \\ cov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) & var(\hat{\beta}_2|\mathbf{X}) \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} \\ &= a^2 var(\hat{\beta}_1|\mathbf{X}) + b^2 var(\hat{\beta}_2|\mathbf{X}) + 2ab cov(\hat{\beta}_1, \hat{\beta}_2|\mathbf{X}) \end{aligned}$$

Opsummering

Opsummering

Meget af hvad vi lærte ved SLR gælder også for MLR

- MLR.1-MLR.4 $\implies$ OLS er middelret.
- MLR.5 $\implies$ Vi kan udregne variansen af OLS.

Frish-Waughs Teorem

- MLR kan fortolkes som SLR, hvor vi "renser" x'erne for deres korrelation med andre variable.

Udeladte og irrelevante variable

- Inklusion af irrelevante variable medfører ikke bias, men risikerer at øge variansen på de øvrige estimater.
- Udeladelsen af relevante variable medfører bias, hvis de er korreleret variablene i modellen.

Opsummering: OLS er BLUE