

Den simple lineære regression model (SLR)

Økonometri A

Bertel Schjerning

September 6, 2024

Program

Definitioner og antagelser (W2.1)

Udledningen af OLS estimatoren (W2.2)

Egenskaber ved OLS (W2.3+W2.6)

Fordelingen af OLS estimaterne (W2.5)

- Middelrethed

- Varians

Måleenheder og funktionelle former (W2.4)

Data visualisering

Motivation

Vi er interesseret i at kende (den kausale) sammenhænge mellem et outcome (y) og en forklarende variable (x)

For eksempel:

- Hvordan påvirker gødning produktionen af sojabønner?
- Hvordan påvirker uddannelse timelønnen?
- Hvordan påvirker “kvaliteten” af den administrerende direktør profitten i en virksomhed?
- Hvordan afhænger væksten i BNP af landenes initiale BNP?

Definitioner og antagelser

Den simple lineære regression model

Med SLR antager vi at sammenhængen mellem y og x i “populationsmodellen” er lineær:

$$y = \beta_0 + \beta_1 x + u \quad (1)$$

y : Afhængig variable

x : Forklarende variable

u : Unobserveret fejllid

β_0 : Intercept (konstantled)

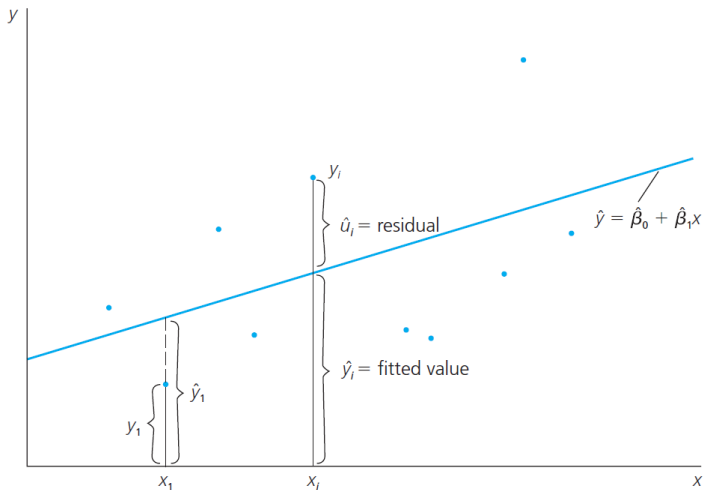
β_1 : Hældningskoefficient

Vi er typisk mest interesseret i β_1 , som måler styrken af sammenhængen mellem y og x):

$$\Delta y = \beta_1 \Delta x \quad \text{hvis} \quad \Delta u = 0$$

Den simple lineære regression model

FIGURE 2.4 Fitted values and residuals.



Den simple lineære regressionsmodel

Vigtige overvejelser ved specifikation af en lineær regressionsmodel:

Spg. 1: Hvordan håndterer vi andre faktorer end x , der påvirker y ?

Spg. 2: Hvilken funktionel form beskriver bedst sammenhængen mellem x og y ?

- Skal y afhænge af $\log(x)$, x^2 , $1/x$, eller en anden funktion af x ?
- Skal vi modellere $\log(y)$ som funktion af x ?
- Kan vi antage, at y er lineær i parametrene?

Spg. 3: Kan β_1 fortolkes som en *ceteris paribus* (kausal) effekt?

Den simple lineære regressionsmodel

Spg. 1: Hvordan håndterer vi, at andre faktorer end x påvirker y ?

Vi antager, at alle andre faktorer, der påvirker y , er indeholdt i fejlleddet u :

$$y = \beta_0 + \beta_1 x + u$$

Diskussion:

- Hvad indeholder fejlleddet u i eksemplet, hvor y er udbyttet af sojabønner, og x er gødningsmængden?
- Hvad indeholder fejlleddet u i eksemplet, hvor y er timeløn, og x er uddannelse (målt i år)?

Spg. 2: Hvilken funktionel form beskriver bedst sammenhængen mellem x og y ?

- Vi antager, at y er en **lineær** funktion af x .
- Denne lineære antagelse kan være restriktiv og ikke altid passende.
- Eksempel: Hvordan ser sammenhængen mellem gødning og sojabønner ud? (Overvej aftagende marginalafkast)

Vi vil senere se, hvordan vi kan lempe antagelsen om linearitet.

Den simple lineære regressionsmodel

- Hvorfor tror I, at sammenhængen mellem temperatur og elforbrug er ikke-lineær?
- Er modellen $y = \beta_0 + \beta_1x + \beta_2x^2 + u$ mere passende?

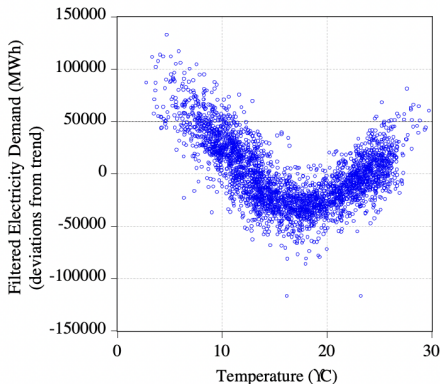


Fig. 2. Non-linearity in electricity demand response to temperature variations.

Den simple lineære regressionsmodel

Spg. 3: *ceteris paribus*/kausale fortolkninger

Kausale fortolkninger kræver ekstra antagelser. β_1 beskriver, hvordan y afhænger af x :

$$\Delta y = \beta_1 \Delta x, \quad \text{hvis} \quad \Delta u = 0$$

Hvornår kan dette være et problem?

- **Eks 1:** Gødning og udbytte? Mere gødning på dårlig jord.
- **Eks 2:** Timeløn og uddannelse? Højere evner kan føre til både mere uddannelse og højere løn.
- I begge tilfælde er ($\Delta u \neq 0$), og vi risikerer at overvurdere effekten af x på y .

Antagelser om u er centrale i økonometri, men ofte svære at validere. 9

Antagelser om fejleddet

Alle andre faktorer end x , som påvirker y , er indeholdt i fejleddet u .

Med SLR laver vi to antagelser om u :

$$E(u) = 0$$

$$E(u|x) = 0$$

Den første antagelse er ret uskyldig – vi normaliserer typisk $E(u)$ til 0

- I sammenhængen mellem timeløn (y) og uddannelse (x) vil evner/intelligens være indeholdt i u .
- Hvis vi normaliserer $E(u) = 0$, vil $\beta_0 = E(y)$ for en person med gennemsnitlig intelligens og uden uddannelse ($x = 0$).

Antagelser om fejleddet

Den anden antagelse er kritisk: $E(u|x) = 0$

Det er antagelsen om (**“zero conditional mean assumption”**)

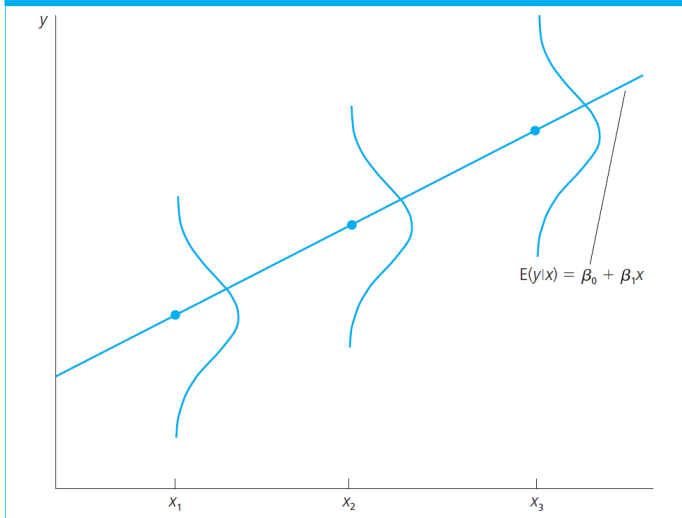
Hvad betyder antagelsen?

- Den forventede værdi af u er uafhængig af x for alle x .
- Det er en stærkere antagelse end at $cov(u, x) = 0$.
- Omvendt antager vi ikke at x og u er generelt uafhængige
- Fx har vi *ikke* antaget $E(u^2|x) = 0$ (konstant varians).
- Generel uafhængighed \Rightarrow alle funktioner af u er uafhængige af x

Med de to antagelser gælder at

$$E(y|x) = \beta_0 + \beta_1 x + \underbrace{E(u|x)}_{=0}$$

FIGURE 2.1 $E(y|x)$ as a linear function of x .



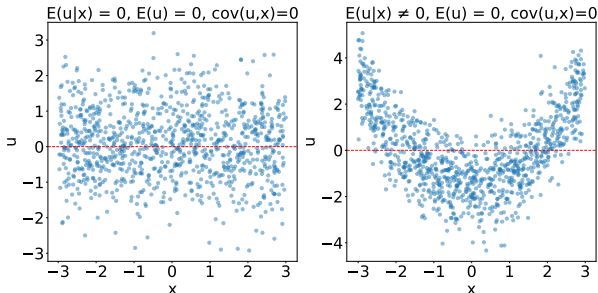
Antagelser om fejleddet

Alle andre faktorer end x , som påvirker y er indeholdt i fejleddet u .

Med SLR vi laver to antagelser om u :

$$E(u) = 0 \quad (\text{scalar})$$

$$E(u|x) = 0 \quad (\text{funktion af } x)$$



Bemærk at $\text{cov}(u, x) = 0 \not\Rightarrow E(u|x) = 0$

Antagelser om fejleddet

Eksempel: Timeløn og uddannelse.

- Model

$$\text{timeløn} = \beta_0 + \beta_1 \text{uddannelse} + u$$

- Fejleddet indeholder intelligens og andre typer evner.
- Kan vi rimeligvis antage at $E[u|x] = 0$?
- Er følgende antagelser rimelige?

$$E(\text{evner} | \text{uddannelse} = 9) = E(\text{evner} | \text{uddannelse} = 17)$$

$$E(\text{evner} | \text{folkeskolen}) = E(\text{evner} | \text{kandidatudd})$$

- Hvad med erhvervserfaring, motivation, familieforhold, uddannelseskvalitet, helbred, jobkarakteristika, geografisk placering, og netværk?

Udledningen af OLS estimatoren

Estimation af parametrene i SLR

Vi ønsker at estimere β_0 og β_1 i **populationsmodellen**:

$$y = \beta_0 + \beta_1 x + u$$

En ligning med 3 ubekendte, som kan estimeres på forskellige måder:

1. Maximum Likelihood Estimation (MLE):

- Antag en fordeling for u , fx normalfordeling.
- Opskriv og maksimer likelihood-funktionen mht. β_0 og β_1 .

2. Method of Moments (MM):

- Kræver kun at $E(u|x) = E(u) = 0$.
- Kræver ingen antagelser om fordelingen af u .

Eksempel: Estimation af den forventede værdi

- Variabel y med $E(y) = \mu$.
- μ er ukendt, og vi ønsker at estimere den.
- Vi har n observationer af y .
- Hvad er en naturlig estimator for μ ?

Den naturlige estimator er den empiriske middelværdi (gennemsnit):

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n y_i$$

Som er en middelfret estimator for μ (se math refresher C2).

Ideen bag MM: Erstat det teoretiske moment (fx middelværdi) med det empiriske moment.

Ideen bag momentestimation

Variansen af y , hvor $E(y) = \mu_y$, er:

$$\text{Var}(y) = \sigma_y^2 = E[(y - \mu_y)^2]$$

MM-estimator:

$$\hat{\sigma}_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2$$

Kovariansen mellem x og y :

$$\text{Cov}(x, y) = \sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$$

MM-estimator:

$$\hat{\sigma}_{xy} = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

Udledningen af OLS-estimatoren

Regression model:

$$y = \beta_0 + \beta_1 x + u$$

Antagelser:

$$E(u) = 0 \quad (2)$$

$$E(u|x) = 0 \quad (3)$$

Det gælder at $E(u|x) = 0 \implies \text{cov}(x, u) = 0$

$$\begin{aligned} \text{cov}(x, u) &\equiv E[(x - E(x))(u - E(u))] \\ &= E[(x - E(x))u] \\ &= E(xu) - E(x)E(u) \\ &= E(xu) = 0 \end{aligned}$$

Udledningen af OLS estimatoren

Vi finder først de relevante populationsmomenter.

Udtryk fejlleddet u ved hjælp af modellen:

$$u = y - \beta_0 - \beta_1 x.$$

Indsæt dette i antagelserne (2) og (4):

$$E(u) = E(y - \beta_0 - \beta_1 x) = 0$$

$$E(xu) = E(x(y - \beta_0 - \beta_1 x)) = 0$$

Dette er de teoretiske momenter (populationsmomenter).

Udledningen af OLS-estimatoren

Vi kan estimere β_0 og β_1 ved at minimere residualkvadratsummen:

$$\hat{\beta}_0, \hat{\beta}_1 = \arg \min_{\beta_0, \beta_1} \underbrace{\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2}_{Q_n(\beta_0, \beta_1)}$$

Deraf navnet **Ordinary Least Squares (OLS)**.

Førsteordensbetingelserne er identiske med momentbetingelserne:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \Rightarrow \sum_{i=1}^n u_i = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \Rightarrow \sum_{i=1}^n u_i x_i = 0$$

Udledningen af OLS estimatoren: Regneregler

Ingredienser til udledningen:

Regneregler for summer:

$$\frac{1}{n} \sum_{i=1}^n x_i = \bar{x}$$

$$\frac{1}{n} \sum_{i=1}^n ax_i = a \frac{1}{n} \sum_{i=1}^n x_i = a\bar{x}$$

$$\frac{1}{n} \sum_{i=1}^n a = a \frac{1}{n} \sum_{i=1}^n 1 = a$$

$$\frac{1}{n} \sum_{i=1}^n x_i(y_i + z_i) = \frac{1}{n} \sum_{i=1}^n x_i y_i + \frac{1}{n} \sum_{i=1}^n x_i z_i$$

Udledningen af OLS-estimatoren: Regneregler

Regneregler til brug i udledningen:

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \frac{1}{n} \left(\sum_{i=1}^n x_i - n\bar{x} \right) = \frac{1}{n} (n\bar{x} - n\bar{x}) = 0$$

$$\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})\bar{x} = \bar{x} \cdot \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) = \bar{x} \cdot 0 = 0$$

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})x_i &= \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x})x_i - (x_i - \bar{x})\bar{x}] \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = \text{Var}(x) \end{aligned}$$

Udledningen af OLS-estimatoren: $\hat{\beta}_0$

Trin 1: Udledning af $\hat{\beta}_0$

$$E(u) = 0 \implies \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{n} \sum_{i=1}^n \hat{\beta}_0 + \frac{1}{n} \sum_{i=1}^n \hat{\beta}_1 x_i \iff \bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Vi har nu udtrykt $\hat{\beta}_0$ som en funktion af $\hat{\beta}_1, \bar{y}$ og \bar{x}

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Udledningen af OLS-estimatoren: $\hat{\beta}_1$

Trin 2: Udledning af $\hat{\beta}_1$

$$E(u \cdot x) = 0 \implies \frac{1}{n} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i = 0$$

Indsæt $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$:

$$\frac{1}{n} \sum_{i=1}^n \left[(y_i - \bar{y}) + \hat{\beta}_1 (\bar{x} - x_i) \right] x_i = 0$$

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

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

Udledningen af OLS-estimatoren

GOOOOAAAAL! Vi har nu udledt OLS-estimatoren for β_1 og β_0 :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\widehat{cov}(x_i, y_i)}{\widehat{var}(x_i)} \quad (5)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (6)$$

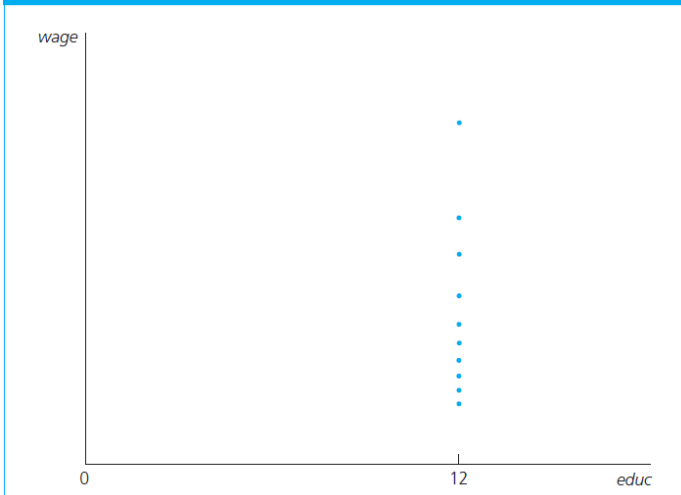
Bemærk: Vi kræver, at $\widehat{var}(x_i) > 0$ for at OLS-estimatoren er veldefineret:

$$SST_x \equiv \sum_{i=1}^n (x_i - \bar{x})^2 > 0$$

- Hvorfor er det vigtigt?
- Hvorfor kan det være en problematisk antagelse?

Udledningen af OLS estimatoren: Variation i x

FIGURE 2.3 A scatterplot of wage against education when $educ_i = 12$ for all i .





- Jupyter Notebook: `02_slr_examples.ipynb`
- Part 1: Timeløn og uddannelse (OLS estimation)

Egenskaber ved OLS

Prædikterede værdier og residualer

Ud fra parameterestimerterne kan vi finde den prædikterede værdi af y :

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residualerne kan beregnes som:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Bemærk: \hat{y}_i er vores bud på $E(y|x_i)$, men \hat{y}_i vil sjældent være lig y .

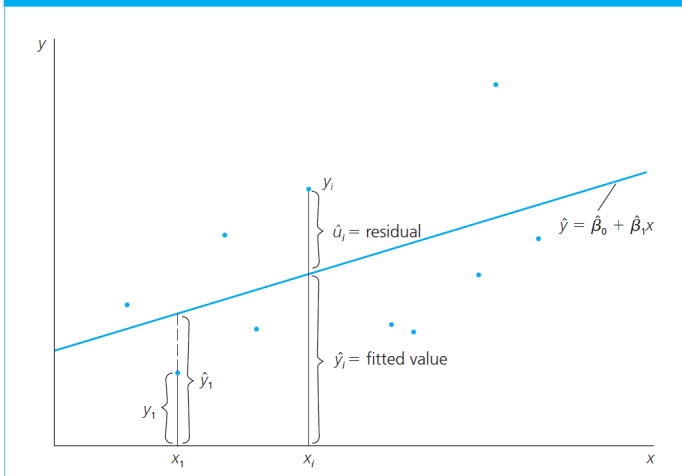
Egenskaber ved OLS residualer:

- $\sum_{i=1}^n \hat{u}_i = 0$
- $\sum_{i=1}^n \hat{u}_i x_i = 0$

Hvorfor er dette ikke så overraskende?

Prædikterede værdier og residualer

FIGURE 2.4 Fitted values and residuals.



Quiz

Lad $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ og $\hat{u}_i = y_i - \hat{y}_i$

Ræk hånden op, hvis du mener følgende er SAND:

$$A : \frac{1}{n} \sum_{i=1}^n \hat{y}_i = \frac{1}{n} \sum_{i=1}^n y_i$$

$$B : \frac{1}{n} \sum_{i=1}^n y_i = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{n} \sum_{i=1}^n x_i$$

$$C : \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i) \hat{u}_i = 0$$

Et relevant spørgsmål:

Hvor meget af variationen i y vi kan forklare med x ?

Til det formål definerer vi følgende:

- Total sum of squares (i y): $SST \equiv \sum_{i=1}^n (y_i - \bar{y})^2$.
- Explained sum of squares: $SSE \equiv \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$.
- Residual sum of squares: $SSR \equiv \sum_{i=1}^n \hat{u}_i^2$.

Det gælder at den totale variation (kvadratsum) kan skrives som

$$SST = SSE + SSR$$

Vi kan således dekomponere SST i en forklaret og i en residual del.

En naturlig måde at beregne, hvor meget af variationen i y vi kan forklare med x , er således

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

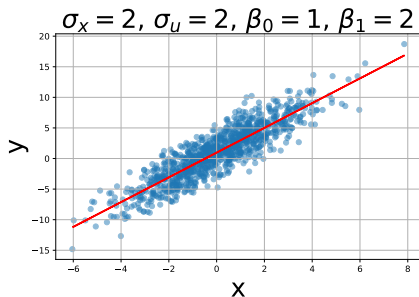
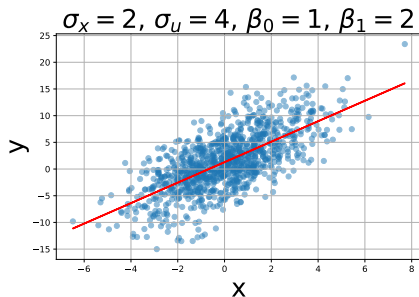
hvor $0 \leq R^2 \leq 1$.

Bemærk at vores model sagtens kan være relevant selvom R^2 er lav.

Goodness of fit

Quiz: I hvilken figur er R^2 størst?

$$y = \beta_0 + \beta_1 x + u, \quad x \sim N(0, \sigma_x^2), \quad u \sim N(0, \sigma_u^2)$$





- Jupyter Notebook: `02_slr_examples.ipynb`
- Part 2: Timeløn og uddannelse (Goodness of fit)

Fordelingen af OLS estimerterne

Fordelingen af OLS estimerterne

OLS estimatoren er en maskine:

Input (Data)

Sample 1: $\{(y_1, x_1), \dots, (y_n, x_n)\}$

\vdots

Sample k: $\{(y_1, x_1), \dots, (y_n, x_n)\}$

\rightarrow **OLS** \rightarrow \vdots

Output (Estimator)

$(\hat{\beta}_0, \hat{\beta}_1)_1$

$(\hat{\beta}_0, \hat{\beta}_1)_k$

Hvad er fordelingen af OLS estimerterne?

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1) \sim F(\theta)$$

- Hvilken fordeling $F()$
(eksempelvis t-fordeling eller normalfordeling)
- og hvilke parametre θ , som $F(\theta)$ afhænger af
(eksempelvis middelværdi, μ og varians σ^2 , frihedsgrader)

Definition (se appendix C2)

Antag vi har en estimator \mathbf{b} for β .

\mathbf{b} er en **middelret** (unbiased) estimator for β , hvis:

$$E(\mathbf{b}) = \beta$$

for alle værdier af β .

- Middelrethed er en statistisk egenskab, der sikrer, at estimatorens forventede værdi er lig med den sande parameter.
- Hvorfor er det vigtigt, at en estimator er middelret?
- Kan en estimator være brugbar, selv hvis den ikke er middelret?

Centrale antagelse for den simple lineære regressionsmodel

SLR.1 Populationsmodellen er lineær i parametrene:

$$y = \beta_0 + \beta_1 x + u.$$

SLR.2 Tilfældig udvælgelse:

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

SLR.3 Variation i x :

I datasættet må x ikke antage den samme værdi for alle observationer.

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

$$E(u|x) = 0.$$

Teorem 2.1: Middelrethed af OLS estimatoren

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

$$E(\hat{\beta}_0) = \beta_0$$

$$E(\hat{\beta}_1) = \beta_1$$

Middelrethed af OLS estimatoren: Bevis (1)

Ingredienser til beviset:

$$\text{SLR.1 } y = \beta_0 + \beta_1 x + u$$

$$\text{SLR.2 tilfældig stikprøve} \implies E(u_i|x) = E(u|x) \text{ for alle } i.$$

$$\text{SLR.3 } \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$

$$\text{SLR.4 } E(u|x) = 0$$

$$\text{OLS estimatoren: } \hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n y_i (x_i - \bar{x})}{SST_x}$$

Regneregler:

- $\sum_{i=1}^n (x_i - \bar{x}) = 0$
- $E(y) = E(E(y|x))$ (Law of iterated expectations)
- $E(a(x)y + b(x)|x) = a(x)E(y|x) + b(x)$ (linearitet af forventning)

Middelrethed af OLS estimatoren: Bevis (2)

Ved brug af **SLR.1** kan vi skrive OLS estimatoren som

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{SST_x} = \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{SST_x} \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{SST_x} + \beta_1 \frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{SST_x} + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \\ &= \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x}\end{aligned}$$

Middelrethed af OLS estimatoren: Bevis (2)

Ved brug af **SLR.1** kan vi skrive OLS estimatoren som

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n y_i(x_i - \bar{x})}{SST_x} = \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i + u_i)(x_i - \bar{x})}{SST_x} \\ &= \beta_0 \frac{\sum_{i=1}^n (x_i - \bar{x})}{SST_x} + \beta_1 \frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{SST_x} + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \\ &= \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x}\end{aligned}$$

Tag forventning på begge sider, betinget på stikprøven

$X = (x_1, x_2, \dots, x_n)$:

$$\begin{aligned}E(\hat{\beta}_1|X) &= E\left(\beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \middle| X\right) \\ &= \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i|X)}{SST_x} \quad (\text{linearitet af forventning})\end{aligned}$$

Middelrethed af OLS estimatoren: Bevis (3)

$$\text{SLR.2} + \text{SLR.4} \implies E(u_i|X) = 0$$

$$\text{SLR.3} \implies \text{SST}_x \equiv \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$

således at

$$E(\hat{\beta}_1|X) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i|X)}{\text{SST}_x} = \beta_1$$

Pga. law of iterated expectations gælder der, at

$$E(\hat{\beta}_1) = E(E(\hat{\beta}_1|X)) = \beta_1$$

SLR.1-SLR.4 $\implies \hat{\beta}_1$ er en middelret estimator for β_1 .

Home run!!!

Middelrethed af OLS estimatoren: Bevis (3)

$$\text{SLR.2} + \text{SLR.4} \implies E(u_i|X) = 0$$

$$\text{SLR.3} \implies \text{SST}_x \equiv \sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$$

således at

$$E(\hat{\beta}_1|X) = \beta_1 + \frac{\sum_{i=1}^n (x_i - \bar{x}) E(u_i|X)}{\text{SST}_x} = \beta_1$$

Pga. law of iterated expectations gælder der, at

$$E(\hat{\beta}_1) = E(E(\hat{\beta}_1|X)) = \beta_1$$

SLR.1-SLR.4 $\implies \hat{\beta}_1$ er en middelret estimator for β_1 .

Home run!!!

Hvad med $\hat{\beta}_0$?

Middelrethed af OLS estimatoren: Bevis (4)

OLS estimatoren for β_0 og SLR.1 giver

$$\begin{aligned}\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ &= \beta_0 + \beta_1 \bar{x} + \bar{u} - \hat{\beta}_1 \bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u}\end{aligned}$$

Tag betinget forventning betinget på stikprøven, $X = (x_1, x_2, \dots, x_n)$:

$$\begin{aligned}E(\hat{\beta}_0|X) &= E(\beta_0 + (\beta_1 - \hat{\beta}_1) \bar{x} + \bar{u}|X) \\ &= \beta_0 + E(\beta_1 - \hat{\beta}_1|X) \bar{x} + E(\bar{u}|X) = \beta_0\end{aligned}$$

Vi har lige har vist $E(\beta_1 - \hat{\beta}_1|X) = 0$

SLR.2 + SLR.4 giver $E(\bar{u}|X) = 0$

Dermed er $\hat{\beta}_0$ en middelret estimator for β_0

$$E(\hat{\beta}_0) = E[E(\hat{\beta}_0|X)] = \beta_0$$

Variansen af OLS-estimatoren

Selvom OLS-estimatoren er middelret, vil $\hat{\beta}_1$ ofte afvige fra β_1 og variere mellem stikprøver.

- $\hat{\beta}_1$ er en stokastisk variabel med en middelværdi (β_1) og en varians.
- Variansen af $\hat{\beta}_1$ givet stikprøven X , $\text{Var}(\hat{\beta}_1 | X)$, bestemmer estimatets præcision.
- Lav varians betyder, at estimerne er tæt på β_1 på tværs af stikprøver, hvilket øger tilliden til resultaterne.

For at udlede variansen af $\hat{\beta}_1$ kræves en ekstra antagelse.

Variansen af OLS-estimatoren

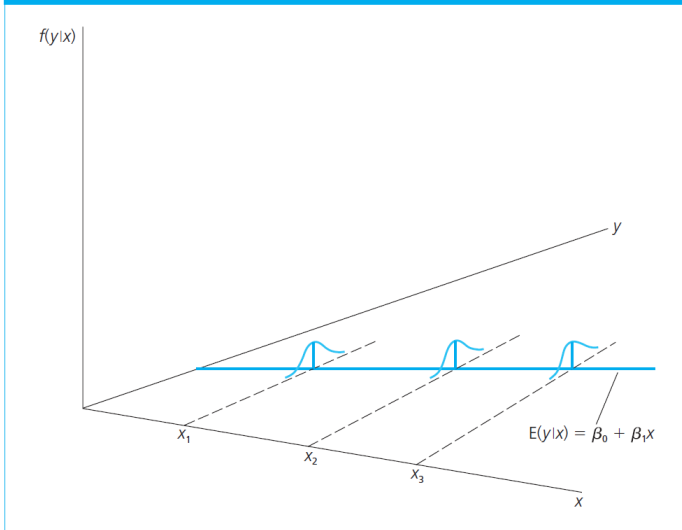
SLR.5 Variansen af fejlleddet er konstant:

$$\text{Var}(u|x) = \sigma^2 \quad (\text{Homoskedasticitet})$$

- **Gauss-Markov:** SLR.1-SLR.5 gør OLS til den bedste lineære, ubiasede estimator (BLUE).
- **Middelrethed:** Kun SLR.1-SLR.4 kræves for, at OLS er middelret.
- **Efficiens:** SLR.5 sikrer, at OLS er efficient med minimal varians.
- Uden SLR.5 er fejlleddene **heteroskedastiske**, hvilket kan gå ud over efficiens, men ikke middelrethed.
- Vi kan let udlede $\text{Var}(\hat{\beta})$ under SLR.5. Uden homoskedasticitet er det mere kompliceret (det ser vi på i kap. 8)

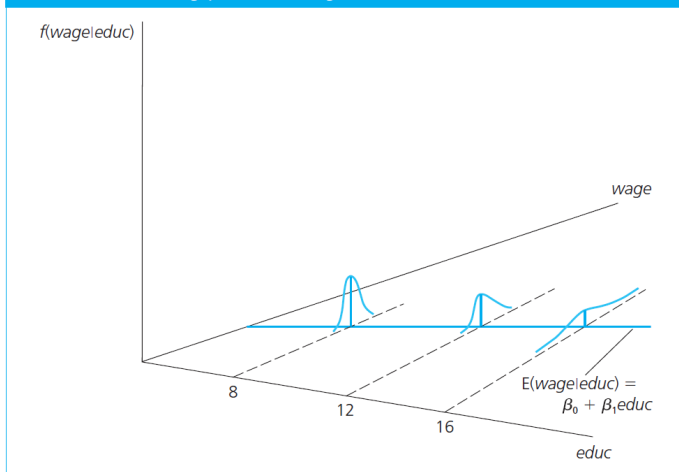
Homoskedasticitet: SLR.5 opfyldt

FIGURE 2.8 The simple regression model under homoskedasticity.



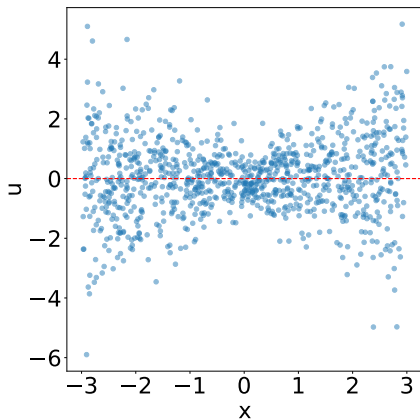
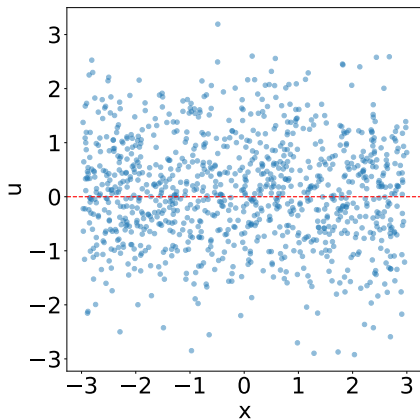
Heteroskedasticitet: SLR.5 ikke-opfyldt

FIGURE 2.9 $\text{Var}(\text{wage}|\text{educ})$ increasing with educ .



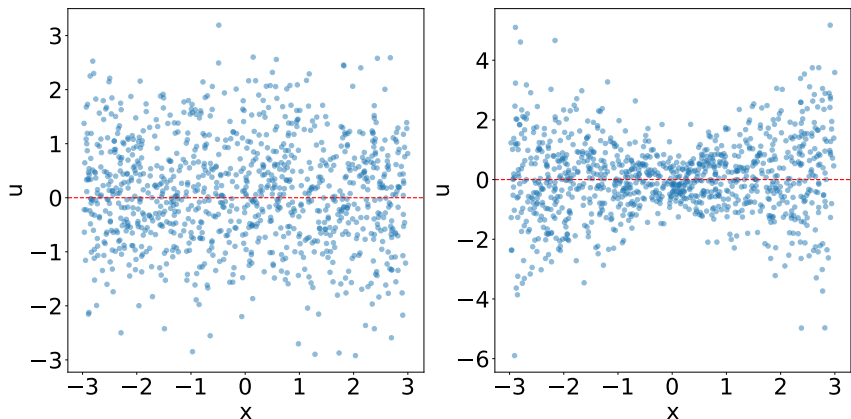
Heteroskedasticitet

Quiz: I hvilken figur er u heteroskedastisk?



Heteroskedasticitet

Quiz: I hvilken figur er u heteroskedastisk?



Left $u_i \sim \mathcal{N}(0, \sigma^2)$, for all i

Right $u_i \sim \mathcal{N}(0, \sigma^2(x_i))$ where $\sigma(x_i) = \sigma_0(1 + \gamma|x_i|)$

Homoskedasticitet

Hvad betyder homoskedasticitet?

- Det betyder at variansen af u beintinget på x er uafhængig af x .

Vi kan vise, at når SLR.5 er opfyldt, så gælder.

$$\begin{aligned} \text{Var}(u|x) &= E([u - E(u|x)]^2|x) \\ &= E(u^2|x) = \sigma^2 \end{aligned}$$

Der gælder også, at:

$$\text{Var}(u) = E(u^2) = E(E(u^2|x)) = \sigma^2$$

Altså den ubetingede varians af fejlleddet er også σ^2 .

Teorem 2.2: Variansen af OLS-estimatorene

Under antagelse af SLR.1-SLR.5 gælder det, at variansen af OLS-estimatorene er:

$$\text{Var}(\hat{\beta}_1 | X) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sigma^2}{SST_x}$$

$$\text{Var}(\hat{\beta}_0 | X) = \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

hvor $X = \{x_1, x_2, \dots, x_n\}$ er de x 'er, vi har i vores data.

Variansen af OLS-estimatoren: Bevis

Ingredienser til beviset:

$$\text{OLS-estimatoren: } \hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Antagelser:

- SLR.1: Model: $y_i = \beta_0 + \beta_1 x_i + u_i$
- SLR.2: Observationerne er uafhængige.
- SLR.3: $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0$
- SLR.4: $E(u_i | x_i) = 0$
- SLR.5: $\text{Var}(u_i | x_i) = \sigma^2$

Regneregler:

- $\text{Var}(a(x)y + b(x) | x) = a(x)^2 \text{Var}(y | x)$

$$\begin{aligned} \text{Var}(\hat{\beta}_1 | X) &= \text{Var} \left(\frac{\sum_{i=1}^n u_i(x_i - \bar{x})}{SST_x} \middle| X \right) \\ &= \frac{1}{SST_x^2} \text{Var} \left(\sum_{i=1}^n u_i(x_i - \bar{x}) \middle| X \right) \quad (\text{brug SLR.2}) \\ &= \frac{1}{SST_x^2} \sum_{i=1}^n \text{Var}(u_i | X)(x_i - \bar{x})^2 \\ &= \frac{1}{SST_x^2} \sum_{i=1}^n \sigma^2(x_i - \bar{x})^2 \quad (\text{brug SLR.5}) \\ &= \frac{\sigma^2}{SST_x^2} \sum_{i=1}^n (x_i - \bar{x})^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{aligned}$$

Variansen af OLS-estimatoren

Variansen af fejlleddet, σ^2 , er ukendt.

Vi kan estimere σ^2 med følgende estimator:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2} \quad (7)$$

Bemærk: Vi dividerer med $n-2$ (ikke n) for at korrigere for at to parametre, $\hat{\beta}_0$ og $\hat{\beta}_1$, er estimeret fra data.

Teorem 2.3: Middelrethed af OLS-variansestimatoren

Under antagelse af SLR.1-SLR.5 er estimatoren for variansen af fejlleddet middeleret:

$$E(\hat{\sigma}^2) = \sigma^2$$

Bevis: Se Wooldridge (Teorem 2.3)

Variansen af OLS-estimatoren

Vi kan nu udregne **standardfejlen** for OLS-estimatoren.

Under antagelserne SLR.1-SLR.5 er standardfejlen for OLS-estimatoren:

$$se(\hat{\beta}_1) = \sqrt{\text{Var}(\hat{\beta}_1 | X)} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (8)$$

Standardfejlen måler variationen i $\hat{\beta}_1$ forårsaget af stikprøvevariation.

Standardfejlen er central i hypotesetestning og konfidensintervaller.

Opsummering

- Givet SLR.1-SLR.4 er OLS-estimatoren middelfret.
- Givet SLR.1-SLR.5 er variansen af OLS-estimatoren:

$$\text{Var}(\hat{\beta}_1 | X) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

hvor variansen kan estimeres som:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

Dvs. vi ved nu:

$$\hat{\beta}_0 \sim ? \left(\beta_0, \frac{\sigma^2 \frac{1}{n} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \quad \text{og} \quad \hat{\beta}_1 \sim ? \left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)$$

Vi har endnu **ikke udledt fordelingen** af $\hat{\beta}_0$ og $\hat{\beta}_1$.



- Jupyter Notebook: `02_slr_examples.ipynb`
- Part 3: Timeløn og uddannelse (Varians og standard fejl)
- Part 4: Simulationsstudie (egenskaber ved OLS estimator)

Regressionsmodel uden konstantled

Nogle gange kan vi have en formodning om at regressionsmodel **ikke** bør indholde et konstantled.

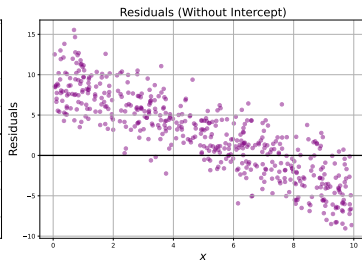
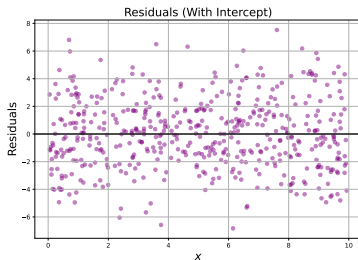
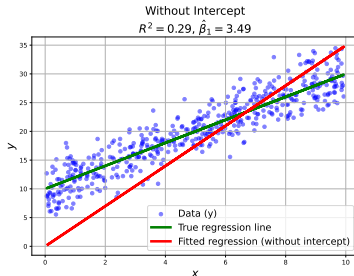
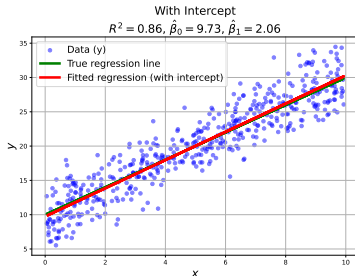
$$y = \beta_1 x_1 + u$$

Uden konstantled holder nogle af OLS egenskaberne dog ikke holder længere:

- Summen af residualerne $\sum_{i=1}^n \hat{u}_i$ er ikke nødvendigvis 0.
- R^2 kan blive negativ.
- Hvis der faktisk er et konstantled i den "sande model", vil OLS estimationen af en model uden konstantled være en **ikke-middelret estimator**.

Regressionsmodel med og uden konstantled

True parameters: $\beta_0 = 10, \beta_1 = 2$



Måleenheder og funktionelle former

Måleenheder og funktionelle former

Skal sammenhængen mellem y og x være lineær for at OLS virker?

Nej.

Det er muligt at lave transformationer af modellen, men parameterestimaterne ændrer sig.

Man behøver ikke at have en lineær relation mellem y og x , kun at modellen er lineær i parametrene, dvs. følgende er tilladt

$$g(y_i) = \beta_0 + \beta_1 f(x_i) + u_i.$$

Fortolkningen af parametrene bliver anderledes.

Vi vil undersøge både *lineære* og *ikke-lineære* transformationer af modellen.

Måleenheder og Funktionelle Former

- Skal sammenhængen mellem y og x være lineær for at anvende OLS? **Nej**.
- Transformationer ændrer parameterestimerterne, men OLS kan stadig bruges korrekt, så længe modellen er *lineær i parametrene*:

$$g(y_i) = \beta_0 + \beta_1 f(x_i) + u_i$$

- Fortolkningen af parametrene ændrer sig.
- Vi undersøger både *lineære* og *ikke-lineære* transformationer.

Lineære transformationer

Regresionsmodel:

$$\text{timeløn}_i = \beta_0 + \beta_1 \text{uddannelse}_i + u_i$$

Typiske lineære transformationer af modellen er ændringer i måleenhederne for den afhængige og/eller uafhængige variabel. Fx

- timelønnen målt i 2010-DKK priser:

$$\text{timeløn}_i^{2010} = \text{timeløn}_i^{1994} \times 1.37$$

- uddannelse målt i måneder i stedet for år:

$$\text{uddannelse}_i^{\text{måned}} = 12 \times \text{uddannelse}_i^{\text{år}}$$

- uddannelse målt relativt til 9. klasse:

$$\text{uddannelse}_i^{9.\text{klasse}} = \text{uddannelse}_i^{0.\text{klasse}} - 9$$

Lineære transformationer: Stata eksempel

```
gen wage2010    = wage * 1.37
gen educ_month  = educ * 12
gen educ_9      = educ - 9
```

```
reg wage educ
estimates store reg_base
reg wage2010 educ
estimates store reg_wage2010
reg wage educ_month
estimates store reg_educ_month
reg wage educ_9
estimates store reg_educ_9
```

```
estimates table reg_base reg_wage2010 reg_educ_month reg_educ_9, stats(N r2)
```

Lineære transformationer: Stata eksempel

```
estimates table reg_base reg_wage2010 reg_educ_month reg_educ_9, stats(N r2)
```

Variable	reg_base	reg_wage2010	reg_educ_month	reg_educ_9
educ	4.2669794	5.8457618		
educ_month			.35558162	
educ_9				4.2669794
_cons	90.336431	123.76091	90.336431	128.73925
N	1078	1078	1078	1078
r2	.0890296	.0890296	.0890296	.0890296

Lineære transformationer

Overordnet gælder, at hvis vi starter fra modellen

$$y = \beta_0 + \beta_1 x + u$$

Så medfører de transformerede variable \tilde{y} og \tilde{x} følgende

$$\tilde{y} = y * a \Rightarrow \tilde{\beta}_0 = a\hat{\beta}_0, \quad \tilde{\beta}_1 = a\hat{\beta}_1$$

$$\tilde{y} = y + a \Rightarrow \tilde{\beta}_0 = \hat{\beta}_0 + a, \quad \tilde{\beta}_1 = \hat{\beta}_1$$

$$\tilde{x} = x * a \Rightarrow \tilde{\beta}_0 = \hat{\beta}_0, \quad \tilde{\beta}_1 = 1/a\hat{\beta}_1$$

$$\tilde{x} = x + a \Rightarrow \tilde{\beta}_0 = \hat{\beta}_0 - a\hat{\beta}_1, \quad \tilde{\beta}_1 = \hat{\beta}_1$$

I kan tjekke ovenstående ved at bruge (\tilde{y}, \tilde{x}) i udledningen af OLS estimatoren.

Påvirker lineære transformationer prædiktioner og goodness of fit?

Lineære transformationer: Standardiserede variable

En særlig form for lineære transformation af en variabel kaldes **standardisering**

$$\tilde{x} = \frac{x_i - \bar{x}}{\hat{\sigma}_x}$$

Dvs. vi omdanner x til at have middelværdi 0 og standardfejl 1.

Hældningskoefficienten angiver effekten på y , når x stiger med en standard afvigelse (hvis MLR.1-4 er opfyldt).

Standardiserede variable bruges often når enheden på x er svært at fortolke. Fx testscore eller IQ-målinger.

Mere om det i Wooldridge kapitel 6.1.

Quiz Betragt følgende estimations model

$$løn = \hat{\beta}_0 + \hat{\beta}_1 uddannelse,$$

hvor vi måler uddannelse i år.

Antag at vi i stedet måler uddannelse i måneder. Hvilket af følgende udsagn er sande?

1. Parameterestimerne er uændret.
2. SSE er uændret.
3. R^2 er uændret.
4. Standardfejlen på β_1 er uændret.

Ikke-lineære transformationer (funktionel form)

I mange studier er vi interesserede i den procentvise effekt på y af at ændre x . Fx det procentvise afkast af et års ekstra uddannelse.

En model med et konstant procentvis afkast er der i udgangspunktet en ikke-lineær sammenhæng mellem x og y .

Vi kan dog stadig estimere det med OLS ved at skrive modellen som

$$\log(\text{timeløn}) = \beta_0 + \beta_1 \text{uddannelse}_i + u$$

Fortolkningen af β_1 :

- Den relative ændring af timelønnen ved at tage et års ekstra uddannelse (ceteris paribus)
- $100\beta_1$ er ca. det procentvise afkast af et års ekstra uddannelse

Ligeledes kan vi også bruge $\log(\text{uddannelse})$ i stedet for uddannelse .

Ikke-lineære transformationer: Stata eksempel

```
gen lwage = log(wage)
gen leduc = log(educ)

reg wage educ
estimates store reg_base
reg lwage educ
estimates store reg_lwage
reg wage leduc
estimates store reg_leduc
reg lwage leduc
estimates store reg_loglog

estimates table reg_base reg_lwage reg_leduc reg_loglog, stats(N r2)
```

Ikke-lineære transformationer: Stata eksempel

```
estimates table reg_base reg_lwage reg_leduc reg_loglog, stats(N r2)
```

Variable	reg_base	reg_lwage	reg_leduc	reg_loglog
educ	4.2669794	.02821304		
leduc			67.098009	.45628662
_cons	90.336431	4.5603896	-24.529995	3.7694963
N	1078	1078	1046	1046
r2	.0890296	.08450376	.11585009	.1159361

Ikke-lineære transformationer (funktionel form)

TABLE 2.3 Summary of Functional Forms Involving Logarithms

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\% \Delta x$
Log-level	$\log(y)$	x	$\% \Delta y = (100\beta_1) \Delta x$
Log-log	$\log(y)$	$\log(x)$	$\% \Delta y = \beta_1 \% \Delta x$

Quiz

Ræk hånden op, hvis du mener følgende modeller er lineære i parametrene:

1. $y_i = \beta_0 + \beta_1\sqrt{x_i} + u_i$

2. $\exp(y_i) = \beta_0 + \beta_1\sqrt{x_i + 3} + u_i$

3. $Y_i = AL_i^{\beta_1} u_i$

4. $\log y_i = \beta_0 x_i^{\beta_1} + u_i$

Er det muligt at omskrive modellerne så de bliver lineære?

Data visualisering

Indtil videre har vi været i en verden, hvor vi har antaget at modellen er lineær:

$$E(y|x) = \beta_0 + \beta_1 x$$

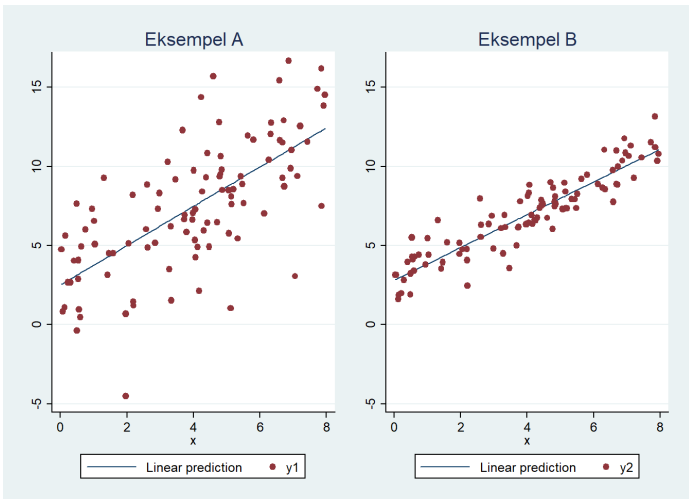
Vi skal senere se formelle test af om den funktionelle form er rigtig.

En mere simpel måde at validere den funktionelle form på er ved visuel inspektion af data.

Data visualisering: Eksempel

Quiz

Ræk hånden op, hvis du mener den funktionelle form er rigtig:



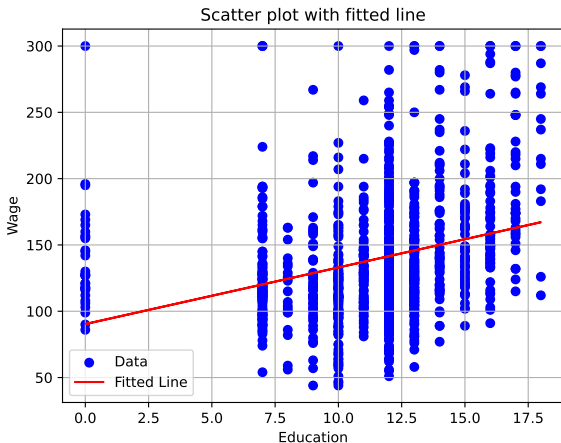
Data visualisering: Stata eksempel

```
reg wage educ  
predict wage_hat, xb
```

```
twoway (scatter wage educ) (line wage_hat educ, sort), ///  
graphregion(color(white)) ylabel(0(50)300)
```

Data visualisering: Stata eksempel

Hvad med her?



Hvad hvis vi havde millioner af observationer?

Vi kan danne et simpelt “ikke-parametrisk” estimator for $E(y|x)$, som

$$E(y|x) = \mu(x)$$

hvor $\mu(x)$ er gennemsnittet af y for personer med $x_i \in [x - \epsilon; x + \epsilon]$.

Vi kan plotte μ_x sammen med den fittede linje OLS uanset hvor mange observationer vi har.

Det kaldes et **bin scatter plot**.

- Bin = et interval af x .
- Dvs. vi beregner gennemsnittet af y indenfor et interval af x .
- I vores eksempel er det mest naturlige at bruge år.

Data visualisering: Stata eksempel

```
reg wage educ
predict wage_hat, xb

preserve
bin = educ
sort bin
collapse wage wage_hat educ, by(educ)

twoway (scatter wage educ) (line wage_hat educ, sort), ///
      graphregion(color(white)) ylabel(0(50)300)

restore
```

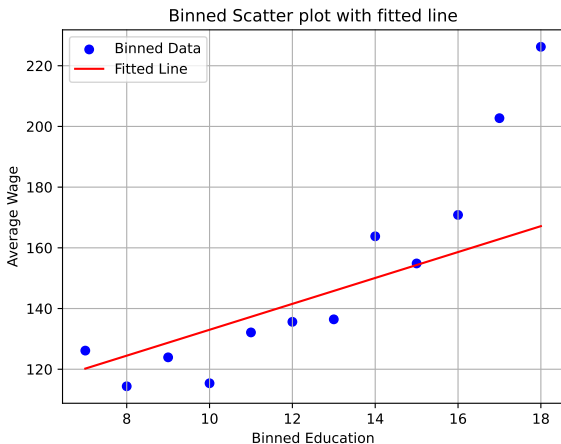
Hvis x er en kontinuert variabel, kan vi danne “bins” ved

```
gen bin = round(x/0.1)*0.1 /*Afrunding til nærmeste 0.1*/
```

Data visualisering: Stata eksempel

Hvad med nu?

Ræk hånden op, hvis du mener den funktionelle form er rigtig:



Opsummering

Opsummering: OLS er en (conditional) mean estimator



Opsummering

OLS model: $y = \beta_0 + \beta_1 x + u$

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

Husk der er forskel på:

- **Populationsparametre:** β_0 og β_1 (de sande værdier i populationen)
- **Estimer:** $\hat{\beta}_0$ og $\hat{\beta}_1$ (beregnet ud fra data ved hjælp af OLS-estimatoren)
- **OLS-estimatoren:** Metoden vi bruger til at beregne estimerne

Og forskel på

- Statistisk antagelse: $E(u|x) = E(u) = 0$
- Mekaniske egenskaber for residualerne: $\sum_{i=1}^n \hat{u}_i = \sum_{i=1}^n \hat{u}_i x_i = 0$.