

Øvelsestime 2

Simpel lineær regression

Agenda

- 1) Introduktion til PS2
- 2) Diskussion: Gruppearbejde og opsamling
- 3) Stataøvelser: Gruppearbejde og opsamling
- 4) Introduktion til teoretisk opgave (lektie)

Tobias Gabel Christiansen
tgc@econ.ku.dk
22-02-2021

KØBENHAVNS UNIVERSITET



1) Planen for semesteret

Dato	Opgave
Uge 7 - 15/02	PS1
Uge 8 - 22/02	PS2
Uge 9 - 01/03	PS3
Uge 10 - 08/03	OO1
Uge 11 - 15/03	PS4
Uge 12 - 22/03	PS5 + gennemgang af OO1
Uge 13 - 29/03	PS6
Uge 14 - 05/04	Påskeferie
Uge 15 - 12/04	OO2
Uge 15 - 19/04	PS7
Uge 16 - 26/04	PS8 + gennemgang af OO2
Uge 17 - 03/05	OO3
Uge 18 - 10/05	PS9
Uge 19 - 17/05	PS10 + gennemgang af OO3
	Spørgetime

← Her er vi :-)

1) Introduktion til problem set 2

- **PS1:** Indledende analyse med fokus på vælge og beskrive variable der kan bruges til at estimere en simpel Engelkurve for mad, tøj og alkohol.
- **PS2:** Simpel lineær regression til at estimere Engelkurver.

1) Introduktion til problem set 2

Version 1 (level-level):

$$xfath_i = \beta_0 + \beta_1 xtot_i + u_i$$

Fobrug og indkomst målt i (canadiske) dollars.

- β_1 : Angiver den forventede ændring i forbrug på mad (målt i \$) når samlet forbrug stiger med 1\$.
- i) Antager konstant sammenhæng mellem absolutte ændringer i $xtot_i$ og $xfath_i$
- β_0 : Angiver regressionslinjens skæring med y-aksen. Med andre ord er det værdien af forbrug ($xfath_i$) når samlet forbrug er nul ($xtot_i = 0$)
 - i) Eneste meningsfulde værdi for konstantled er derfor 0
 - ii) Selv hvis sande værdi for $\beta_0 = 0$ kan vi sagtens få et estimat $\widehat{\beta}_0 \neq 0$, da vi kun arbejder med et sample fra populationen. I vores tilfælde har vi fx ingen husholdninger med et samlet forbrug tæt på nul ($xtot_i \neq 0$) så vi har svært ved at sige noget om sammenhængen mellem samlet forbrug og forbrug på mad for disse værdier.

1) Introduktion til problem set 2

Kort note om konstantled:

- I nogle tilfælde giver det slet ikke mening at tolke β_0
- Hvorfor så inkludere det?

1) Dels sikrer konstantleddet at vi harmløst kan antage at $E(u) = 0$

Antag fx at vi har følgende model uden konstantled:

$$y_i = \beta_1 x_i + v_i$$

Hvis $E(v_i) \neq 0$ kan vi omskrive modellen (ved at lægge nul til):

$$\begin{aligned} y_i &= \beta_1 x_i + v_i + E(v_i) - E(v_i) \\ &= E(v_i) + \beta_1 x_i + (v_i - E(v_i)) \\ &= \beta_0 + \beta_1 x_i + u_i \end{aligned}$$

Hvor $\beta_0 = E(v_i)$ og $u_i = v_i - E(v_i)$. I modellen med konstantled gælder det derfor at $E[u_i] = 0$. Intuitivt så fanger konstantleddet alle de ting der ikke varierer. Da middelværdien for fejlledet er konstant fanges denne altså også.

Konstantleddet sikrer desuden (per konstruktion) at residualerne summer til 0 (se slide 23)

1) Introduktion til problem set 2

2) Generelt undgår vi desuden én potentiel kilde til bias i estimatoren for hældningsparameteren ($\widehat{\beta}_1$). $\widehat{\beta}_1$ vil generelt være biased hvis vi udelader konstantleddet og $\beta_0 \neq 0$:

Antag at sande model er givet ved:

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

I stedet estimerer vi modellen uden konstantled:

$$y_i = \beta_1 x_i + v_i$$

Estimatoren for denne model er givet ved:

$$\widehat{\beta}_1 = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum x_i (\beta_1 x_i + v_i)}{\sum x_i^2} = \beta_1 + \frac{\sum x_i v_i}{\sum x_i^2} = \beta_1 + \frac{\sum x_i (\beta_0 + u_i)}{\sum x_i^2} = \beta_1 + \beta_0 \frac{\sum x_i}{\sum x_i^2} + \frac{\sum x_i u_i}{\sum x_i^2}$$

Under antagelse at $E(u_i | x_i) = 0$ finder vi at:

$$E[\widehat{\beta}_1 | x] = \beta_1 + \beta_0 \frac{\sum x_i}{\sum x_i^2}$$

Så $\widehat{\beta}_1$ er biased hvis $\beta_0 \neq 0$ ($\sum x_i$ er generelt ikke lig 0). Med andre ord så er $\widehat{\beta}_1$ biased hvis vi tvinger regressionslinjen igennem origo når den sande skæring med y-aksen ikke er origo.

3) Omvendt hvis $\beta_0 = 0$ og vi inkluderer et konstantled vil det ikke føre til bias i $\widehat{\beta}_1$ men blot føre til at estimatoren har lidt højere varians. Dvs. inklusionen af et konstantled er i værste fald ret harmløs.

Bemærk: Der findes tilfælde hvor det også er harmløst at udelade konstantleddet. Et eksempel er når man arbejder med standardiseret data hvor konstantleddet er lig 0 per design.

1) Introduktion til problem set 2

Version 2 (level-log):

$$\frac{xfath_i}{xtot_i} = \delta_0 + \delta_1 \log\left(\frac{xtot_i}{price_i}\right) + v_i$$

Forbrugsandele og den naturlige logaritme til (deflateret) samlet forbrug.

- δ_1 : Angiver den forventede ændring i forbrugsandelen (målt i procentpoint) når samlet (deflateret) forbrug stiger med 1% (approximation der er mest præcis for små ændringer)
 - i) Antager konstant sammenhæng mellem relative ændringer i $\frac{xtot_i}{price_i}$ og absolutte ændringer $\frac{xfath_i}{xtot_i}$.
 - ii) I absolutte termer betyder det at der skal en større og større ændring i $\frac{xtot_i}{price_i}$ til for at producere samme effekt på $\frac{xfath_i}{xtot_i}$.
- δ_0 : Angiver nu værdien for $\frac{xfath_i}{xtot_i}$ når $\frac{xtot_i}{price_i} = 1$
 - i) Giver ikke mening at tolke i dette tilfælde

Bemærk: $\frac{\delta_1}{100}$ angiver ændringen i $\frac{xfath_i}{xtot_i}$ når $\frac{xtot_i}{price_i}$ stiger med 1%. Da $\frac{xfath_i}{xtot_i}$ er en andel (værdi mellem 0 og 1), kan vi gange ændringen (fx 0,05) med 100 for at tolke det i pp. Det gælder altså at $100 \cdot \frac{\delta_1}{100} = \delta_1$ angiver den forventede ændring i pp. når vores afhængige variable er en andel.

Fortolkning af parameter i level-log model: $y = \delta_0 + \delta_1 \log(x) + u$

Under antagelse at $\Delta u = 0$ (alt andet lige antagelse) vil en ændring i $\log(x)$ fører til:

$$\Delta y = \delta_1 \Delta \log(x) \leftrightarrow$$

$$\Delta y = \delta_1 (\log(x_1) - \log(x_0)) \leftrightarrow$$

$$\Delta y = \delta_1 \log\left(\frac{x_1}{x_0}\right) \leftrightarrow$$

Læg 0 til

$$\Delta y = \delta_1 \log\left(1 + \frac{x_1}{x_0} - 1\right) \leftrightarrow$$

Sæt på fælles brøk

$$\Delta y = \delta_1 \log\left(1 + \frac{x_1 - x_0}{x_0}\right) \leftrightarrow$$

Brug approksimation: $\log(1 + z) \approx z$

$$\Delta y \approx \delta_1 \frac{(x_1 - x_0)}{x_0} \leftrightarrow$$

Gang med $\frac{100}{100}$

$$\Delta y \approx \frac{\delta_1}{100} \frac{(x_1 - x_0)}{x_0} \cdot 100 = \frac{\delta_1}{100} (\% \text{-vis ændring i } x)$$

Gå selv igennem tilfældet med en *log-level* model og se at $100 \cdot \delta_1$ angiver den forventede *relative* ændring (målt i pct.) i y når x øges med én enhed.

Gruppearbejde

To-do næste 15 minutter:

- Besvar spørgsmål a. og b.
 - a. Discuss model (2). What is the interpretation of δ_1 when $\delta_1 > 0$ and $\delta_1 < 0$?
(Hint: luxury versus necessity goods)
 - b. What is likely to be included in the error term in model (2)?
(Hint: What other variables may influence the dependent variable, $xfath/xtot$, besides the explanatory variable?)

OBS: Brug for hjælp? Skriv Jeres gruppenummer i Padlet så kommer jeg hurtigst muligt (link i chatten)

2) Gruppearbejde - opsamling

- a. Discuss model (2). What is the interpretation of δ_1 when $\delta_1 > 0$ and $\delta_1 < 0$?
(Hint: luxury versus necessity goods)

Ved at **deflatere** med individuelt forbrugerindeks tages der højde for

- regionale prisforskelle (ødemark i Canada vs. Toronto)
- prisforskelle der skyldes at data ikke er samlet på samme tidspunkt (sommeren vs. vinter)

Ved at tage **logaritmen** fanger vi at sammenhængen mellem stigninger i samlet (deflateret) forbrug og budgetandelen for en vare sandsynligvis ikke er lineær. Empirisk er level-log modellen den funktionelle form der giver det bedste resultat (fanger at effekten af indkomststigninger på forbrugsandele er aftagende)

δ_1 angiver den forventede ændring i forbrugsandelen (målt i procentpoint) når samlet (deflateret) forbrug stiger med 1%

2) Gruppearbejde - opsamling

- a. Discuss model (2). What is the interpretation of δ_1 when $\delta_1 > 0$ and $\delta_1 < 0$?
(Hint: luxury versus necessity goods)

Ved at bruge **budgetandele** ser vi lettere om der er tale om et luksusgode eller nødevendighedsgode

$\delta_1 > 0$: Forbrugsandelen vokser når samlet (deflateret) forbrug stiger → Luksusgode

$\delta_1 < 0$: Forbrugsandelen falder når samlet (deflateret) forbrug stiger → Nødevendighedsgode

2) Gruppearbejde - opsamling

b. What is likely to be included in the error term in model (2)?

(Hint: What other variables may influence the dependent variable, $xfath/xtot$, besides the explanatory variable?)

Fejlleddet indeholder alt information om ændringer i budgetandelen på en vare, som ikke kan forklares af en konstant og den procentvise ændring i det samlede reale forbrug.

Det kan være:

- Udeladte variable (fx alder, køn, uddannelse, regionale forskelle)
- Målefejl
- Sæsonudsving

To potentielle ulemper ved at udelade en relevant variable (dvs. en variabel der har en effekt på $\frac{xfath}{xtot}$):

- Hvis $\log(\frac{xtot}{price})$ ikke samvarierer med den udeladte variabel kan vi reducere variansen på vores estimator ($\widehat{\delta_1}$) ved at inkludere den udeladte variabel.
- Hvis $\log(\frac{xtot}{price})$ samvarierer med den udeladte variabel bryder vi med SLR.4 og estimatoren ($\widehat{\delta_1}$) er biased hvis vi ikke kontrollerer for den udeladte variabel.

3) Introduktion til Stata øvelser

- I dag skal vi bruge OLS til at estimere en simpel Engelkurve i Stata
- Kommando i Stata:

reg y x

Start altid med at lave en do-file og gem denne i en mappe sammen med datasættet (brug det datasæt i gemte sidst)

```
1  *PROBLEM SET 2
2  clear all
3  cls
4  cd "C:\Users\45414\OneDrive\Dokumenter\Økonomi mm\Holdunderviser\Økonometri I\Problem sets - Slides\PS 2"
5
6  *QUESTION 1: BEGIN STATA LOG FILE
7  cap log close
8  log using PS2, replace text
9
10 *READ IN DATA FROM LAST WEEK
11 use engel
```

4) Gruppearbejde: Stata øvelser

To-do indtil 12:15

- Lav Stata øvelserne (anden halvdel af PS2)
- Herefter samler vi op i fællesskab inden jeg introducerer Jer til dagens lektie.
- **OBS:** Brug for hjælp? Skriv Jeres gruppenummer i Padlet så kommer jeg hurtigst muligt (link i chatten)

5) Opsamling på Stata øvelser

1. Brugen af log-fil

En Stata log-fil er en tekstfil der gemmer alle brugte kommandoer og alle resultater

```
1
2 *PROBLEM SET 2
3 clear all
4 cls
5 cd "C:\Users\45414\OneDrive\Dokumenter\Økonomi mm\Holdunderviser\Økonometri I\Problem sets - Slides\PS 2"
6
7 *QUESTION 1: BEGIN STATA LOG FILE
8 cap log close
9 log using PS2, replace text
10
11 *READ IN DATA FROM LAST WEEK
12 use engel
```

- **Log using PS2:** Danner en tekstfil (log-fil) der hedder PS2 i mappen angivet i begyndelsen (cd)
- **Log close:** Lukker loggen (hvis den er åben for ændringer). Undgår fejlmeddelse hvis vi kører koden to gange
- **Cap:** Fortæller Stata at den skal ignorere linjen hvis den medfører en fejlmeddelse (fx vil Stata brokke sig hvis vi burger **log close** før vi har dannet log-filen, dvs. første gang koden køres).

5) Opsamling på Stata øvelser

2. Estimation af simpel Engelkurve: $xfath_i = \beta_0 + \beta_1 xtot_i + u_i$

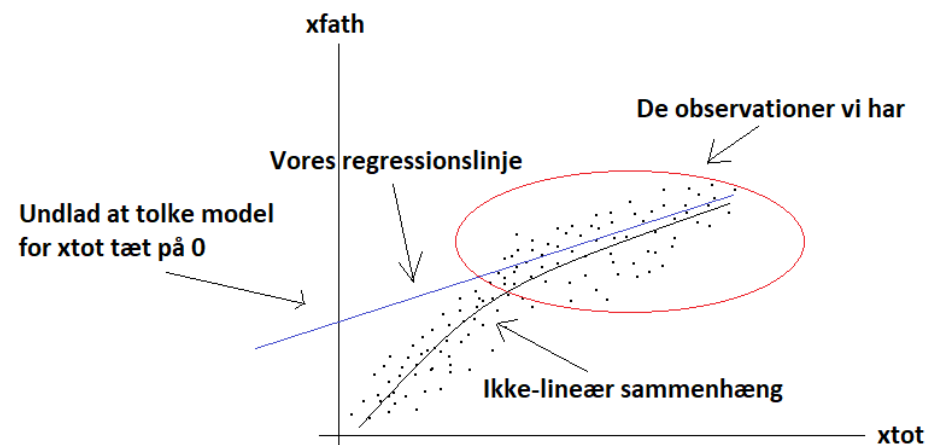
- i. What is the interpretation of the slope β_1 and intercept β_0 in model (1)?

β_1 : Når samlet forbrug stiger med \$1 så er den forventede stigning i udgifter til mad β_1

β_0 : Når samlet forbrug er 0 så er forbruget på mad β_0 (meningsløst hvis forskellig fra nul)

Men vi ser på forbrug som er forskellig fra 0 (ingen observationer med $xtot_i = 0$). Her kan sammenhæng godt være lineær, og i så fald er det vigtigt at inkludere konstantled for at undgå bias i vores hældningsparameter for den del af populationen vi fokuserer på. Lineariteten holder formentlig ikke omkring et forbrug på nul.

Stilistisk eksempel:



I det stilistiske eksempel ovenfor vil vi fx overestimere hældningen på den lineære sammenhæng mellem $xtot$ og $xfath$ for *den del af populationen vi fokuserer på* hvis vi udelader β_0 (prøv at tegn en linje der går igennem (0,0) og som fitter de observationer vi har og se hvad der sker med hældningen)

5) Opsamling på Stata øvelser

ii. What is the estimate of the slope? And the intercept?

Source	SS	df	MS	Number of obs	=	258
Model	14140148.2	1	14140148.2	F(1, 256)	=	13.11
Residual	276117266	256	1078583.07	Prob > F	=	0.0004
				R-squared	=	0.0487
				Adj R-squared	=	0.0450
Total	290257414	257	1129406.28	Root MSE	=	1038.5

xfath	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xtot	.0488408	.0134891	3.62	0.000	.0222771	.0754046
_cons	1443.072	195.0856	7.40	0.000	1058.895	1827.249

$\widehat{\beta}_1 = 0,049$: Vi forventer at forbrug på mad stiger med \$0,049 når samlet forbrug stiger med \$1

$\widehat{\beta}_0 = 1443,072$: Regressionslinjen skærer y-aksen over 0.

5) Opsamling på Stata øvelser

- iii. What is the total variation in the dependent variable, SST? The explained variation in the dependent variable, SSE? The variation in the residuals, SSR?

Source	SS	df	MS	Number of obs	=	258
Model	14140148.2	1	14140148.2	F(1, 256)	=	13.11
Residual	276117266	256	1078583.07	Prob > F	=	0.0004
Total	290257414	257	1129406.28	R-squared	=	0.0487
				Adj R-squared	=	0.0450
				Root MSE	=	1038.5

xfath	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xtot	.0488408	.0134891	3.62	0.000	.0222771	.0754046
_cons	1443.072	195.0856	7.40	0.000	1058.895	1827.249

Total sum of squares (SST): $\sum_i^n (y_i - \bar{y})^2 = 290257414$

Explained sum of squares (SSE): $\sum_i^n (\hat{y}_i - \bar{y})^2 = 14140148$

Sum of squared residuals (SSR): $\sum_i^n (y_i - \hat{y}_i)^2 = 276117266$

5) Opsamling på Stata øvelser

iv. Find the coefficient of determination, R^2 . How can it be calculated from the three measures from the previous question? How would you interpret the calculated R^2 ?

Source	SS	df	MS	Number of obs	=	258
Model	14140148.2	1	14140148.2	F(1, 256)	=	13.11
Residual	276117266	256	1078583.07	Prob > F	=	0.0004
				R-squared	=	0.0487
				Adj R-squared	=	0.0450
Total	290257414	257	1129406.28	Root MSE	=	1038.5

xfath	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xtot	.0488408	.0134891	3.62	0.000	.0222771	.0754046
_cons	1443.072	195.0856	7.40	0.000	1058.895	1827.249

$$R^2 = \frac{SSE}{SST} = \frac{14140148,2}{290257414} = 0,0487$$

Modellen forklarer ca. 4,9% af den observerede variation i forbrug på mad (xfath).

Husk: En lav R^2 skal ikke forstås som at modellen er dårlig. Vi er oftest interesseret i effekten fra en enkel af de uafhængige variable (x) på den afhængige variabel (y). Hvis vi troværdigt kan estimere (alt andet lige)-effekten fra x på y kan denne viden sagtens bruges til fx policy anbefalinger på trods af at vi kun forklarer en smule af variationen y.

5) Opsamling på Stata øvelser

v. What is the estimate of the variance of the error term, σ^2 ?

Source	SS	df	MS	Number of obs	=	258
Model	14140148.2	1	14140148.2	F(1, 256)	=	13.11
Residual	276117266	256	1078583.07	Prob > F	=	0.0004
				R-squared	=	0.0487
				Adj R-squared	=	0.0450
Total	290257414	257	1129406.28	Root MSE	=	1038.5

xfath	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
xtot	.0488408	.0134891	3.62	0.000	.0222771	.0754046
_cons	1443.072	195.0856	7.40	0.000	1058.895	1827.249

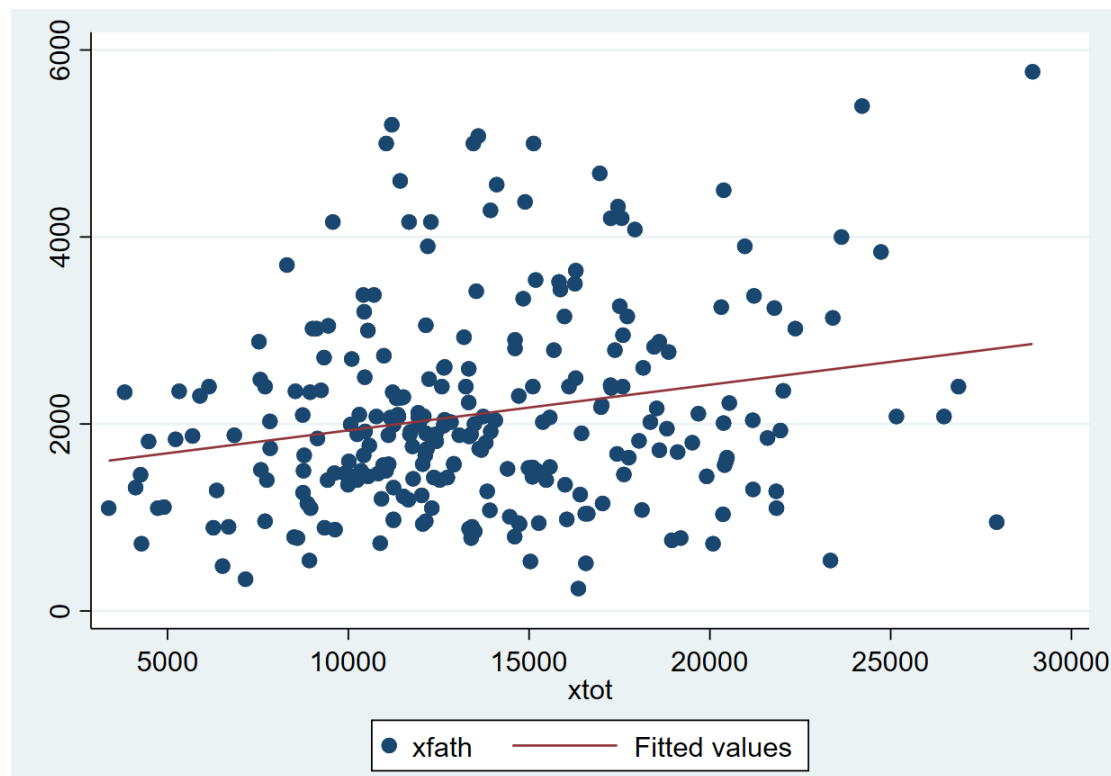
$$\hat{\sigma}^2 = \frac{SSR}{n - k} = \frac{276117266}{256} = 1078583,07$$

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{1078583,07} = 1038,5$$

5) Opsamling på Stata øvelser

vi. Illustrate the estimated Engel curve in a graph together with the actual data observations. This code should be useful:

```
twoway (scatter xfath xtot if dmale==1) (lfit xfath xtot if dmale==1)
```



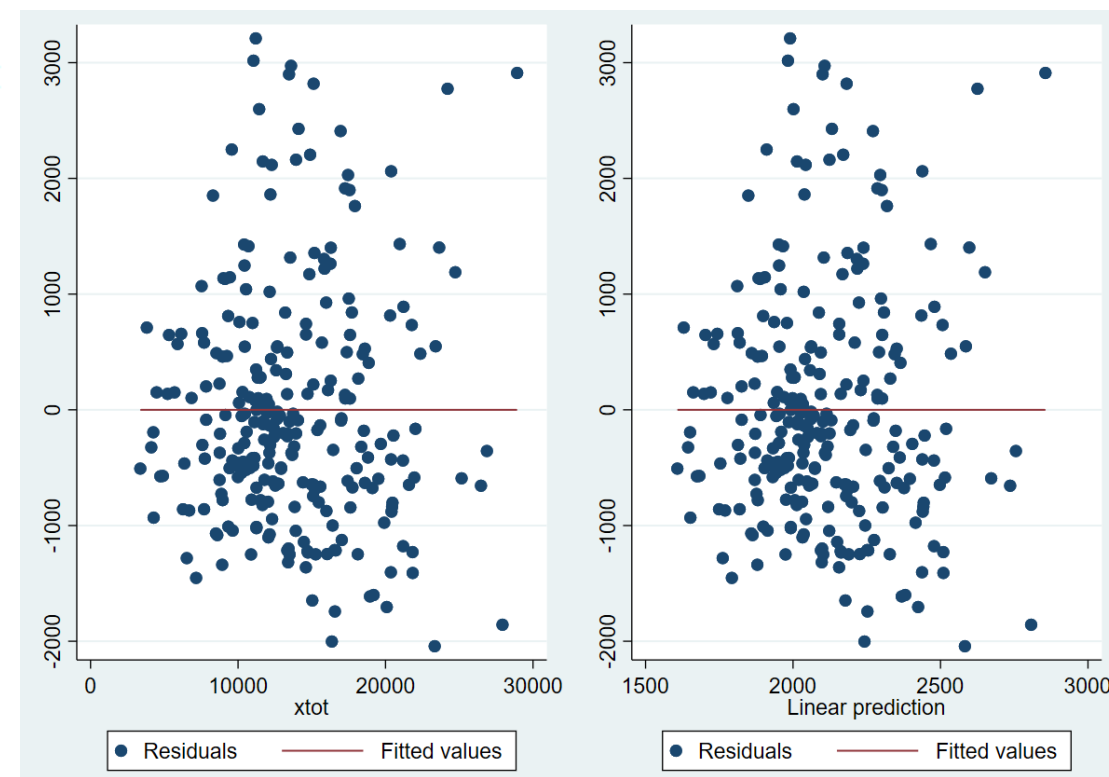
Twoway kombinerer grafer (i det her tilfælde "scatter" og "lfit")

5) Opsamling på Stata øvelser

vii. Perform scatterplots of the residuals from the regression against *xtot* and the predicted value of food consumption, separately.

```
regress xfath xtot if dmale==1
predict yhat if e(sample)==1, xb
predict uhat if e(sample)==1, residuals
*SCATTER PLOTS OF RESIDUALS, PREDICTED VALUES AND THE EXPLANATORY VARIABLE
twoway (scatter uhat xtot) (lfit uhat xtot), name(fig1, replace)
twoway (scatter uhat yhat) (lfit uhat yhat), name(fig2, replace)
graph combine fig1 fig2
```

- $yhat = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 xtot_i$
- $uhat = \hat{u}_i = y_i - \hat{y}_i$
- `e(sample)==1` angiver at vi kun benytter den del af data som modellen er baseret på (`dmale==1`)



5) Opsamling på Stata øvelser

- vii. Perform scatterplots of the residuals from the regression against *xtot* and the predicted value of food consumption, separately.

Per konstruktion gælder det altid at:

$$\sum_i^n \hat{u}_i = 0 \quad \text{og} \quad \sum_i^n \hat{u}_i x_{tot_i} = 0$$

Mekanisk egenskab for OLS: Stammer fra første-ordens betingelserne når vi løser for $\widehat{\beta}_0$ og $\widehat{\beta}_1$.

Da $\widehat{xfath}_i$ er en lineær kombination af x_{tot_i} (og dermed perfekt korreleret med denne) gælder desuden

$$\sum_i^n \hat{u}_i \widehat{xfath}_i = 0.$$

Tjekker vi efter finder vi at:

*CORRELATIONS AMONG RESIDUALS, PREDICTED VALUES AND THE EXPLANATORY VARIABLE
 *ARE THEY CONSISTENT WITH THE OLS ESTIMATOR'S MECHANICAL PROPERTIES?

`correlate` uhat xtot
`correlate` uhat yhat
`correlate` yhat xtot

	uhat	xtot		uhat	yhat		yhat	xtot
uhat	1.0000		uhat	1.0000		yhat	1.0000	
xtot	0.0000	1.0000	yhat	-0.0000	1.0000	xtot	1.0000	1.0000

5) Opsamling på Stata øvelser

- vii. Perform scatterplots of the residuals from the regression against $xtot$ and the predicted value of food consumption, separately.

Husk at disse mekaniske egenskaber kun gælder for residualerne. Med andre ord er det **ikke** ensbetydende med at:

$$E(u_i) = 0 \quad \text{og} \quad E(u_i x_{tot_i}) = 0$$

Vi observerer aldrig fejlleddet (u_i) så vi ved ikke om disse holder. Vi kan blot argumentere/antage at de gør.

5) Opsamling på Stata øvelser

3. Construct the variables which are needed to estimate model (2). $\frac{xfath_i}{xtot_i} = \delta_0 + \delta_1 \log\left(\frac{xtot_i}{price_i}\right) + v_i$

- Sidste gang (PS1) dannede vi:

```
15 *QUESTIONS 3+4: CONSTRUCT RELEVANT VARIABLES
16 generate xtot=xfath+xrest+xhhop+xwcloth+xmcloth+xtran+xrecr+xtob+xalc+xcare+xcaruse
17 generate wfath = xfath/xtot
18 generate wwcloth = xwcloth/xtot
19 generate wmcloth = xmcloth/xtot
20 generate walc = xalc/xtot
```

- De eneste nye variable er derfor:

```
*QUESTION 3: CONSTRUCT NEW VARIABLES
generate wcloth=(xwcloth+xmcloth)/xtot
generate lnrxtot=log(xtot/price)
```

- Bemærk: Vi ved fra sidst at mænd kun købte herretøj og kvinder kun købte dametøj. Vi kan derfor blot samle det til én variabel (wcloth) som gjort ovenfor . Giver samme resultat som hvis vi brugte wwcloth og wmcloth separat når vi betinger på køn i vores regressioner.

5) Opsamling på Stata øvelser

4. Estimate model (2) by OLS for the budget shares of food, clothing and alcohol for men and women, separately. Provide a table with the estimated slope parameters (for men and women in all three expenditure categories).

```
*QUESTION 4: ESTIMATE ENGEL CURVE MODELS IN EXPENDITURE SHARES
```

```
regress wfath lnrxtot if dmale==1  
outreg2 using results.xls, replace ctitle(FOOD - MALE)
```

```
regress wcloth lnrxtot if dmale==1  
outreg2 using results.xls, append ctitle(CLOTH - MALE)
```

```
regress walc lnrxtot if dmale==1  
outreg2 using results.xls, append ctitle(ALCOHOL - MALE)
```

```
regress wfath lnrxtot if dmale==0  
outreg2 using results.xls, append ctitle(FOOD - FEMALE)
```

```
regress wcloth lnrxtot if dmale==0  
outreg2 using results.xls, append ctitle(CLOTH - FEMALE)
```

```
regress walc lnrxtot if dmale==0  
outreg2 using results.xls, append ctitle(ALCOHOL - FEMALE)
```

- Installér outreg2 ved at skrive: **ssc install outreg2** i kommandovinduet
- Vi vælger selv navnet på filen ("navn.xls"), og filtypen (.xls, .tex, .text, .doc osv.)
- ctitle('TITEL') angiver titlen på kolonnen med den tilhørende regression.
- Bemærk vi skriver "... results.xls, *replace*..." ved første regression. Undgår at Stata brokker sig over at der allerede eksisterer en fil der hedder "results.xls" hvis vi kører koden igen.
- Ved efterfølgende regressioner skriver vi "... results.xls, *append*...". Fortæller Stata at disse regressioner skal tilføjes til filen "results.xls".
- Se "help outreg2" for flere funktioner.

5) Opsamling på Stata øvelser

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	FOOD - MALE	CLOTH - MALE	ALCOHOL - MALE	FOOD - FEMALE	CLOTH - FEMALE	ALCOHOL - FEMALE
lnrxtot	-0.117*** (0.0139)	0.0218** (0.00915)	0.0191 (0.0125)	-0.111*** (0.0163)	0.0744*** (0.0112)	0.00732 (0.00885)
Constant	0.633*** (0.0552)	-0.00662 (0.0363)	-1.31e-06 (0.0498)	0.594*** (0.0628)	-0.154*** (0.0433)	0.00833 (0.0342)
Observations	258	258	258	192	192	192
R-squared	0.217	0.022	0.009	0.196	0.188	0.004
Standard errors in parentheses						
*** p<0.01, ** p<0.05, * p<0.1						

Hældningskoefficienten (lnrxtot) er den centrale parameter. Hvis den er positiv er det et luksusgode og hvis den er negativ er det et nødvendighedsgode. Konstanten er som tidligere nævnt svært at give en økonomisk fortolkning.

Husk: Hvis samlet forbrug stiger med 1 pct. så er den forventede stigning forbrugsandelen approksimativt $\widehat{\beta}_1$ pp.

- Mad: $\widehat{\beta}_1 < 0$ for begge køn $\rightarrow$ Nødvendighedsgode
- Tøj: $\widehat{\beta}_1 > 0$ for begge køn $\rightarrow$ Luksusgode
- Alkohol: $\widehat{\beta}_1 > 0$ for begge køn $\rightarrow$ Luksusgode (men statistisk insignifikant)

Generelt **små effekter** (økonomisk/praktisk signifikans lav)

6) Introduktion til lektie

1. Write up the linear regression model with a constant term and an explanatory variable in matrix form for n observations.

Antagelser og definitioner:

- n individer
- afhængig variabel y
- forklarende variabel x
- fejllid ϵ
- parametervektor β

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$



$$\underset{n \times 1}{Y} = \underset{n \times 2}{X} \underset{2 \times 1}{\beta} + \underset{n \times 1}{\epsilon}$$

6) Introduktion til lektie

Modellen i matrixform er derfor givet ved:

$$\underset{n \times 1}{Y} = \underset{n \times 2}{X} \underset{2 \times 1}{\beta} + \underset{n \times 1}{\epsilon}$$

Løses for $\hat{\beta}$ ved brug af OLS (dvs. ved at minimere sum of squared residuals):

$$\min_{\hat{\beta}} \hat{\epsilon}'\hat{\epsilon} \quad , \quad \hat{\epsilon} = Y - X\hat{\beta}$$

Finder man at OLS-estimatoren er givet ved:

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

Alternativt til OLS kan vi finde $\hat{\beta}$ gennem Method of Moments [ved at løse den empiriske udgave af $E(X'u)$] Giver samme resultat. I dette tilfælde er $\hat{\beta}$ en (2 x 1) vektor. Det er jeres opgave at finde frem til udtrykkene for de to elementer i:

$$\hat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \end{bmatrix}$$

6) Introduktion til lektie

Hints til at løse for $\hat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{bmatrix}$:

1. Find $X'X$ (2x2 matrix) og $X'Y$ (2x1 matrix)
2. Find $(X'X)^{-1}$ ved at invertere $X'X$ med reglen givet i PS2
3. Omskriv brøken foran matrixen til $\frac{1}{n\sum(x_i - \bar{x})^2}$ og gang $\frac{1}{n}$ ind i matrixen
4. Indsæt udtrykkene for $(X'X)^{-1}$ og $X'Y$ i estimatoren $\hat{\beta} = (X'X)^{-1}X'Y$
5. Gang matricerne ud (ignorer brøken foran)
6. Læg $\bar{y}\bar{x}\sum x_i - \bar{y}\bar{x}\sum x_i = 0$ til første række i vektoren og omskriv
7. Gang brøken foran ind.

7) Gennemgang af teoretisk opgave

Vi skal løse $\hat{\beta} = (X'X)^{-1}X'Y$:

$$\hat{\beta} = \begin{bmatrix} \bar{y} - \bar{x} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \end{bmatrix}$$

1) Find relevante matricer:

$$X'X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i x_i \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i y_i x_i \end{bmatrix}$$

- Inverter $(X'X)$ vha. regneregler

$$\begin{aligned}(X'X)^{-1} &= \begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i x_i \end{bmatrix}^{-1} = \frac{1}{n \sum_i x_i x_i - \sum_i x_i \sum_i x_i} \begin{bmatrix} \sum_i x_i x_i & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \\&= \frac{1}{n \sum_i x_i x_i - n \sum_i x_i \bar{x}} \begin{bmatrix} \sum_i x_i x_i & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \quad (\text{omskriv sum til gennemsnit}) \\&= \frac{1}{n \sum_i x_i (x_i - \bar{x})} \begin{bmatrix} \sum_i x_i x_i & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \quad (\text{omskriv nævner i brøk}) \\&= \frac{1}{n \sum_i (x_i - \bar{x})^2} \begin{bmatrix} \sum_i x_i x_i & -\sum_i x_i \\ -\sum_i x_i & n \end{bmatrix} \quad (\text{omskriv nævner vha. (A.7) og (A.8)})\end{aligned}$$

- Fortsat fra forrige slide:

$$(X'X)^{-1} = \frac{1}{\sum_i (x_i - \bar{x})^2} \begin{bmatrix} \frac{\sum_i x_i x_i}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \quad (\text{divider igennem med } n)$$

- OLS kan derfor skrives som:

$$\hat{\beta} = \frac{1}{\sum_i (x_i - \bar{x})^2} \underbrace{\begin{bmatrix} \frac{\sum_i x_i x_i}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}}_{(*)} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

- Fokus på (*)

$$\begin{aligned} (*) &= \begin{bmatrix} \frac{\sum_i x_i x_i}{n} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{n} \sum_i y_i \sum_i x_i x_i - \bar{x} \sum_i x_i y_i \\ -\bar{x} \sum_i y_i + \sum_i x_i y_i \end{bmatrix} \quad (\text{gang matricerne sammen}) \\ &= \begin{bmatrix} \bar{y} \sum_i x_i x_i - \bar{x} \sum_i x_i y_i + (\bar{y}\bar{x} \sum_i x_i - \bar{y}\bar{x} \sum_i x_i) \\ \sum_i y_i (x_i - \bar{x}) \end{bmatrix} \quad (\text{trick: læg nul til}) \\ &= \begin{bmatrix} \bar{y} \sum_i x_i (x_i - \bar{x}) - \bar{x} \sum_i x_i (y_i - \bar{y}) \\ \sum_i y_i (x_i - \bar{x}) \end{bmatrix} \quad (\text{diverse omskrivninger}) \\ &= \begin{bmatrix} \bar{y} \sum_i (x_i - \bar{x})^2 - \bar{x} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_i (y_i - \bar{y})(x_i - \bar{x}) \end{bmatrix} \quad (\text{brug (A.7) og (A.8)}) \end{aligned}$$

- Sætter vi dette tilbage i udtrykket fås:

$$\hat{\beta} = \frac{1}{\sum_i (x_i - \bar{x})^2} \begin{bmatrix} \bar{y} \sum_i (x_i - \bar{x})^2 - \bar{x} \sum_i (x_i - \bar{x})(y_i - \bar{y}) \\ \sum_i (y_i - \bar{y})(x_i - \bar{x}) \end{bmatrix}$$

$$= \begin{bmatrix} \bar{y} - \bar{x} \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \end{bmatrix}$$

- Som var det ønskede.

Tips og huskeregler til opgave 3 i OO1

Opgave 3.1

- I PS2 gennemgik vi en lektie hvor vi løste for $\hat{\beta}_0$ og $\hat{\beta}_1$ ved brug af matrixregning. I den opgave var modellen givet ved:

$$y_i = \beta_0 + \beta_1 x_{1i} + \epsilon_i$$

- Dette kunne vi skrive på matrixform som:

$$\underset{n \times 1}{Y} = \underset{n \times 2}{X} \underset{2 \times 1}{\beta} + \underset{n \times 1}{\epsilon} \quad \longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Tips og huskeregler til opgave 3 i OO1

Opgave 3.1

- I OO1 skal vi i stedet løse for $\hat{\beta}_1$ og $\hat{\beta}_2$ hvor modellen er givet ved

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- Dette kunne vi skrive på matrixform som:

$$\underset{n \times 1}{Y} = \underset{n \times 2}{X} \underset{2 \times 1}{\beta} + \underset{n \times 1}{\epsilon} \quad \longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

Tips og huskeregler til opgave 3 i OO1

Opgave 3.1

- I OO1 skal vi i stedet løse for $\hat{\beta}_1$ og $\hat{\beta}_2$ hvor modellen er givet ved

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- Dette kan vi skrive på matrixform som:

$$\begin{matrix} Y \\ n \times 1 \end{matrix} = \begin{matrix} X \\ n \times 2 \end{matrix} \begin{matrix} \beta \\ 2 \times 1 \end{matrix} + \begin{matrix} \epsilon \\ n \times 1 \end{matrix} \quad \longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

- Brug disse matricer til at finde: $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X'X)^{-1}X'y$
- Det kan være en hjælp at dele beregningerne op I bidder som vi gjorde i PS2. Bemærk dog at vi benyttede forskellige tricks i udledningen i PS2 som i ikke behøver at bruge denne gang.