

Øvelsestime 3

Hypotesetest og multiple lineære regression

Agenda

- 1) Introduktion til PS3
- 2) Gruppearbejde: Diskussion
- 3) Opsamling på Diskussion
- 4) Gruppearbejde: Stataøvelser
- 5) Opsamling på Stataøvelser

Tobias Gabel Christiansen
tgc@econ.ku.dk
01-03-2021

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 :-)

Første aflevering

- Første aflevering er blevet gjort tilgængelig i dag og skal afleveres online d. 15/02
- Består af fire opgaver skrevet som én rapport (aflever rapport som .pdf og do-file som .txt)
- Tænk på hver opgave som del af en større **helhed**:
 - 1) Deskriptiv analyse -> Empirisk / Teoretisk analyse -> Konklusion
 - 2) Overvej hvad det overordnede formål med rapporten er og hav dette for øje når I løser de individuelle opgaver
 - 3) Brug information fra tidligere i opgaven hvis det er relevant (referer fx tilbage til den deskriptive analyse, eller overvej hvordan resultater fra nye modeller er forskellige fra forrige)
- Husk at **fortolke** Jeres resultater undervejs.
- Skriv **relevante** antagelser/pointer kort og præcist.

Disse råd gælder generelt (også til eksamenen)

Første aflevering

- Næste holdtime er sat af til at arbejde på afleveringen. Her vil jeg være til rådighed, og kan vejlede Jer hvis I går i stå.
- Arbejd og aflever sammen I jeres grupper. Det er god træning inden **eksamenen**.
- Alle afleveringer skal godkendes. Jeg forventer at I alle gør Jeres bedste og løser / forsøger at løse **alle** opgaver.
- Grupper på tværs af hold? Upload kun besvarelsen til én holdunderviser (skriv en mail til mig hvis I afleverer til en anden end mig).

2) Introduktion til Problem set 3

Vi forlader mikroøkonomien (Engelkurver) og kaster os over makroøkonomien i PS3. Her vil vi analysere forskellige former for konvergens:

Absolut konvergens:

- På langt sigt vil BNP pr. arbejder konvergere mod én og samme vækstrate for alle lande, så alle lande konvergere mod samme niveau for BNP pr arbejder.
- Lande med relativt lav BNP pr arbejder ved et initialt tidspunkt vil opleve relativt højere vækst.

Betinget konvergens:

- På langt sigt vil BNP pr. arbejder konvergere mod en lande-specifik vækstrate som er bestemt ud fra strukturelle egenskaber (fx investeringsrater for fysisk og human kapital).
- BNP pr. arbejder konvergerer mod samme niveau for lande der strukturelt set er ens.
- Lande med relativ lav BNP pr. arbejder ved et initialt tidspunkt oplever højere for givne strukturelle niveauer

2) Introduktion til Problem set 3

- Data: Penn World Tables (PWT) (samme som der bliver brugt i Makro I)
- Datasættet indeholder 97 observationer (lande) med tilhørende 11 variable for hvert land:

Name	Label
country	
iso3	
y03	Real GDP per worker relative to the USA in year 2003
y60	Real GDP per worker relative to the USA in 1960
gy	Average annual growth rate of real GDP per worker, 1960-2003. Lo...
sk	Average investment rate in physical capital (investment share of G...
sh	Average investment rate in human capital, 1960-1995.
u	Average number of schooling years in population above the age of...
isi	Index of social infrastructure
n	Average annual labour force growth rate, 1960-2003.
grade	Grading of data reliability from Penn World Table 6.1

3) Gruppearbejde: Diskussion

To-do næste 10 minutter:

Hvordan undersøger vi om *absolut* konvergens finder sted?

- Specificér en model vi kan bruge ud fra de variable vi har tilgængeligt i data.
- Opstil en hypotese (dvs. angiv H_0 og H_a) som kan bruges til at teste om der tale om absolut konvergens.

Hint: Modellen skal have følgende form,

$$\frac{\ln(y_{i,T}) - \ln(y_{i,0})}{T} = K + \beta \ln(y_{i,0})$$

4) Opsamling på gruppearbejde

Specificér en model vi kan bruge ud fra de variable vi har tilgængeligt i data:

4) Opsamling på gruppearbejde

Specificér en model vi kan bruge ud fra de variable vi har tilgængeligt i data:

- Den empiriske model til at teste for absolut konvergens kan opskrives som:

$$gy_i = \beta_0 + \beta_1 \log(y60_i) + u_i$$

gy_i = gennemsnitlig vækstrate i BNP pr. arbejder fra 1960 til 2003

$y60_i$ = BNP pr. arbejder (relativt til USA) i 1960.

β_1 : Når initial BNP pr. arbejder vokser med 1% så er den forventede ændring i den gns.

vækstrate i BNP pr. arbejder $100 \cdot \frac{\beta_1}{100} = \beta_1$ pp. (Bemærk: vi dividerer med 100 for at tolke det som en %-vis ændring i $y60$ og ganger med 100 for tolke det som en pp. ændring i gy)

- $\beta_1 < 0$: Absolut konvergens. Højere BNP-vækst for lande med lavere initialindkomst.
- $\beta_1 = 0$: Ingen sammenhæng mellem initialindkomst niveau og vækstrate.
- $\beta_1 > 0$: Divergens. Højere BNP-vækst for lande med højere initialindkomst

OBS: I dette problem set ser vi bort fra divergens ($\beta_1 > 0$)

4) Opsamling på gruppearbejde

Opstil en hypotese (dvs. angiv H_0 og H_a) som kan bruges til at teste om der tale om absolut konvergens:

4) Opsamling på gruppearbejde

Opstil en hypotese (dvs. angiv H_0 og H_a) som kan bruges til at teste om der tale om absolut konvergens.

Da vi ser bort fra tilfældet med divergens er den relevante hypotesetest én-sidet:

$$H_0: \beta_1 = 0 \qquad H_a: \beta_1 < 0$$

H_0 : Ingen konvergens

H_a : Absolut konvergens

- Hvis nulhypotesen kan forkastes er der empirisk belæg for absolut konvergens.
- Omvendt, hvis nulhypotesen ikke kan forkastes er der ikke empirisk belæg for absolut konvergens.
- **Husk:** Vi tester værdier for populationsparametrene (β) og ikke estimatet ($\hat{\beta}$)

5) Gruppearbejde: Stataøvelser

To-do indtil 12:25 minutter:

- Arbejd på øvelserne I Stata

Hints:

Opg. 1: Man kan fjerne variable ved at skrive: drop if grade == "D"

Opg. 4: Variablen er defineret som $y60_i = \frac{BNP_i^{1960}}{BNP_{USA}^{1960}}$ hvor $BNP_{USA}^{1960} = 31691$. Omskriv $y60_i$ ud fra dette. Hvad ændrer sig når I estimerer den nye model? Kan I ud fra modellen $gy = \beta_0 + \beta_1 \log(y60_i) + u_i$ se hvorfor det er tilfældet?

Opg. 6: Retningen af omitted variable bias af $\widehat{\beta}_1$ når x_2 udelades fra modellen er givet ved:

	$\text{cov}(x_1, x_2) > 0$	$\text{cov}(x_1, x_2) < 0$
$\beta_2 > 0$	Opad biased	Nedad biased
$\beta_2 < 0$	Nedad biased	Opad biased

OBS: Skriv I Padlet hvis I har brug for hjælp (link i chatten)

6) Gennemgang af Stata øvelser

1) Load og beskriv data

```
clear all
cls
cd "C:\Users\45414\OneDrive\Dokumenter\Økonomi mm\Holdunderviser\Økonometri I\Problem sets - Slides\PS 3"
use pwt
describe
summarize
```

- Describe: 97 observationer og 11 variable
- Summarize: Pletvist manglende værdier for sk, sh, u og isi
- BNP pr. arbejder i 2003 (y03) mere spredt end BNP pr. arbejder i 1960 (y60). Taler umiddelbart imod absolut konvergens (dvs. imod at lande konvergerer mod samme niveau)

summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
iso3	0				
y03	97	.3268375	.3311104	.0189353	1.735
y60	97	.306645	.2779749	.0263556	1.03
gy	97	.0162376	.0135159	-.0107736	.0557359
sk	97	.1570792	.0819589	.0242046	.4359967
sh	83	.0662433	.0361047	.0052366	.1464919
u	81	6.478321	2.831886	.839	12.049
isi	96	10.90923	102.0116	.19339	1000
n	97	.0219104	.0089357	.0007946	.0479241
grade	0				

6) Gennemgang af Stata øvelser

1) Load og beskriv data

```
summarize if grade == "D"
summarize if grade!="D"
drop if grade=="D"
```

- 11 observationer frasorteres når vi fravælger lande med `grade == D`
- Frasortede lande er kendetegnet ved lavere BNP pr. arbejder *og* lavere vækstrater.
- Vi frasorterer altså lande der umiddelbart går imod hypotesen om absolut konvergens (potentielt sample selection bias)
- Omvendt er datakvalitet er dårlig og potentielt upålideligt for disse lande.
- Betinget på `grade≠D`: Stadig missing value for `sh`, `u` og `isi`, men ikke længere for `sk` som vi burde senere.

```
summarize if grade == "D"
```

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
iso3	0				
y03	11	.1211692	.1827999	.0189353	.62778
y60	11	.1007046	.1029403	.0303274	.3947546
gy	11	.0125377	.0182847	-.0052427	.0557359
sk	11	.1086792	.0534549	.026438	.1741021
sh	6	.0328445	.0458306	.007126	.1242247
u	7	3.258	2.794059	.839	8.764
isi	11	.3514455	.1773043	.22283	.76689
n	11	.0238392	.0051754	.0157218	.0303094
grade	0				

```
summarize if grade!="D"
```

Variable	Obs	Mean	Std. Dev.	Min	Max
country	0				
iso3	0				
y03	86	.3531439	.3372188	.0211414	1.735
y60	86	.3329862	.2825464	.0263556	1.03
gy	86	.0167108	.0128449	-.0107736	.0529002
sk	86	.1632699	.0831171	.0242046	.4359967
sh	77	.0688458	.0342542	.0052366	.1464919
u	74	6.782946	2.656771	.876	12.049
isi	85	12.27553	108.4095	.19339	1000
n	86	.0216636	.0092997	.0007946	.0479241
grade	0				

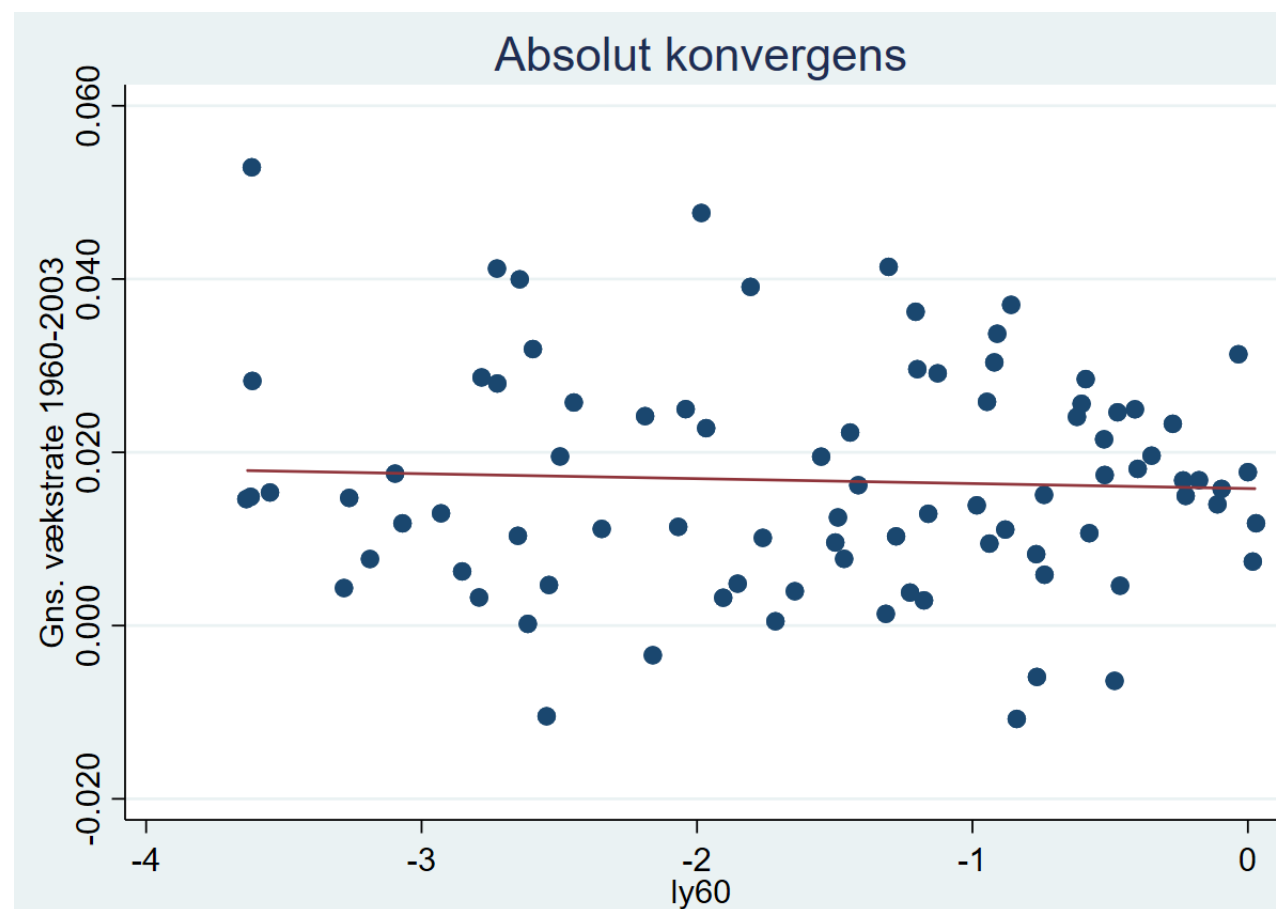
6) Gennemgang af Stata øvelser

2) Plot data

*QUESTION 2: CONSTRUCT LOG INCOME AND MAKE PLOT

```
generate ly60=log(y60)
twoway (scatter gy ly60) (lfit gy ly60), legend(off) title("Absolut konvergens") ytitle("Gns. vækstrate 1960-2003") name(fig1, replace)
graph export PS3.png, replace
```

- Korrelation mellem initial BNP pr. arbejder og gns. vækstrate umiddelbart 0 eller svagt negativ.
- Ingen tydelige tegn på absolut konvergens indtil videre.



6) Gennemgang af Stata øvelser

3) Regression: Absolut konvergens

- Når initial BNP pr. arbejder stiger med 1% er det forventede fald i den gns. Vækstraten 0,0006 pp.
- Effekten af initial BNP pr. arbejder er hverken *statistisk* eller *økonomisk* signifikant.
- Stata rapporterer automatisk to-sidet tests hvor $H_0: \beta_1 = 0$ vs. $H_a: \beta \neq 0$ for alle parameter.
- Modellen forklarer 0,22% af variationen i gns. vækstrater ($R^2 = 0,0022$)

```
*QUESTION 3: ESTIMATE GROWTH MODEL
regress gy ly60
```

Source	SS	df	MS	Number of obs	=	86
Model	.000031102	1	.000031102	F(1, 84)	=	0.19
Residual	.013993109	84	.000166585	Prob > F	=	0.6668
Total	.014024212	85	.000164991	R-squared	=	0.0022
				Adj R-squared	=	-0.0097
				Root MSE	=	.01291

gy	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ly60	-.0005696	.0013183	-0.43	0.667	-.0031911	.0020519
_cons	.0158234	.0024808	6.38	0.000	.0108901	.0207568

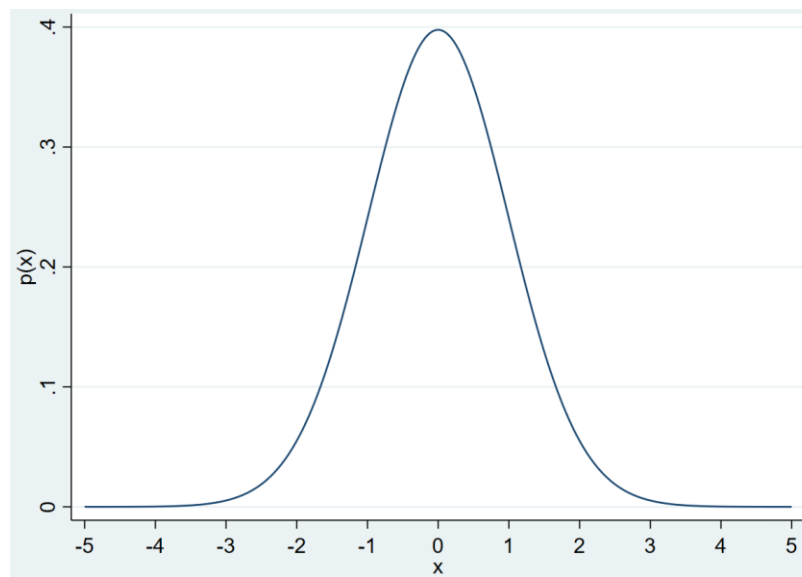
- Den én-sidede hypotese fra tidligere er givet ved:

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 < 0$$

- Vi finder t-værdien (denne er den samme for én-sidet og to-sidet test):

$$t = \frac{(\widehat{\beta}_1 - 0)}{se(\widehat{\beta}_1)} = \frac{-0.0006}{0.0013} = -0,43$$

- **Under antagelse at MLR.1-6 er opfyldt** ved vi at t-statistikken følger en t-fordeling med 84 (n-k-1) frihedsgrader **under H_0** . Vi kan derfor vurdere om H_0 er sandsynlig ved at tage stilling til, om det er sandsynligt at den værdi for t-statistikken som vi har beregnet (ud fra vores specifikke sample), er trukket fra denne fra denne fordeling:

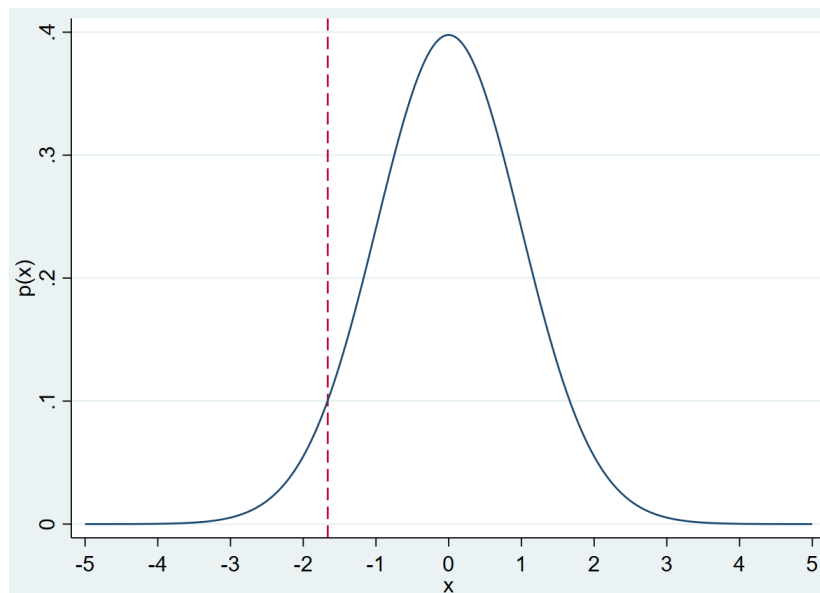


- Hvis det er sandsynligt at den værdi vi har beregnet med vores sample kan være trukket fra denne fordeling, taler det til *forde*l for at t rent faktisk følger ovenstående fordeling, og dermed til *forde*l for H_0 (husk t -statistikken følger kun denne fordeling under H_0)
- Hvis det er derimod er ssandsynligt at den værdi vi har beregnet er trukket fra denne fordeling, taler det *imod* at t følger ovenstående fordeling, og dermed *imod* H_0 (til *forde*l for H_a)

Bemærk: Ovenstående ræsonnement bygger på antagelsen at MLR.1-6 er opfyldt!

Hypotesetest

- Hvordan vurderer vi om $-0,43$ er sandsynlig eller ej?
- En måde er ved at vælge et **signifikansniveau** (α). Signifikansniveauet definerer en **kritiskværdi** hvor sandsynligheden for at trække en værdi der er mere ekstrem end denne er $\alpha\%$. Vælger vi fx et 5% signifikansniveau ($\alpha = 0,05$) for vores én-sidede test er den kritiske værdi $-1,66$ ["*display invttail(84,0.95)*"]



- Under H_0 vil $(1 - \alpha)\% = 95\%$ af random samples fra populationen resulterer i en t -værdi der er mindre ekstrem end $-1,66$ (dvs. mere positiv i dette tilfælde).
- Hvis $t < -1,66$: Vi vurderer at det er usandsynligt at t følger ovenstående fordeling og forkaster $H_0: \beta_1 = 0$
- Hvis $t > -1,66$: Vi vurderer at det er sandsynligt at t følger fordelingen og kan ikke forkaste $H_0: \beta_1 = 0$
- **Bemærk:** Signifikansniveau angiver derfor hvor ofte vi er villige til at forkaste H_0 ved en fejl ($\alpha\%$ af random samples fra populationen vil rent faktisk resultere i t værdier mere ekstreme den kritiske værdi *under* H_0)

Hypotesetest

- I vores tilfælde finder vi at:

$$t = -0,44 > -1,66$$

- Vi kan derfor ikke forkaste at $\beta_1 = 0$ på et 5% signifikansniveau. Vi kan ikke forkaste at det negative fortegn på vores estimat $\widehat{\beta}_1$ skyldes sampling error og vi finder derfor ikke empirisk belæg for absolut konvergens.
- Et **alternativ** til på forhånd at definere et signifikansniveau og benytte den tilhørende kritiske værdi til at vurdere hvorvidt H_0 er sandsynlig eller usandsynlig er at beregne t -statistikens p -værdi.
- P -værdien kan fortolkes på to måder:
 - Det laveste signifikansniveau hvor vi kan forkaste H_0
 - Sandsynligheden for at observere en t -værdi lige så, eller mere, ekstrem som den vi har observeret (under H_0).
- Vi forkaster H_0 hvis p -værdien er lavere end vores signifikansniveau (fx 5%)
 - $p > 0.05$: Der er mere end 5% sandsynlighed for at observere en t -værdi så ekstrem som den vi har observeret under H_0
 - $p < 0.05$: Sandsynligheden for at vi har observeret t -værdien under H_0 er mindre end 5% så vi forkaster H_0 .

$$p_{ensidet} = P(t_{86-2} \leq -0,43) = 0,33$$

- Vi kan derfor (stadig) ikke forkaste H_0 på et 5% signifikansniveau. Vi skulle vælge et signifikansniveau højere eller lig 33% for at forkaste H_0 .

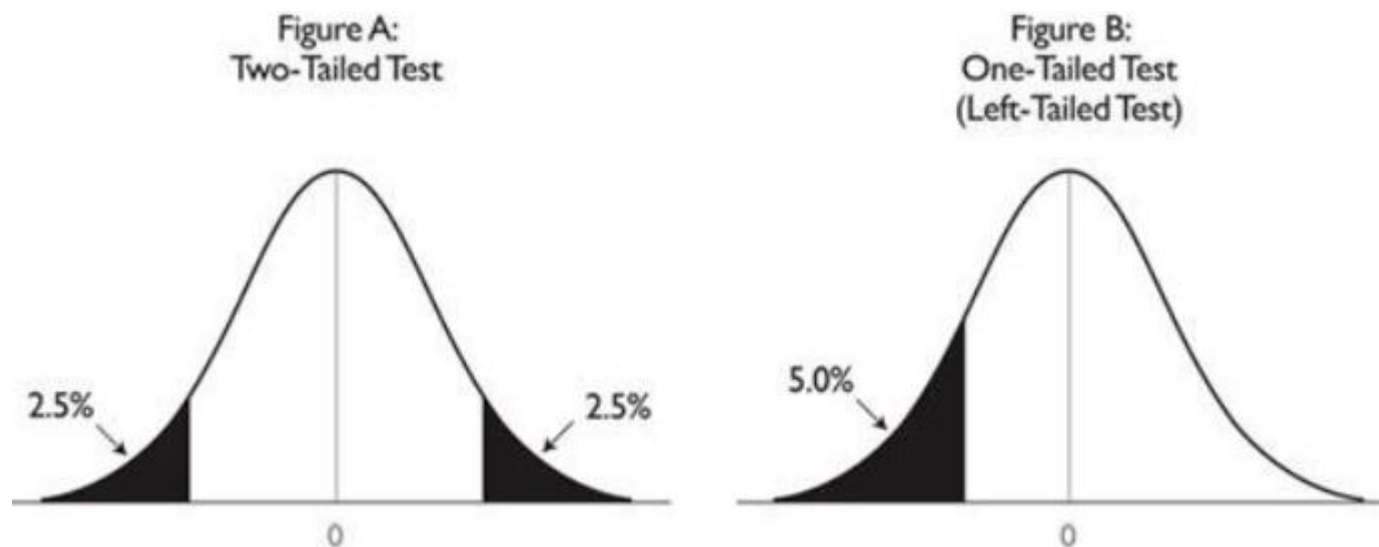
P-værdi: én-sidet vs. to-sidet

Pga. symmetri omkring 0

$$p_{tosidet} = P(|t_{86-2}| \geq |-0,43|) = P(t_{86-2} \leq -0,43) + P(t_{86-2} \geq 0,43) = 2P(t_{86-2} \leq -0,43) = 0,66 = 2p_{ensidet}$$

Genvej: P-værdi for én-sidet test er halv så stor som p-værdi for to-sidet test.

Stata: Rapporterer automatisk p-værdier for to-sidet test om $\beta = 0$



Ekstra note om én-sidet hypotesetest:

- Ved en én-sidet hypotesetest bør man opskrive nul- og alternativhypotese som:

$$H_0: \beta_1 \geq 0 \quad H_a: \beta_1 < 0$$

På den måde inkluderer H_0 og H_a alle tænkelige værdier for β_1 . **Men** hvis vi kan forkaste H_0 under $\beta_1 = 0$, så kan vi også forkaste H_0 for alle værdier hvor $\beta_1 > 0$. Med andre ord er $\beta_1 = 0$ den værdi som ligger tættest på H_a og dermed sværest at forkaste. I praksis skriver vi derfor ofte blot:

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 < 0$$

- Det er "**lettere**" at forkaste H_0 når vi bruger en én-sidet hypotesetest end ved en to-sidet såfremt fortegnet på estimatet er i samme retning som den vi tester i. Det skyldes at vi lægger al sandsynlighedsmassen i den ene hale af fordelingen. Det kommer på bekostning af at en ...
- ... én-sidet hypotesetest kun informerer os om effekten i **én retning** (i ovenstående tilfælde om β_1 er signifikant negativ eller ej). Hvis vi eksempelvis estimerede $\widehat{\beta}_1 = 1000$ med tilhørende $se(\widehat{\beta}_1) = 0.0001$, ville vi med den ovenstående én-sidede hypotesetest ikke kunne forkaste $H_0: \beta_1 = 0$! Med andre ord har vi ingen statistisk power i den *modsatte* retning end den vi tester i. Brug derfor kun én-sidet hypotesetest hvis I er sikre på retningen af effekten, eller hvis I ikke interesserer jer for resultater i modsat retning.
- Basér aldrig hvorvidt I bruger én-sidet test fremfor en to-sidet test på baggrund af den estimerede værdi $\widehat{\beta}_1$. Det invaliderer testen da det signifikansniveau I angiver ikke vil være retvisende (I vil være mere tilbøjelige til at forkaste H_0 ved en fejl end det fremgår). Hypotesen bør altid opstilles **før** I har set data.

6) Gennemgang af Stata øvelser

Huskeregler til hypotesetest (også til eksamen):

1. Husk altid eksplicit at skrive både H_0 og H_a (herunder om I benytter en to-sidet eller én-sidet test)
2. Husk at angive hvilket signifikansniveau I bruger (fx 5% / $\alpha = 0,05$)
3. Argumentér for, eller antag at, modellen opfylder de egenskaber der er krævet for at testen er valid (fx MLR.1-6)
4. Angiv derefter hvilken fordeling test-statikken følger under H_0 og find tilhørende kritiske værdier / p-værdier ud fra denne.
5. Husk at fortolke det I finder frem til. Hvilken betydning har det – i ord – at I fx forkaster H_0 ?

6) Gennemgang af Stata øvelser

4) Skaler variabel

Vi omskriver variabelen fra relativ til absolutte værdi i 1960:

$$y_{60} \cdot BNP_{USA}^{1960} = \frac{BNP_i^{1960}}{BNP_{USA}^{1960}} BNP_{USA}^{1960} = BNP_i^{1960}$$

Hvor $BNP_{USA}^{1960} = 31691$. Estimerer vi modellen med den nye variabel finder vi:

```
*QUESTION 4: SPECIFY INCOME IN ABSOLUTE TERMS AND RE-ESTIMATE MODEL
generate ly60abs=log(y60*31691)
regress gy ly60abs
```

gy	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ly60abs	-.0005696	.0013183	-0.43	0.667	-.0031911	.0020519
_cons	.0217268	.0116917	1.86	0.067	-.0015234	.044977

Det er kun konstantleddet der ændrer sig, mens hældningen er uændret.

6) Gennemgang af Stata øvelser

4) Skaler variabel

Med andre ord har det ingen betydning for estimatoren af β_1 , men slår i stedet ud som en ændring i konstantleddet. Det skyldes at vi tager logaritmen til indkomsten:

$$\begin{aligned} gy &= \beta_0 + \beta_1 \ln(y60 * 31691) + u \\ &= \beta_0 + \beta_1 \ln(31691) + \beta_1 \ln(y60) + u \\ &= \widetilde{\beta_0} + \beta_1 \ln(y60) + u \end{aligned}$$

Intuition: Hvorvidt $y60$ er skaleret eller ej har ingen betydning når vi ser på relative ændringer.

6) Gennemgang af Stata øvelser

5) Betinget konvergens

- Den udvidede model er givet ved:

$$gy_i = \beta_0 + \beta_1 \log(y60_i) + \beta_2 struc_i + u_i$$

Hvor $struc_i = \ln(sk_i) - \log(n_i + 0,075)$, hvor n angiver befolkningstilvæksten og 0,075 angiver afskrivningsraten på fysisk kapital.

- Den relevante (én-sidede) hypotese for betinget konvergens er:

$$H_0 : \beta_1 = 0$$

$$H_A : \beta_1 < 0$$

- Hvis vi kan forkaste H_0 er der empirisk belæg for betinget konvergens.

6) Gennemgang af Stata øvelser

5) Betinget konvergens

- Forklaringsgraden højere: $R^2 = 0.43$. Dvs. 43% af variationen i vækstraten (gy) forklares af *ly60* og *struc*.
- $\widehat{\beta}_1$: Når initialindkomsten stiger med 1% så er det forventede fald i vækstraten 0.006 pp. for *givne strukturelle karakteristika*.
- Tester vi på et 5% signifikansniveau ($\alpha = 0,05$) finder vi at *ly60* har en signifikant og negativ effekt på *gy* ($t = -4,66 < -1,66$)
- H_0 afvises til fordel for betinget konvergens. Der er empirisk belæg for at landene konvergerer mod det samme BNP pr. arbejder betinget på at landene strukturelt set er ens.
- En anden afskrivingsrate ville lede til andre resultater da denne indgår additivt (ikke multiplikativt). Ændringer slår altså ikke længere blot ud som ændringer i konstantleddet.

*QUESTION 5: TEST THE HYPOTHESIS OF CONDITIONAL CONVERGENCE

generate struc=log(sk)-log(n+0.075)

regress gy ly60 struc

Source	SS	df	MS	Number of obs	=	86
Model	.006007669	2	.003003834	F(2, 83)	=	31.10
Residual	.008016543	83	.000096585	Prob > F	=	0.0000
				R-squared	=	0.4284
				Adj R-squared	=	0.4146
Total	.014024212	85	.000164991	Root MSE	=	.00983

gy	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ly60	-.0055361	.0011858	-4.67	0.000	-.0078947	-.0031776
struc	.0156729	.0019924	7.87	0.000	.0117101	.0196357
_cons	.0021904	.0025636	0.85	0.395	-.0029084	.0072893

6) Gennemgang af Stata øvelser

6) Til forelæsningerne (forelæsning 5) så I at bias ved at udelade x_2 når den sande model er $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ er givet:

$$E[\widehat{\beta}_1|x] - \beta_1 = \beta_2 \frac{\text{cov}(x_2, x_1)}{\text{var}(x_1)}$$

I vores model svarer det til:

$$\text{bias} = \beta_2 \frac{\text{cov}(\log(y60), \text{struc})}{\text{var}(\log(y60))}$$

Så retningen af bias afhænger af fortegnet på β_2 og $\text{cov}(\log(y60), \text{struc})$. Ifølge Solowmodellen må det forventes at begge fortegn er positive.

	$\text{cov}(x_1, x_2) > 0$	$\text{cov}(x_1, x_2) < 0$
$\beta_2 > 0$	Opad biased	Nedad biased
$\beta_2 < 0$	Nedad biased	Opad biased

Bemærk: Bias er kun et problem når den udeladte variabel både er korreleret med variablene i modellen ($\text{cov}(x_1, x_2) \neq 0$) og har en effekt på den afhængige variabel ($\beta_2 \neq 0$)

6) Gennemgang af Stata øvelser

6) Vi kan omskrive udtrykket for bias (bruge definitionen for korrelation)

$$\begin{aligned} bias &= \beta_2 \frac{cov(\log(y60), struc)}{var(\log(y60))} = \beta_2 \frac{corr(\log(y60), struc) std(\log(y60)) std(struc)}{var(\log(y60))} \\ &= \beta_2 \frac{corr(\log(y60), struc) std(struc)}{std(\log(y60))} \end{aligned}$$

Vi kender aldrig sande bias. Men vi kan udregne bias i sample ved at bruge estimatet for $\widehat{\beta}_2$ fra den udvidede model i opgave 5) samt estimater for korrelationer ("*correlate*" i Stata) og standardfejl ("*summarize*" i Stata).

$$\begin{aligned} &= .0156729 \times 0.5324 \times \frac{.6320516}{1.061958} \\ &= 0.004966 \\ &= \widehat{\beta}_1^{SLR} - \widehat{\beta}_1^{MLR} \end{aligned}$$

Sample bias er altså positivt og lig forskellen mellem parametrene i SLR og MLR.

OBS: Ikke nødvendigvis sande bias, som vi aldrig observerer (fx kan der være andre udeladte variable).

Tak for i dag

- Næste holdundervisning er sat af til at arbejde på første obligatoriske opgave.
- Jeg vil kunne **vejlede** Jer hvis I går i stå.
- Brug tid på opgaven – det er en god øvelse til **eksamen**.