Økonometri I

Indholdsfortegnelse

[Matrice regneregler 6](#_Toc64977372)

[Den simple lineære regressionsmodel 6](#_Toc64977373)

[Ordbog 6](#_Toc64977374)

[Antagelser for SLR-modellen 6](#_Toc64977375)

[Method of Moments – MM (Momentestimation) 6](#_Toc64977376)

[Udledning af OLS-estimatoren 7](#_Toc64977377)

[Antagelser for den simple lineære regressionsmodel (SLR.1-4): 7](#_Toc64977378)

[Variansen af OLS-estimatoren 7](#_Toc64977379)

[Antagelse om varians på fejlledet (SLR.5) 7](#_Toc64977380)

[Theorem 2.2 9](#_Toc64977381)

[Standardfejlen for OLS-estimatoren 9](#_Toc64977382)

[Definition af MLR-modellen 9](#_Toc64977383)

[Teoretiske momenter: 10](#_Toc64977384)

[OLS-estimatoren for MLR-modellen 10](#_Toc64977385)

[Fuld rang: 10](#_Toc64977386)

[Fortolkning af 10](#_Toc64977387)

[Sammenhæng mellem MLR og SLR (Frish-Waugh Teoremet) 10](#_Toc64977388)

[OLS residualer 11](#_Toc64977389)

[Forklaringsgrad: 11](#_Toc64977390)

[MLR-antagelserne – Gauss-Markov antagelser 12](#_Toc64977391)

[Teorem 3.1: Middelrethed af OLS-estimatoren 12](#_Toc64977392)

[Overspecificeret regressionsmodel 13](#_Toc64977393)

[Underspecificeret regressionsmodel (udeladt variabel bias k=2) 13](#_Toc64977394)

[Udeladt variabel bias for OLS-estimatoren med k>2 14](#_Toc64977395)

[Varians af OLS-estimatoren 14](#_Toc64977396)

[Misspecificerede modeller 15](#_Toc64977397)

[Estimator af variansen af fejlleddet: 15](#_Toc64977398)

[Teorem 3.3 15](#_Toc64977399)

[Gauss-Markov teoremet (OLS er blot et vægtet gennemsnit) 15](#_Toc64977400)

[OLS-estimatorens fordeling 16](#_Toc64977401)

[Teorem 4.1 16](#_Toc64977402)

[Teorem 4.2: 16](#_Toc64977403)

[Hypotesetest: 17](#_Toc64977404)

[Hypotesetest ved simple hypoteser: 17](#_Toc64977405)

[Kogebog til hypotesetest: 17](#_Toc64977406)

[P-værdi (to-sidet anternativ) 17](#_Toc64977407)

[Konfidensintervallet 18](#_Toc64977408)

[Økonomisk versus statistisk signifikans 18](#_Toc64977409)

[Den restrikterede model 18](#_Toc64977410)

[F-test – multible lineære restriktioner 19](#_Toc64977411)

[OLS-estimatorens asymptotiske egenskaber 19](#_Toc64977412)

[Eksakte vs. asymptotiske egenskaber 19](#_Toc64977413)

[Konsistens 19](#_Toc64977414)

[Store tals lov 20](#_Toc64977415)

[Regneregler for konvergens i sandsynlighed 20](#_Toc64977416)

[Teorem 5.1: 20](#_Toc64977417)

[Teorem 5.2 – Asymptotisk normalfordeling af OLS-estimatoren 20](#_Toc64977418)

[Asymptotiske standardfejl for OLS-estimatoren 21](#_Toc64977419)

[OLS-estimatorens egenskaber 21](#_Toc64977420)

[Lagrange Multiplier (LM) test 22](#_Toc64977421)

[Måleenheder 23](#_Toc64977422)

[Skalering af højresidevariable 23](#_Toc64977423)

[Skalering af venstresidevariable 23](#_Toc64977424)

[Logaritmisk reskalering 23](#_Toc64977425)

[Standardiserede variable 23](#_Toc64977426)

[Funktionelle former 24](#_Toc64977427)

[Logtransformationer 24](#_Toc64977428)

[Kvadratiske led 24](#_Toc64977429)

[Interaktionsled (uden dummyvariabel) 25](#_Toc64977430)

[Variabel udvælgelse 25](#_Toc64977431)

[Alternativt goodness-of-fit measure 25](#_Toc64977432)

[Prædiktioner 25](#_Toc64977433)

[Kvalitativ information og dummy variable 25](#_Toc64977434)

[Dummy variabel med m=2 (binær variabel) 26](#_Toc64977435)

[Fortolkning af (dummy-estimatet) 26](#_Toc64977436)

[Politik evalueringer gennem dummy variable 26](#_Toc64977437)

[Kvalitativ information med m>2 kategorier 27](#_Toc64977438)

[Ulemper ved dummy variable 27](#_Toc64977439)

[Chow test 27](#_Toc64977440)

[Chow test med g=2: 28](#_Toc64977441)

[Chow test med g=m>2: 28](#_Toc64977442)

[Linear Probability Model 28](#_Toc64977443)

[Heteroskedasticitet 29](#_Toc64977444)

[Definition 29](#_Toc64977445)

[Konsekvenser af heteroskedasticitet 30](#_Toc64977446)

[Hvornår og hvor 30](#_Toc64977447)

[White: Heteroskedasticity-Consistent Covariance Matrix Estimator 31](#_Toc64977448)

[Standard overfor robuste standardfejl 31](#_Toc64977449)

[En heteroskedastisk robust hypotesetest 32](#_Toc64977450)

[Multiple restriktioner 32](#_Toc64977451)

[Test af Heteroskedasticitet 32](#_Toc64977452)

[Grafiske test 33](#_Toc64977453)

[Breusch-Pagan test 33](#_Toc64977454)

[White’s test 34](#_Toc64977455)

[Heteroskedasticitet – kommentarer: 34](#_Toc64977456)

[(In)efficiens 34](#_Toc64977457)

[WLS (weighted least squares) 35](#_Toc64977458)

[Linear Probability Model (Lineære sandsynlighedsmodel) 35](#_Toc64977459)

[Feasible GLS (FGLS) 36](#_Toc64977460)

[Procedure for FGLS (kogebog) 36](#_Toc64977461)

[Konklusion 36](#_Toc64977462)

[Misspecifikation af den funktionelle form 37](#_Toc64977463)

[Statistisk procedure: 37](#_Toc64977464)

[Formel test for misspecifikation: 37](#_Toc64977465)

[RESET – REgression Specification Error test 38](#_Toc64977466)

[Test af ikke nestede alternativer 38](#_Toc64977467)

[Tilgang 2: 38](#_Toc64977468)

[Tilgang 3: 38](#_Toc64977469)

[Konklusioner for ikke nestede alternativer 39](#_Toc64977470)

[Proxy variable 39](#_Toc64977471)

[Instrumental variabel (IV) estimation 40](#_Toc64977472)

[Korrelation versus kausalitet 40](#_Toc64977473)

[Instrumental variable (IVs) 40](#_Toc64977474)

[IV-estimatioon for den simple regression 41](#_Toc64977475)

[IV-estimatorens egenskaber 42](#_Toc64977476)

[Inferens med IV-estimatoren 42](#_Toc64977477)

[OLS versus IV-estimation 42](#_Toc64977478)

[Eksakt identifikation med multiple regressorer 43](#_Toc64977479)

[Multipel lineære regressionsmodel med endogene variable 43](#_Toc64977480)

[IV antagelser 44](#_Toc64977481)

[IV-estimatoren i det eksakt identificerede tilfælde 44](#_Toc64977482)

[Two-stage least squares (2SLS): Det simple tilfælde 45](#_Toc64977483)

[Hvornår er IV-estimatoren den foretrukne estimator? 46](#_Toc64977484)

[Vigtige pointer fra IV-estimatoren 46](#_Toc64977485)

[Overidentifikation med multiple regressorer 47](#_Toc64977486)

[Sammenligning af OLS og IV 47](#_Toc64977487)

[Two-stage Least Sqares (2SLS) 47](#_Toc64977488)

[Inferens 47](#_Toc64977489)

[Konklusion om overidentificeret IV-estimator 48](#_Toc64977490)

[Test af eksogenitet 48](#_Toc64977491)

[6-trins procedure for IV-estimation 49](#_Toc64977492)

[Stationære tidsserier 50](#_Toc64977493)

[Datatyper: Tværsnitsdata vs. Tidsserier 50](#_Toc64977494)

[Tidsserier 50](#_Toc64977495)

[Stokastisk proces 51](#_Toc64977496)

[Stationaritet 51](#_Toc64977497)

[Tidsafhængighed 52](#_Toc64977498)

[Statistisk tidsseriemodel (regressionsmodel): 53](#_Toc64977499)

[Ikke stationære tidsserier 53](#_Toc64977500)

[Detrending 53](#_Toc64977501)

[Differender 54](#_Toc64977502)

[Kointegration 54](#_Toc64977503)

[Statistisk model 54](#_Toc64977504)

[Autoregressiv model 54](#_Toc64977505)

[Autoregressive distributed lag model 55](#_Toc64977506)

[Antagelser for den lineære regressionsmodel: 55](#_Toc64977507)

[OLS-estimatoren: 56](#_Toc64977508)

[Middelrethed: 56](#_Toc64977509)

[OLS-estimatoren asymptotiske fordeling 56](#_Toc64977510)

[Autokorrelation i fejlleddet: 57](#_Toc64977511)

[Vigtige pointer: 57](#_Toc64977512)

[Regressionsmetoder med ikke stationære variable 58](#_Toc64977513)

[Misspecifikationstest 58](#_Toc64977514)

[Misspecifikationstest 1 – Ingen autokorrelation i fejlleddet 58](#_Toc64977515)

[Misspecifikationstest 2 59](#_Toc64977516)

[Misspecifikationstest 3 59](#_Toc64977517)

[Gentagne tværsnitsdata 59](#_Toc64977518)

[Gentagne tværsnitsdata i modsætning til tværsnitsdata 59](#_Toc64977519)

[Gentagne tværsnitsdata 60](#_Toc64977520)

[Test for strukturelle ændringer 61](#_Toc64977521)

[Politikanalyse med gentagne tværsnitsdata 61](#_Toc64977522)

[Politikanalyse med gentagne tværsnitsdata 63](#_Toc64977523)

[Dif-in-dif: Gennemsnit vs. Regression 64](#_Toc64977524)

[Grafisk illustration ad dif-in-dif 64](#_Toc64977525)

[Naturlige eksperimenter 65](#_Toc64977526)

[Paneldata for T=2 65](#_Toc64977527)

[Model med uobserveret heterogenitet 65](#_Toc64977528)

[Antageler for model med uobserveret heterogenitet 66](#_Toc64977529)

[Korreleret uobserveret heterogeitet 66](#_Toc64977530)

[First-difference estimation: 66](#_Toc64977531)

[Fordele og ulemper ved paneldata 67](#_Toc64977532)

[Vigtige pointer 67](#_Toc64977533)

[Paneldata 67](#_Toc64977534)

[Simpel panel data regressionsmodel for T>2 67](#_Toc64977535)

[Panel data regressionsmodel med uobserveret heterogenitet 68](#_Toc64977536)

[FD estimator 68](#_Toc64977537)

[FE estimator 68](#_Toc64977538)

[RE 69](#_Toc64977539)

[Inferens 69](#_Toc64977540)

[Argumenter til at foretrække en tilgang frem for den anden 70](#_Toc64977541)

[Huskeliste til at vælge mellem: 70](#_Toc64977542)

# Matrice regneregler

# Den simple lineære regressionsmodel

## Ordbog

* Afhængig variabel (venstresidevariabel)
* Forklarende variabel (højresidevariabel)
* Uobserveret fejlled (udeladte variable og målefejl)
* Skæring (konstantled)
* Hældningsparameter (hældningskoefficient)

Konstantled og x kaldes regressorer.

## Antagelser for SLR-modellen

**(2.5)** Middelværdien af u er 0:

**(2.6)** Den betingede middelværdi af u er:

## Method of Moments – MM (Momentestimation)

Idéen med MM er, at erstatte det teoretiske moment med det empiriske moment.

Man tager derfor et gennemsnit.

Middelværdien af y:

MM-estimatoren af middelværdien er:

Variansen af y, hvor :

MM-estimatoren af variansen er:

Kovariansen mellem x og y:

MM-estimatoren af kovariansen er:

## Udledning af OLS-estimatoren

Benytter tre betingelser:

1. Regressionsmodel:
2. Antagelser

## Antagelser for den simple lineære regressionsmodel (SLR.1-4):

1. Antager en statistisk model for y der er en lineær regressionsmodel af x og parametre. (En antagelse om, hvordan verden har genereret y).
2. At der er data på x og y. Der skal være n observationer, og vi trækker observationer, der er uafhængige af hinanden.
3. Der er variation i x, da hvis x er konstant kan y ikke være en funktion af denne.
4. Den betingede middelværdi af fejlledet givet x er 0. Der er ingen systematisk sammenhæng mellem fejlledet og vores regresser x.

MM-estimatoren løser to ligninger med to ubekendte:

OLS-estimator kan også udeledes ved at minimere summen af kvadrerede residualer:

**W2.5-W2.6**

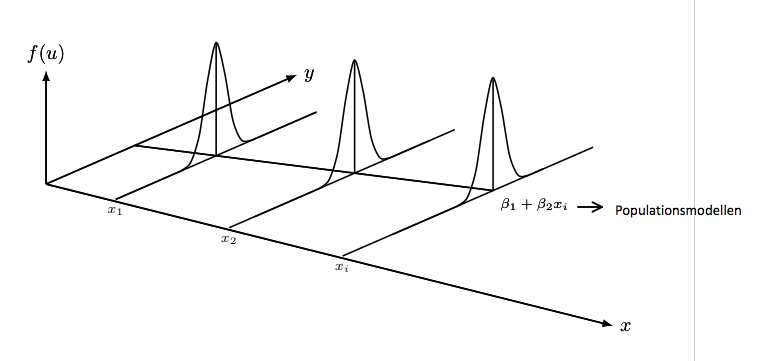
## Variansen af OLS-estimatoren

Siger noget om præcisionen af vores estimator. Variansen på tværs af de mange hypotetiske stikprøver. På baggrund af en stikprøve beregnes et estimat af en varians.

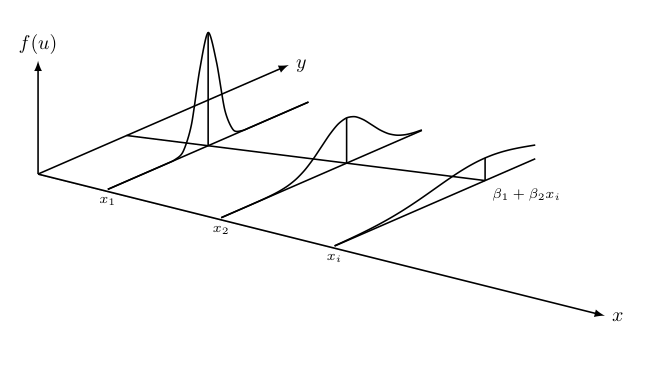
## Antagelse om varians på fejlledet (SLR.5)

**Homoskedasticitet:** Antager variansen af u er konstant.Afledt af græsk og betyder spredning er den samme uanset, hvilken x vi kigger på.

Variansen af u givet x er et udsagn om variansen, hvilket er forskellig fra SLR.4, hvor vi snakker om forventningen til fejlledet u. Det modsatte af SLR.5 havde vi stadig, at OLS-estimatoren var middelret. Kræver kun SLR.1-4 er opfyldt.



**Hetroskedasticitet** er SLR.1-4 opfyldt, men SLR.5 er ikke opfyldt. Dvs. variansen af u afhænger af x. Dette ses nedenfor:



I hver x forstiller vi os mange forskellige træk med samme x værdi og variation i fejlledet. Middelværdien ligger i den sande populationsmodel, dvs. middelværdi lig nul. Alle fordelinger har nogenlunde samme form og har samme varians. Dvs. homoskedacitet har vi samme varians på tværs af x-værdier.

*SLR.1-5 kaldets* ***Gauss-Markov*** *antagelserne.*

kan vi tænke på som den ubetingede varians af fejlledet.

## Theorem 2.2

Variansen af OLS-estimatoren givet x er variansen i tælleren og den kvadrerede afvigelse fra gennemsnittet af x i nævneren.

Se bevis for dette theorem på L4 slide 9-10.

**En estimator for variansen** af fejlledet er:

## Standardfejlen for OLS-estimatoren

Standardfejlen er et mål for variation i OLS-estimatoren (på tværs af stikprøver).

Skal helst være så lille som muligt.

Beregnes automatisk i STATA. Man kalder dette for standardfejlen for OLS-estimatoren. Vi kalder ikke denne for standardafvigelse, da vi har erstattet den ukendte varians med et estimat. Den sande standardafvigelse for OLS-estimatoren er kvardratroden af . Tallet i sig selv er ikke så meningsfyldt, men dette mål er byggesten ift. hypotesetest.

Forskellige faktorer påvirker standardfejlen:

* Når der er mere variation i x, falder standardfejlen.
* Når der er mere variation i stiger standardfejlen.
* Når der er flere observationer, falder standardfejlen.
* Hvis n øges, og da x er lig gennemsnittet sker der intet i nævneren af standardfejlen. Summen af de kvadrerede residualer er også uændret. Det eneste der ændres er, at antallet af observationer nu er n+1. Derfor bliver mindre, derfor reduceres standardfejlen fro OLS-estimatoren. Med flere observationer øger vi generelt præcisionen af OLS-estimatoren. I princippet kan man have tilføjet en outlier, hvorfor det kan være muligt, at man ved at tilføje en observation øger standardfejlen.

**W.3.1-W.3.2 +D1-D3, E1?**

## Definition af MLR-modellen

* Y: er den afhængige variable
* er de k forklarede variable
* : konstantled
* k hældningskoefficienter
* u: fejlleddet
* k: antallet af forklarende variable
* k+1: paramter i alt

Man kan bruge flere fleksible funktionelle former så længe modellerne er lineære i parametrene!

MLR-modellen kan skrives i matrixnotation:

OLS-estimatoren kan udledes som en momentestimator (MM).

## Teoretiske momenter:

Middelværdien af en vilkårlig forklarende variabel også er lig 0.

OLS-estimatoren for MLR-modellen

Kan også udledes som:

## Fuld rang:

Søjlerne i X skal være lineære uafhængige, dvs. de forklarende variable må godt være indbyrdes korreleret i den multiple regressionsmodel, så længe de ikke er perfekt korrelerede. Dette skyldes, at vi aldrig kan estimere de partielle effekter af variable, som varierer 1-til-1.

## Fortolkning af

Hvis alle øvrige forklarende variable holdes konstante undtagen , gælder der at:

Tillader, at vi estimerer partielle effekter (ceteris paribus betragtninger). Dette kan gøres selvom data ikke kan holde alle faktorer faste.

Hvis vi tager log til ovenstående ses let, at vores betaværdier er den forventede effekt af en marginal ændring i effekt på y, hvis vi holder alt andet konstant.

## Sammenhæng mellem MLR og SLR (Frish-Waugh Teoremet)

I tilfælde med k=2 ser den statistiske model ud som følger:

Man kan replikere MLR ved at beregne to simple regressionsmodeller.

1. Regressere og gemme residualerne . Dvs. opstiller en regressionsmodel, hvor man estimerer hjælperegressionen:

Residualerne beregnes som .

1. Regressere y på residualerne Dvs. opstiller en regressionsmodel, hvor y er venstresidevariabel og regresserer y på Residualerne har et gennemsnit på nul, hvorfor OLS-estimatoren kan fås som;

.

Residualerne er ortogonale og er dermed perfekt ukorrelerede med Hvorfor residualerne kan fortolkes som den resterende variation i efter at have kontrolleret for .

## OLS residualer

Samme definition som i SLR

Mekaniske egenskaber:

Punktet ligger på regressionslinjen.

## Forklaringsgrad:

Følger samme definitioner som introduceret tidligere.

Variation i y kan dekomponeres i to komponenter:

1. Total variation: SST
2. Forklaret variation: SSE
3. Residual variation: SSR

Forklaringsgraden, defineres som:

Stiger mekanisk når vi medtager flere x-variable. Dette skyldes, at vores model bliver bedre jo flere forklarende variable, dvs. residualerne bliver alle større jo flere x’er. Modellen bliver altså bedre til at forklare y. Dette gælder ikke, hvis man tager en irrelevant x-variabel med, da vil være uændret.

Man skal ikke bruge til at bestemme, hvilke variable der skal indgå i modellen! Man skal være påpasselig med at sammenligne på tværs af modeller, hvor venstresidevariablene er forskellige.

Eksempel: Hvis vil det betyde at man forklare 10% af variationen.

**W3.3-W3.4+E2**

## MLR-antagelserne – Gauss-Markov antagelser

MLR.1: Lineær i parametrene:

På matrix form:

Aldrig automatisk opfyldt, men det er en antagelse om, hvordan verden har genereret y.

MLR.2 - Tilfældig stikprøve:

Vi har tilfældigt udvalgte observationer

fra populationen. Dvs. tilfældig stikprøve (X,y) med n observationer.

MLR.3 - Fuld rang af X:

Der er variation i x-variablene og de er ikke perfelt korreleret. Udelukker, at individuelle kontrolvariable er perfekt korreleret. Dvs. X har rang k+1.

MLR.4 - Den betingede middelværdi af fejleddet er 0:

u er ukorreleret med alle funktioner af kontrolvariable. Dette er en meget restriktiv antagelse. Hvis MLR.4 er opfyldt, er de forklarende variable eksogene. Hvis er korreleret med u (brud på MLR.4) kaldes endogen. Hård antagelse om, hvad vi kan observere i fejlledet og i vores forklarende variable.

MLR.5 – Homoskedasticitet (Gauss-Markov antagelserne)

På matrix form:

hvor er en (n x n) identitetsmatrix.

Hvis MLR.5 ikke er opfyldt, er fejlleddet heteroskedastisk. F.eks. .

MLR.1-6 kaldes **de klassiske antagelser for MLR-modellen**. Dette kaldes de klassiske, da udgangspunktet engang altid var normalfordelte fejlled.

MLR.6 - u er uafhængig af og normalfordelt:

* u har en betinget middelværdi givet x lig nul (jf. MLR.4) og et homoskedastisk fejlled.
* u indeholder alle faktorer som ikke er med i modellen.
* Er det realistisk, at u antages at være normalfordelt? Måske
* Men: Svært at forene fx en diskret y-variabel med MLR.6. F.eks. hvis y antager ti forskellige diskrete værdier, dvs. på venstresiden har vi noget diskret og på højresiden har vi noget, der er kontinuert på højresiden jf. MLR.6. Dette kan være svært at forene.

## Teorem 3.1: Middelrethed af OLS-estimatoren

Antag vi har en estimator b (stokastisk variabel)

En estimator b for β er middelret hvis:

for alle værdier af β

Når antagelse MLR.1-4 er opfyldt gælder,

OLS-estimatoren vil i gennemsnit give den rigtige værdi (på tværs af mange stikprøver) Med kun en stikprøve beregnes kun et estimat, der afviger fra den sande værdi.

Middelrethed viser altså om estimaterne fordeler sig ’rundt om’ den sande værdi, og dermed ikke kun ligger over eller kun ligger under.

## Overspecificeret regressionsmodel

Hvad er konsekvensen af at inkludere en irrelevant variabel?

En irrelevant variabel er en variabel, der i den sande model har en parameter lig nul, dvs. .

OLS er fortsat middelret, i gennemsnit vil OLS beregne de sande parameterværdier. Det sande svar for xj er 0. Det har altså ingen konsekvenser for om estimatoren er middelret. Det har dog den konsekvens, at det påvirker (øger) variansen på OLS-estimatoren.

9

## Underspecificeret regressionsmodel (udeladt variabel bias k=2)

Hvad er konsekvensen af, at ekskludere en relevant variabel?

Hvis vi kun kan regresserer y mod en forklarende variabel for lidt. Dette problem står vi overfor ved ikke eksperimentel data, hvilket vi typisk arbejder med.

OLS er ikke længere middelret (udeladt variabel bias k=2):

Hvis er forskellig fra nul er OLS biased/ikke længere middelret. Der er dermed to undtagelser, hvor vi stadig har en middelret estimator:

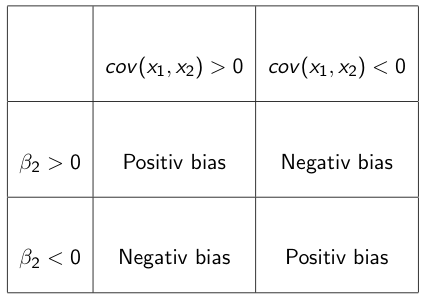
1. i populationen.
2. dvs. hvis er en relevant variabel, men de to forklarende variable er perfekt ukorrelerede, så er OLS-estimatoren stadig middelret.

Vi kan bestemme **retningen** af bias på baggrund af argumenter vedrørende ukendte størrelser i populationen. Vi forholder os til kovarians mellem den ukendte udeladte variabel og de resterende forklarende variable.

**Størrelsen** af bias afhænger bl.a. af : jo mere variation i jo mindre bias.

Udeladt variabel bias for OLS-estimatoren med k=2:

Når estimatoren er ikke-middelret



## Udeladt variabel bias for OLS-estimatoren med k>2

Model:

Antager, at er en udeladt variabel.

Forventnungen ved en udeladt variabel bias:

hvor måler den partielle korrelation mellem og .

Dette betyder:

Dvs. man skal tænke på partielle korrelationer og ikke blot kovariansen. Dvs. hvis der er en udeladt variabel bias er der bias i alle OLS-estimater.

## Varians af OLS-estimatoren

Under antagelse af MLR. 1-5

Se bevis på slide 6 s. 16.

Med data på x-variable kan vi finde ovenstående. Vi skal dog have et estimat for .

Hvis vi tager udgangspunkt i matrix-formen fra før kan man omskrive diagonalelementerne vha. simpel algebra

er forklaringsgraden for en hjælperegression af på de øvrige kontrolvariable. Denne fanger, hvor stor samvariation, der er mellem de forskellige x-variable. Hvis er meget lille er tæt på at være ukorreleret med alle andre x-variable og omvendt. er et mål for hvor stor lineær samvariation, der er mellem x-variable. Hvis x-variablene korrelerer meget højt bliver det alt andet lige svære at få den partielle effekt af, hver enkelt variabel. Variansen for den generelle OLS-estimator er større end eller lig variansen for den simple.

OBS: Middelrethed er vigtigere end varians!

Der er tre komponenter, der påvirker variansen af OLS:

En højere varians af fejlleddet =>

Mere variation i

Samvariation blandt de forklarende variable:

* Jo tættere R^2 er på 1, dvs. jo større samvariation
* Hvis

Irrelevante variable øger variansen af OLS-estimatoren!

**W3.5-W3.6 +D4-D5, E2**

## Misspecificerede modeller

Forskellige scenarier ift. hvornår modeller er misspecificerede:

* Hvis og er ukorreleret:

## Estimator af variansen af fejlleddet:

## Teorem 3.3

Hvis Gauss-Markov antagelserne (MLR.1-5) er opfyldte, er en middelret estimator

Standardfejlen for OLS-estimatoren i det generelle tilfælde.

Standardfejlen giver et mål for variation i OLS-estimator (på tværs af stikprøve). Standardfejlen er vigtig ift. hypotesetest.

## Gauss-Markov teoremet (OLS er blot et vægtet gennemsnit)

Hvis MLR.1-5 holder, er OLS-estimatoren en bedste lineære middelrette estimator i klassen af lineære middelrette estimatorer. Den bedste betyder, at den har den mindste varians. Man siger OLS er BLUE når MLR.1-5 er opfyldt:

* Best (dvs. mindste varians)
* Lineear
* Unbiased
* Estimator

Vi vil altid bruge denne til regressionsanalyse, da den er BLUE.

OLS estimatoren beregner i princippet et vægtet gennemsnit af vores observationer på den afhængige variabel.

Vi kan omskrive OLS-estimatoren for SLR-modellen:

For disse vægte gælder, der nogle **egenskaber** (hvor ):



Alternative lineære estimatorer til OLS kan findes ved at benytte andre vægte. Vægtene skal opfylde ovenstående egenskaber, da dette sikrer estimatoren er middelret. De er dog ikke lige så gode, da OLS har den mindste varians. Den har mindste standardfejl, hvorfor man også kan få de bedste indblik i partielle effekter vha. denne estimator.

Anternative estimatorer refereres til som:

A matricen er en funktion af X. Dvs. samlet alle vægte i en skalar kaldet A.

**W4**

## OLS-estimatorens fordeling

For at lave hypotesetest er det vigtigt at kende fordelingen af .

OLS-estimatoren er lineært relateret til fejlledet

Dette er fundet ved MLR.1 antagelsen. Hvis vi laver en antagelse om, at u er normalfordelt ”arver” betinget på X samme fordeling.

### Teorem 4.1

Under MLR.1-MLR6 gælder:

hvor

Denne kan normaliseres således:

Da, er ukendt estimeres denne varians af fejlledet.

### Teorem 4.2:

Under MLR.1-MLR6 er den standardiserede OLS-estimator t-fordelt:

t fordelt med antallet af frihedsgrader (n-k-1), hvo k+1 er antallet af parametre vi estimerer.

t-fordelingen konverger mod N(0,1), når antallet af frihedsgrader (df) er stor (n>120). Når vi skal afgøre om der er et stort datasæt eller ej er svært at afgøre, men vi følger Woldridge, hvor over 120 i en stikprøve er stor.

# Hypotesetest:

## Hypotesetest ved simple hypoteser:

Vi forsøger at gætte på, hvad den sande parameter er lig. Dette postulerer vi under en nulhypotese, dvs. postulerer en bestemt værdi af OLS-estimatet. Hypotese vedrørende en parameter og dermed kun én restriktion:

Vi vil se, hvordan vores OLS-estimat afviger fra vores hypotesetest. Hvis der er stor afvigelse synes det at foreslå, at nulhypotesen er forkert og omvendt.

t-teststørrelsen måler afvigelsen i forhold til standardfejlen:

Antaget at nulhypotesen er sand har størrelsen, t, en fordeling. Denne vil være t-fordelt med antal frihedsgrader (n-k-1). . Ligner normalfordelingen, så alle værdier vi kan beregne for vores teststørrelse fra () kan forenes med denne fordeling.

Hypotesetesen kræver, at der opstilles en alternativ hypotese. Der findes to typer:

* Ensidet alternativ: Dvs. vi er interesseret i at teste ift. afvigelser i en bestemt retning.
* Tosidet alternativ: . Hvis man ikke er interesseret i at teste afvigelsen i en bestemt retning, men afvigelser i begge retninger vil være evidens for alternativet.

STATA rapporterer altid et standard signifikans test, hvor der testes mod et tosidet alternativ til nulhypotesen.

## Kogebog til hypotesetest:

For at fastlægge, hvilken hypotese, der holder gøres følgende:

* Fastlæg signifikansniveauet = sandsynligheden for at afvise selvom den er sand. Vi bruger typisk et signifikansniveau på 5%.
* Fastlægger en nulhypotese og en alternativhypotese (evt. flere nulhypoteser).   
  q lineære restriktioner på modellens parametre (antal lighedstegn i både og ).
* Beregner teststørrelse. F.eks. t-test.
* Bestem det kritiske område givet signifikansniveauet.
* Afvis , hvis teststørrelsen ligger i det kritiske område.

Afvis , hvis blot en af de q restriktioner ikke holder.

* Alternativt udregn p-værdien.

## P-værdi (to-sidet anternativ)

Sandsynligheden for at observere en mere ekstrem t-teststørrelse under antagelse af, at er sand.

Man beslutter om nulhypotesen kan afvises som følger.

* Små p-værdier er evidens i mode
* Store p-værdier er evidens i mod

Findes i STATA vha. kommandoen ttail(df,t)

## Konfidensintervallet

Konfidensintervallet for er , hvor:

hvor c er 97,5% fraktilen i en i distributionen.

**Fortolkning:** Det beregnede konfidensinterval indeholder den sande parameter i 95% af tilfældene (på tværs af mange stikprøver). I 5% af tilfældene vil konfidensintervallet ikke indeholde den sande parameter.

Hvis vi tester overfor , vil vi altid forkaste til frodel for , hvis

## Økonomisk versus statistisk signifikans

De ovenstående er ikke ens. Dv s. statistisk signifikans er ikke det samme som økonomisk signifikans.

Økonomisk signifikans vedrører størrelsen af de parametre vi har estimeret. Disse estimater skal vurderes relativt til:

* Den gennemsnitlige værdi af den afhængige variabel, y
* Tidligere studier (dette kan ikke gøres til eksamen, men kan bruges til en BA)
* Andre relevante sammenligninger (fx afkast af uddannelse ift. afkast af erfaring)

Fejl, der kan opstå:

* Man skal altid fortolke .
* Statistisk insignifikans udelukker ikke økonomisk signifikans.

Dvs. selvom OLS-estimatet er statistisk insignifikant er det stadig det bedste OLS-estimat, men det er bare behæftet med stor usikkerhed.

* Det er også vigtigt at huske under hvilke antagelser inferens er gyldig.
* Muligt at have lavet Type-1 og Type-2 fejl. Dvs. vi kan afvise nulhypotesen selvom denne faktisk er sand.

## Den restrikterede model

Den restrikterede model er den oprindelige model med pålagt.

F.eks.

Den restrikterede model bliver da under ovenstående nulhypotese:

Denne model er stadig lineær i parametrene og vi kan estimere denne ved OLS. Vi har nu en ny venstresidevariabel lig Dette vil kræve, at der genereres en ny variabel i STATA.

## F-test – multible lineære restriktioner

Dvs. ved test af multible lineære restriktioner bruges F-test:

Vi bruger informationer fra både urestrikteret og retrikteret model.

Summen af de kvadrerede residualer er større i den restrikterede model end i den urestrikterede.

Den restrikterede model kan ikke forklare y bedre end den oprindelige model => F>0.

* q er antal restriktioner.
* n-k.1 er antal frihedsgrader i den urestrikterede model.

Hvis MLR.1-MLR.6 er opfyldt, gælder:

Hvor , hvis en restriktion testes overfor et tosidet alternativ.

F fordelt er forholdet mellem to -fordelte variable.

**W5**

# OLS-estimatorens asymptotiske egenskaber

## Eksakte vs. asymptotiske egenskaber

* OLS er middelret, hvis MLR.1-4 gælder
* OLS er BLUE, hvis MLR.1-5 gælder.
* OLS er normaltfordelt, og t og F-tests kan bruges til inferens, hvis MLR.1-6 gælder.

Ovenstående egenskaber gælder for et hvilket som helst fast n. Dvs. hver strikprøve, der trækkes har samme antal observationer.

Hvis vores stikprøver derimod stiger, dvs. hvis vil estimatoren kollapse og blive lig den sande værdi, hvis estimatoren er konsistent.

## Konsistens

er en estimator af baseret på

Så er en konsistent estimator af , hvis:

Estimatoren konverger i sandsynlighed mod den sande værdi:

Grænsesandsynligheden, , er lig den sande værdi når n går mod uendelig såfremt estimatoren er konsistent.

Minimumskrav for en meningsfuld estimator.

OLS er **inkonsisten** hvis:

Inkonsistens (eller den asymptotiske bias i den simple model (størrelser fra populationen ikke fra stikprøven):

Dette problem forsvinder ikke, selvom n går mod uendelig. Stort datasæt kan stadig have asymptotiske bias og endogenitetsproblemer, hvis fejlledet er korreleret med en regressor.

Eksempler på inkonsistens:

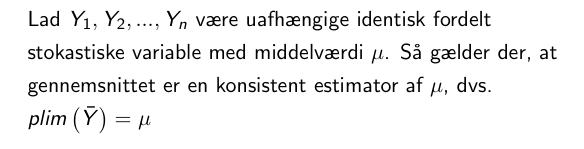
* Variansen ikke går mod nul når n går mod uendelig
* Estimatoren koncentreres om en anden værdi end den sande værdi når n går mod uendelig.

**Middelrethed versus konsistens:**

En middelret estimator er ikke nødvendigvis konsistent. Det kræver, at estimatoren varians går mod nul når n går moduendelig.

Kan godt have en estimator, der er konsistent men ikke middelret. Dette ser vi på senere med IVestimatoren.

## Store tals lov



De empiriske momenter konvergerer i ssh. mod de sande momenter når n går mod uendelig.

## Regneregler for konvergens i sandsynlighed

1. For en kontinuert funktion g gælder:
2. Hvis og gælder følgende:

## Teorem 5.1:

OLS-estimatoren, , er en konsistent estimator af , når MLR.1-4 er opfyldt.

Når vi beviser konsistens kræver vi kun, at MLR.4’ er opfyldt. Dvs. mindre restriktiv end at bevise middelrethed.

MLR.4:

## Teorem 5.2 – Asymptotisk normalfordeling af OLS-estimatoren

Under MLR.1-5 gællder der, at den er assymptotisk normalfordelt.

, hvor og

Dette resultat bygger på den centrale grænseværdisætning (C.12 i Woldridge). Kvadratroden af n sikrer vi konverger mod en asymptotisk normalfordeling.

er residualerne fra en regression af på de øvrige forklarende variable.

er den sædvanlige OLS standardfejl.

I kap. 4 antog vi normalfordelte fejlled. Nu bruger vi teorem 5.2, der siger den asymptotiske fordeling approksimativt er normalfordelt.

Ingen faste regler for, hvornår n er stor nok til at anvende asymptotiske resultater.

## Asymptotiske standardfejl for OLS-estimatoren

Asymptotisk varians for OLS:

Se hvordan komponenter er i grænsesandsynligheden - plims:

konverger mod 0 med hastigheden 1/n, hvorfor standardfejlen konverger mod nul med hastigheden (da jf. ovenstående beregninger er n i nævneren).

## OLS-estimatorens egenskaber



Assymtotisk efficient er pendant til Gauss-Markow teoremet med n, der er voksende.

## Lagrange Multiplier (LM) test

Bruges til test af multible lineære restriktioner:

* F-test
* Lagrange multiplier test

LM-testet kræver ikke, at den urestrikterede model estimeres.

Testet er gyldigt, hvis MLR.1-MLR.5 er opfyldt. Kun valid assymptotisk dvs. 🡪

Model:

Nulhypotese:

y er dermed ikke en funktion af disse variable.

Restrikteret model (pålagt nulhypotesen):

Under

er ukorrelerert med de udeladte variable .

Tager vores residualer og regresserer dem på alle x-variable. Dette vil sige os noget om, der er korrelation mellem højresidevariable og residualer.

Noter fra hjælperegressionen.

Beregn LM-teststørrelsen. LM=

Assymptotisk fordelt LM, hvor q (antal frihedsgrader)er antallet af restriktioner. Vores estimator er normalfordelt, hvorfor residualerne er normalfordelt. er summen af de kvadrerede residualer 🡪 fordelt stokastisk variabel.

De to teststørrelser er assymptotisk ækvivalente, men vil være forskellige i endelige stikprøver.

Den ene er ikke bedre end den anden.

**W6**

# Måleenheder

Interesserede i, hvordan vi påvirker OLS-estimator når vi reskalerer.

Mincer regression:

Forgangsmand til empirisk arbejdsmarkedsforskning. Vi kigger på, hvad der sker, hvis vi ændrer skala på højre eller venstresidevariable.

Denne opfylder MLR.3, da exper ikke er lineært korreleret med .

## Skalering af højresidevariable

Ændres uddannelse til at være årsbasis i stedet for månedsbasis fås:

Vi får års afkast af uddannelse er månedsafkast gange antal måneder på et år.

**Estimater og standard fejlen bliver reskaleret men alt andet forbliver konstant!**

Dvs. t-test forbliver også konstant.

Generelt set er skala af x-variable ikke et problem.

## Skalering af venstresidevariable

Reskaleres timelønninger (ikke log timeløninger):

Vi reskalerer vores OLS-estiamtor og vores standard fejl.

SSR, SST (variation i y), SSE og er reskaleret.

og t-test er uændrede. Intuitivt giver det mening, da skalaen på y ikke skal ændre forklaringsgraden af modellen på y. t-test er invariante overofr, hvordan vi vælger at skalere y.

## Logaritmisk reskalering

Fører til samme statistiske model, men med en ny skæring lig . Når vi reskalerer log-transiformerede variable ændres skæringen alene.

*Se flere mellemregninger på slide.*

## Standardiserede variable

Hvis relevant skala ikke er helt oplagt kan man standardisere

Hvis der ikke findes en naturlig skala normaliserer man ift. standardafvigelsen som referencepunkt.

Ved at standardisere variable fås:

Når x ændres med en standard afvigelse, forventes y at ændres med standardafvigelse.

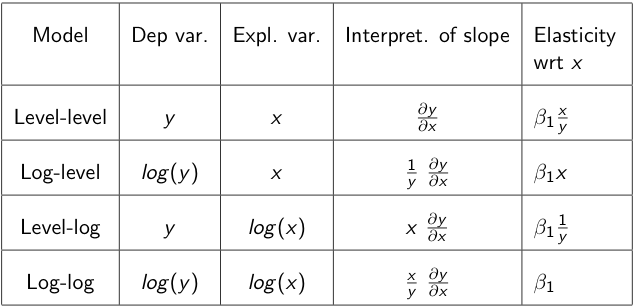
## Funktionelle former

MLR antager modellen er lineær i parametrene ikke i variablene!

Funktionelle former har implikationer for fortolkninger. Eksempler:

* Log-transformerede variable relative ændringer
* Kvadratiske størrelser kan bruges til at estimerer stigende eller aftagende marginale effekter
* Kan gerelaisere sin funktionelle form vha. interaktionsled.

Fortolkning af parametre:



### Logtransformationer

Inflydelsen på ekstreme observationer bliver markant mindre ved log-transformationer. Dvs. outliers bliver en mindre bekymring.

Problemet ved log-transformationer er, at vi ikke kan tage logaritmer til negative værdier eller 0.

### Kvadratiske led

Når vi inkluderer ikke lineære led besværliggør vi fortolkning af parametre. Man skal evaluerer afkast af x ved relevante værdier af x for ta kunne opnår forklaring. F.eks. kan relevante værdier være gennemsnit af x fra data.

Man skal sikre sig, at man ikke ekstrapolerer resultater.

### Interaktionsled (uden dummyvariabel)

Hvordan samspil mellem og påvirker y-variablen.

Interaktion kan f.eks. være produkt mellem to x-variable.

Her skal vi ligesom med kvadratiske led evaluere med relevante værdier for x.

OBS: Pas på om resultater ekstrapoleres.

## Variabel udvælgelse

Hvordan vælger vi de x-variable, der skal indgå på højresiden af vores model. Dette kan f.eks. gøres ved:

* eller udføre relevante statistiske test; . Dette er en hundrede procent statistisk tilgang.   
  Man skal estimerer sin model ift. den hypotese man er interesseret i at teste!
* Man skal også bruge fortolkning af de parametre til variablene, der indgår i modellen. En mere økonomisk tilgang.

Man skal bruge både statistik og økonomi til at udvælge relevante højresidevariable.

## Alternativt goodness-of-fit measure

mekanisk egenskab, at den altid stiger når vi medtager flere x-variable. Varierer afhængigt af, hvor vigtig den medtagende variable er ift. forklaring af y.

Kan bruge alternativt - mål – justeret :

Den straffer modellen når man øger k. Dvs ikke længere den systematik, at når man øger k øges .

Kun hvis x-variablen vi medtager er nogenlunde signifikant vil stige.

Husk: er normalt nødvendigt for at afvise .

## Prædiktioner

Kan vi prædiktere ting på baggrund af vores regressionsmodeller. Ja det kan vi, vores   er et estimat for vores sande med ukendte forventning.

**W7**

# Kvalitativ information og dummy variable

Indtil nu har vi arbejdet med kvantitative variable.

Vi medtager kvalitativ information f.eks. køn, postnummer, lønforskelle mellem privat og offentlig, deltids- eller fuldtidsbeskæftiget eller helbred. For at kunne tage disse type forklarende variable med skal vi igennem lidt detaljer.

## Dummy variabel med m=2 (binær variabel)

En dummy variabel, buges til at forklare kvalitativ information med to udfald. Dvs. det er en binær kan variabel, der enten kan være aktiveret, 1, og deaktiveret, 0.

Referencekategorien er defineret som der, hvor dummyvariablen er deaktiveret. Referencekategoriens tal afhænger af hvilken fortolkning af OLS-estimater man ønsker senere ved f.eks. hypotesetest. Denne er dog oftest sat lig 0.

Kan alternativt kaldes indikatorer, binære variable, men disse er ækvivalente med dummy-variable.

Man medtager dummy variable i MLR ligesom normale regressorer (x-variable).

Begge dummy-variable kan ikke medtages på samme tid. Dette skyldes, at der gælder identiteten . Dette kaldes også dummy-fælden. Se L12 slide 10.

Benyttes den modsatte dummy variabel i regressionsmodellen får man samme resultat blot med modsat fortegn. Dette er svagheden ved approksimationer ved log-level, da man blot får modsat fortegn og der kan være forskel i niveauer mellem de to udfald af variablen. Her skal man bruge eksakt procentivse forskel se længere nede.

### Fortolkning af (dummy-estimatet)

Dummy tillader et forskelligt niveau/en forskellig skæring mellem to typer f.eks. mand og kvinde. Forskelle i løn mellem mand og kvinde kaldes løngabet.

Fortolkning af parameteren ift. dummyvariablen er den forventede forskel mellem to kategorier betinget på vores resterende x-variable er fastholdt. Dvs. en alt andet lige betragtning.

Effekten af en dummyvariabel kan illustreres grafisk som et skift i skæringen (de resterende variable er restrikteret fastholdt).

Hvis den afhængige variabel er i niveauer, er parameteren fortolket som den absolutte forskellen mellem de to kategorier.

Hvis den afhængige variabel er i logaritmer, er parametren approksimativt fortolket som den procentvise forskel mellem de to kategorier. Den eksakte procentvise forskel mellem de to kategorier er givet som:

### Politik evalueringer gennem dummy variable

Disse dummy variable anvendes oftest ved politik evalueringer f.eks. jobtræning på indtjeningsmuligheder.

Her kan man tænke på dummy variable som en treatment gruppe; de der er aktive i et program og de, der ikke er en del af programmet, som også kaldes kontrolgruppen.

Hvad er den kasuale effekt af programet, der skal man undersøge om LR.1-4 er opfyldt. Hvis de er angiver -værdien effekten af at have deltaget i programmet.

Oftest er der kæmpe endogenitetsproblemer, da man oftest skal melde sig til sådan nogle programmer, og der derfor er individuelle præferencer, der er uobserverbare, hvilket medføre MLR.4 ikke er opfyldt. Dette giver anledning til bias.

## Kvalitativ information med m>2 kategorier

Hvis kvalitative variable med m kategorier:

* m-1 dummy variable kan medtages
* Den kategori uden en dummy variabel er referencekategorien (kun vigtig ift. fortolkning af parametrene, ændres denne skes en mekanisk ændring i parameterestimaterne)
* Hvis m dummy variable medtages ender vi i en dummy fælde

Parametrene for dummyvariablene angiver forskellen mellem den pågældende kategori og referencekategorien alt andet lige.

Valg af referencekategori er kun vigtig for fortolkningen af parameterestimaterne men ikke vigtig for estimationen og prædiktionen.

## Ulemper ved dummy variable

Den store fordel er, at vi kan bruge meget mere fleksible funktionelle former ved dummy variabel specifikation.

Ulempen er, at vi medtager mange forklarende variable, hvorfor k bliver neget stor, hvilket er uhensigtsmæssigt, hvis data ikke er tilstrækkeligt.

## Chow test

Kan vi antage, at alle parametre er forskellige på tværs af alle grupper?

Chow test gør det muligt at at opstille en hypotese om, at parametre er ens i to specifikationer og teste denne hypotese. Hvis man forventer heterogene effekter - parametrene varierer på tværs af grupper – kan man bruge denne test. Man skal være påpasselig med, at man øger chancen for type 1 fejl jo flere specifikationer, dvs. hvis man forventer 0 effekt skal man ikke bruge denne.

Hvis vi er i et tilfælde, hvor vi har to grupper f.eks. en kvinde variabel. Her angiver man parametre med to fodtegn. Det første er hvilken gruppe man indgår i (f.eks. gruppespecifik skæring). Dvs. alle modellens parametre fuld heterogenitet, dvs. parametre er forskellig fra y afhængig af, hvilken gruppe man er i.

Man undersøger nulhypotesen om, at der ikke er nogen forskel på tværs af de to grupper. Dvs. alle parametre er de samme på tværs af de to grupper. Der er multiple restriktioner.

Alternativhyposen er blot, at ikke er sand.

Vi tester k+1 restriktioner ved at bruge en Chow test. Denne er i overensstemmelse med en F-test.

Man skal estimerer regressionsmodellen for tre forskellige samples. Bruge SSR for samlet model og gruppespecifikke:

* Regression af gruppe 1 alene for at få
* Regression af gruppe 2 alene for at få
* Regression for samlede model 8begge grupper) fpr at få

Kønsvariablen indgår ikke som en separat regresser, da den ikke har variation indenfor en gruppe.

### Chow test med g=2:

Har man disse tre størrelser kan man opstille en Chow test med g=2 grupper:

Hvor er svarende til den urestrikterede model.

Under nulhypotesen er den statistiske test F-fordelt med frihedsgrader.

Hvis der antages at variansen af u (fejlledet) er den samme i de to grupper, er Chow testen gyldig under de samme antagelser (MLR.1-5 dvs. Gauss-Markow) som for F-testen.

**Den fulde interaktionsmodel:**

Man kan omskrive Chow testen med en dymmyvariabel, der er interageret med alle x’er. Denne kaldes den fulde interaktionsmodel. Mereffekter er forskellen mellem gruppespecifikke parametre. Denne hypotese kan testes vha. en almindelig F-test.

Nulhypotesen siger, at alle mereffekterne mellem de to modeller er lig nul. Alternativhypotesen siger blot, at de er forskellige fra nul:

Når vi antager inferrans (dvs. hvis MLR.5 ikke er opfyldt) er det lettere at benytte denne test end Chow testen, da denne er mere robust.

### Chow test med g=m>2:

Den generaliserede Chow test med g=m>2 grupper:

Denne model bruges når der er en diskret afhængig variabel med to værdier. Lineær regression har store vanskeligheder, hvis man har en y-værdi, der er diskret.

## Linear Probability Model

Den afhængige variabel kan antage enten .

Hvis MLR.4 er opfyldt er den betingede forventning lig den lineære regressionsmodel på højresiden.

Den lineære sandsynlighedsmodel ser ud som følger:

Generalt for lineære variable:

Giver os en model for respons sandsynligheden, p(x):

Da sandsynligheder summerer til en har vi, at

**Fortolkning af parametrene en LPM:**

* y er en diskret variabel
* Parameteren kan ikke blive fortolket som den marginale ændring i y givet, at ændres med én enhed.
* Parametrene giver ændring i sandsynligheden for y=1 når den forklarende variabel ændres med en enhed:

Skriv 10, 11 og 12

Hvis y er diskret er fejlledet diskret – dette følger definitorisk. u er lig minus responssandsynligheden, hvis y=0 og lig en minus responssandsynligheden, hvis y=1. Vi har altså:

Variansen af u givet x er (udledning se forelæsning 13 slide 12):

u er heteroskedastisk, da den afhænger af x. Dette bryder med MLR.5.

Skriv slide 13 ned.

**W8.1-W8.3**

# Heteroskedasticitet

Betyder forskellig spredning. Vi tillader nu, at fejlledet har forskellig varians ved forskellige observationer. Vi vil udvikle metode til når MLR.5 ikke er opfyldt.

## Definition

Den generelle definition er (tager udgangspunkt i en regressionsmodel):

Variansen varierer på tværs af observationer.

MLR.5 er antagelsen om homoskedasticitet og denne er et specialtilfælde af heteroskeasticitet:

Med MLR.5 er OLS simpel, men den mest præcise estimator.

Antagelsen om homoskedasticitet er restriktiv, da den antager variansen af fejlledet er konstant.

De uobserverbare choks til forbrugsadfærd, der kan man forvente variansen af disse varierer på tværs af indkomster. Hvis man mener forbrugschoks ikke afhænger af indkomster, kan man argumentere for, at variansen er konstant på tværs.

## Konsekvenser af heteroskedasticitet

Hvis vi antager MLR-1-4 (SLR.1-4) så er OLS-estimatoren middelret og konsistent. Det at vi dropper MLR.5 har det ikke konsekvenser for de statistiske egenskaber.

Givet Gauss-Markow antagelserne MLR.1-5, OLS er efficient og dens varians er:

Her er MLR.5 nødvendig for at udlede disse formler.

Generelt udtryk for variansen af OLS-estimatoren, hvis vi havde MLR.5 med:

Variansen af OLS-estimatoren vægter forskelligt.

Vi skal have en ny metode til at beregne variansen af estimatoren. Vi ved, at OLS ikke længere er BLUE, hvorfor det ikke længere er den bedste estimator.

Hvis fejlledende er heteroskedastiske:

* OLS er unbiased og konsistent givet MLR.1-5
  + Den normale expression for estimering af variansen af OLS estimatoren er ikke længere gyldigt.
  + T og F er ikke længere t og F fordelt, hvorfor de ikke er troværdige.
  + OLS er ikke længere BLUE.
  + OLS er ikke længere assymptotisk efficient.

Vi benytter I stedet White’s estimator af variansen af OLS-estimatoren, som er konsistent under heteroskedasticitet.

Det er muligt at få valide statistiske test uden at specificere den præcise type af heteroskedasticitet.

## Hvornår og hvor

Hvis man arbejder med data, hvor enheder er meget forskellige i størrelse, der skal man overveje om MLR.5 er opfyldt. F.eks. størrelse på lande eller firmaer.

Når data består af gennemsnit over forskellige antal observationer. Heterogene enheder med meget forskellig størrelse:

* Per capita værdier for lande
* Karaktergennemsnit mellem skoler

Ukorrekt funktionel form:

* y har måske heteroskedastiske fejlled, hvor log(y) måske ikke er.

Heteroskedasticitet er dog specifik for hver model og datasæt.

## White: Heteroskedasticity-Consistent Covariance Matrix Estimator

White har bevist at en konsistent estimator for OLS variansen er:

Udtrykket er samlet set en konsistent estimator af OLS-estimatoren.

De kvadrerede residualer er ikke den sande varians for en given observation i. Dette skyldes, at vi kun har en observation for observation i. For flere observationer får vi blot flere variansparametre.

Vi kræver en stor stikprøve og n skal gå mod uendelig. Hvis vi har en stor stikprøve er Whites standardfejl valide.

Robuste standardfejl er sensitive overfor ekstreme observationer (ingen store outliers).

Vi kan nu udregne robuste standardfejl, der er konsistente overfor arbitrær heteroskedasticitet.

Implementerer dette i STATA ved at bruge robust:

**Regress y x, robust**

### Standard overfor robuste standardfejl

Vi har nu to metoder til at beregne standardfejl – konventionelle metode fra kapitel 2 og robuste standardfejl fra kapitel 8.

Hvis der er heteroskedasticitet kan Whites estimator bruges til at beregne standardfejl .

Hvordan vil homoskedasticitet påvirke de konventionelle standardfejl?

De konventionelle standardfejl vil være positivt biased og dermed være større end de generelle standardfejl såfremt følgende gælder:

Hvis vi har varians af fejlled, der er lavere for de i observationer, der ligger langt væk fra x’s middelværdi, der vil det, der står på højresiden typisk være mindre end det, der står på venstresiden. Og omvendt.

Dvs. negativ bias, hvis det omvendte gælder.

* Hvis vi har data som en cigar er robuste standardfejl mindre en de konventionelle.
* Hvis vi har data, som en butterfly er de robuste standardfejl større end de konventionelle.
* Hvis vi har en bølge, kan de to standardfejl være ens. Der er stadig heteroskedasticitet.

Det er dermed ikke garanteret om de robuste standardfejl er større end de konventionelle. Det afhænger af, hvordan heteroskedasticiteten ser ud.

Man skal være påpasselig med robuste standardfejl for små stikprøver.

## En heteroskedastisk robust hypotesetest

Hypotesen er,

Heteroskedastsik robust t test:

t er asymptotisk standard normalfordelt.

Ikke sikkert, at ovenstående er overholdt for en lille stikprøve og man skal være varsom med inferens ved robuste standardfejl.

STATAS ’robust’ implementerer den robuste t test.

## Multiple restriktioner

Hypotesen:

R holder styr på, hvilke kombinationer af beta er vi interesserede i at lave multible hypoteser om.

Heteroskedatsik robust F test kan beregnes ved at bruge den robuste kovarians matrix:

Hvor q er antallet af restriktioner, er Whites kovarians matrix estimator og W er Wald testen. Dette er uden MLR.5.

F er asymptotisk F fordelt.

Med små n (lille stikprøve), er den robuste F test ikke nødvendigvis tæt på den begrænsede fordeling af F(q,n-k-1) – vær påpasselig.

Hvis STATA kommandoen robust bruges findes ovenstående robuste F test.

# Test af Heteroskedasticitet

Betragter regressionsmodellen:

Vi antager MLR.1-4 er opfyldt 🡪 den er velspecificeret.

Hypotesen:

Alternativ hypotese:

Denne, , svarer definitorisk til .

Hvis nulhypotesen er falsk, da er en funktion af x’erne. Afhænger variansen af fejlled af x-værdier. Dette vil bryde nulhypotesen og være i modstrid med MLR.5.

# Grafiske test

Estimerer den statistiske model vha. OLS-estimatoren. Beregner residualer Plotter de kvardererede residualer mod estimaterne af de afhængige variable.

Dette er en lidt upræcis metode.

# Breusch-Pagan test

Antager en simpel lineære sammenhæng:

Dette er den sande statistiske model.

Nulhypotesen:

Alternativhypotesen:

Homoskeadsticitet gør, at varansen af fejlled ikke afhænger af x. Det vil medføre, at nulhypotesen er opfyldt og MLR.5 er opfyldt. Hvis bare, der er brud på én vil der være tale om heteroskedasticitet.

Erstatter de sande standardfejl med OLS residualerne:

Denne hjælperegression kan estimeres. Der er multiple restriktioner, hvorfor denne kan testes med en F test eller en LM test.

Denne test implementeres som en ordinær F test eller en LM test. Dette skyldes når vi laver hypotesetest kigger vi på fordelingen af teststørrelsen under antagelse af, at nulhypotesen er opfyldt. Dvs. under antagelse af, at MLR.5 er opfyldt. Vi kan derfor bruge den ordinære F test.

For store datasæt kan F test og LM test følge den sædvanlige fordeling selv når de sande fejlled er erstattet med OLS residualer.

F test:

F-testen er assymptotisk -fordelt undeer nulhypotesen.

LM test:

LM er assymptotisk -fordelt under H0.

Note: Hvis du forventer en variabel er skyld i heteroskedasticitet, kan man teste ved brug af en simplificeret Breush-Pagan test. Her tester man altså mod specifikke brud på MLR.5.

# White’s test

Antagelsen kan erstattes med den svagere condition:

Alle kvadrerede fejlled er ukorreleret med forklarende variable og funktioner af de forklarende variable. White er dermed mere generel og tjekker for ikke linære forhold, der kan give brud på MLR.5.

Har vi en simpel model, hvor k=3:

Regressionen er da:

Jo flere k jo flere led. Her er der 9 forklarende variable.

Se nulhypotese på slide 15 s. 8.

Den simplificerede White’s test:

Når denne beregnes regner man samlet set krydsprodukter og kvadratiske led ved at se om kvadrerede residualer er en funktion af hhv. og .

# Heteroskedasticitet – kommentarer:

* MLR.1-4 sikrer at OLS er unbiased og konsistent.
* Under MLR.1-4, MLR.5 er testbare gennem test af Breush-Pagan og White.
* Hvis MLR.4 ikke er opfyldt, er disse test ikke valide og det er muligt, at forkaste nulhypotesen om homoskedasticitet selv, hvis MLR.5 er opfyldt.
* Dette medfører at det at forkaste MLR.5 kan skyldes mere generelle misspecifikationer.
* Generel misspecifikation f.eks. forkert funktionel form eller endogenitet.

# (In)efficiens

Når der er heteroskeadsticitet:

* OLS er ikke længere BLUE
* - OLS er ikke længere asymptotisk efficient.

Hvorfor?

OLS minimerer summen af de kvadrerede residualer:

Den giver samme vægt til alle residualer. Dette giver mening når der er samme varians for alle fordelinger. OLS estimatoren er inefficient når der er forskellige varianser.

En efficient estimator giver forskellige vægte, der er omvendt prportiobnale til variansen af

# WLS (weighted least squares)

Forestiller os nu en varians af u, der ikke længere er arbitrær, men en parametrisk form for variansen af fejlledet u har en baseline , der er ens for alle funktioner af x samt en multiplikativ faktor, der er en funktion af den forklarende variabel x. Den afhænger dermed af den konkrete værdi af x.

h(x) er positiv for alle værdier af x. (varianser er altid positive).

er altid en ukendt parameter.

Hvis man kender den eksakte type af heteroskedasticitet får vi mulighed for at transformere regressionsmodel sådan, at den transformerede model har homoskedastiske fejlled 🡪 dvs. får os tilbage til en model, hvor MLR.5 er opfyldt.

Formelt:

Når h er en kendt funktion kan den beregnes for hver individuel observation: .

Fejlledet i den transformerede model har følgende karakteristika:

1. Den forventede middelværdi er 0
2. Den forventede varians er konstant

I STATA skal vi implementere en regressionsmodel, hvor der ikke beregnes et konstantled.

Parametrene er de samme som i den oprindelige model!

Vi har mulighed for at estimere efficient, da MLR.1-5 er opfyldt i den transformerede model.

”Weighet least squares” giver forskellig vægt til forskellige observationer. Meget vægt til observationer med lav varians og omvendt.

WLS er bare OLS regression på den transformerede model, der minimerer den vægtede sum af kvadrerede residualer.

WLS hører under GLS. Hvis h-funktionen ikke er kendt bruger vi Feasible GLS.

# Linear Probability Model (Lineære sandsynlighedsmodel)

Se slide 15 s. 18

# Feasible GLS (FGLS)

I mange tilfælde er den eksakte type heteroskedasticitet ukendt, men h kan være modeleret og efterfølgende estimeret (h er ukendt).

Vi vælger en fleksibel funktionel form for h og estinerer den. En måde at gøre det på er, at antage at variansen af u har følgende funktionelle form:

Variansen er proportionel med . Dette mefører forskellige x-værdier har forskellige varianser af deres uobserverbare fejlled.

Næste step er at erstatte de ukendte ’er med estimater. Bruger sædvanlige værktøjer som fra WLS.

Dette er et forsøg på at lave en approksimation.

## Procedure for FGLS (kogebog)

1. Estimer den originale model ved brug af OLS
2. Beregn OLS residualerne og konstruer . Da logaritmer giver en lineær regressionsmodel.
3. Estimer hjælperegressionen:
4. Beregn de forventede værdier af fra regressionen i step 3.
5. Beregn – dvs. approksimeret sin h-funktion.
6. Estimer modellen:

Ved brug af WLS med vægtene .

Ved at implementere FGLS er vores standardfejl valide, da de vil opfylde MLR.5. Forskellen er, at FGLS ikke er gyldig i små stikprøver – kun asymptotisk valid. Dette skyldes, at vi approksimerer h.

## Konklusion

Konklusionen er, at en efficient estimator vægter korrekt. Derfor er WLS BLUE ved heteroskedastiske fejlled. Når vi laver transformation er det med det formål at lave en valide parameterestimater.

Vi undersøger om homoskeastisk eller heteroskedastisk fejlled ud fra enten Breusch-Pagan eller Whites test.

Hvis heteroskedasticitet er fundet kan forskellige rettelser gøres

* ”Den moderne tilgang”: Bruge OLS med robuste standardfejl, der er valide i store stikprøver.
* ”Den gamle tilgang”: Antag en specifik form for og benyt denne til at transformere regressionsmodellen sådan, at de transformerede fejlled er homoskedastiske:
  + WLS: h(x) er en kendt funktion. Mindre gennemsigtighed af resultaterne grundet frihedsgraden fra valget af h. Svært at vurdere troværdigheden.
  + FGLS: h(x) er en ukendt funktion og estimeres. Denne fungerer i store stikprøver.

Hvis man har en lille stikprøve og er bekymret for brud på MLR.5 så er den eneste mulighed WLS.

**W9.1-2**

# Misspecifikation af den funktionelle form

Når vi antager en fejlagtig funktionel form:

* OLS vil generelt set være biased og ikke være konsistent
* Fejlspecificeret funktionel form er ens med udeladte variable. Vi har udeladt led og dermed indgå i fejlledet, Hvis der er ting i fejlledet, der er korreleret med x-variable, så er OLS estimator biased.

Økonomisk teori giver os mange muligheder for funktionelle former, men den giver næsten aldrig præcise instruktioner for, hvordan vi stiller den rigtige empiriske model op.

Eksempler på misspecificerede funktionelle former:

* Den afhængige variable har den forkerte funktionelle form
* Forklarende variable har forkerte funktionelle former

Silver lining: De er muligt at løse problemet så længe vi observerer de relevante variable. Vi kan dernæst prøve os frem os se, hvilken, der giver korrekte resultater.

Brud på MLR.4 er vi mest nervøse overfor udeladte variable vi ikke kan observere, der vil give bias.

Forket funktionel for er anset som værende mindre seriøst end udeladte variable (som typisk ikke kan observeres), da denne fejl kan løses og den korrekte model kan implementeres.

Hvis vi har angivet et simpelt lineært forhold mellem y og x-variable. Denne funktionelle misspecifikation kan tænkes som værende en simpel lineær approksimation til en ikke lineær model. Denne approximation er god for små x-værdier.

## Statistisk procedure:

* Estimer modellen med OLS
* Beregn residualer
* Plot residualer mod de forklarende variable
* Beregn formel test.

## Formel test for misspecifikation:

1. Antag en regressionsmodel, hvor det antages at MLR.1-4 er opfyldt.
2. Hvis et kvadratisk led tillægges skulle den blive insignifikant.
3. Det er generelt muligt at approksimere en ukendt funktionel form med polynomier.
4. En simpel test for misspecifikation er:
   1. Tillæg en kvadratisk led for den forklarende variabel
   2. Test samlet signifikans dvs. en (robust) F-test.
   3. Hvis ikke signifikant vil den originale model specifikation ses som værende reasonable.

## RESET – REgression Specification Error test

Man tester om ikke lineære funktioner af x hare n forklaringsgrad. Hvis denne har vil den oprindelige funktionelle form ikke fange det fulde billede. Den vil fortælle om modellen er signifikant. Den guider ikke i retning af, hvilken vej der er korrekt, hvis den funktionelle form ikke var signifikant.

1. Antager en regressionsmodel, der opfylder MLR.1-4
2. RESET tillægger et polynomium u de forventede værdier for y
3. Testen af den korrekte funktionelle form kan ses som en nulhypotese:
4. Test statistikken er approksimativt F fordelt med (2,n-k-3).

Tester over et generelt tilfælde, der indeholder den rigtige model.

## Test af ikke nestede alternativer

Hvordan tester man specifikationer, der ikke nødvendigvis er nestede?

Nestede modeller:

Ikke nestede modeller:

Vi forsøger at lave en stor samlet model og finde om det er komponenterne fra den ene eller den anden model, der har størst forklaringsgrad.

Tilgang 2: Estimer den udvidede model:

Test de følgende to hypoteser:

* Model 1:
* Model 2:

Tilgang 3: Kigger på en hjælperegression, hvor man medtager den prædikterede y cvariabel fra model 2, hvor x1 og x2 er logaritmetransformationer.

Hypotesen:

Opstiller en hjælperegression på naggrund af model 2. Undersøger om prædiktionen fra x1 har nogen forklaringsgrad i hjælperegressionen baseret på model 2.

Hypotesen:

### Konklusioner for ikke nestede alternativer

Vi kan ende i konklusioner om, hvilke modeller vi skal foretrække frem for hinanden er cirkulære eller begge modeller afvises. Hvis begge afvises må man forsøge sig med andre kriterier. Der kan altid være en tredje mulighed, der er den sande funktionelle form.

# Proxy variable

Variable, der er interessante for modellen, men som er uobserverbare. I stedet kan man medtage proxyer for de interessante variable, der måske bør indgå i modellen.

Den sande model:

er ikke observerbar, men er.

Proxy variable erstatter de sande uobserverbare variable.

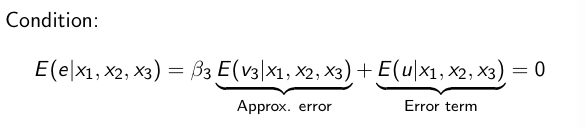
Idéen er at minimere problemet med udeladt variabel bias.

Proxy-variable markeres med en stjerne, så man eksplicit kan vise at variablen er uobserverbar. Vi antager at den sande x\* har en lineær form ift. vores proxyvariabel x:

Hvor er en standardfejl, der indikerer, at vores proxyvariabel ikke forklarer hele variationen i den sande x-værdi.

Vi er interesseret i at estimere partielle effekter, hvor vi ved hjælp af proxyvariable forsøger at minimere bias fra et evt. uobserverbart x.

Den alternative model er:



Når man bruger proxyvariable og man skal argumentere for, at den betingede forventning af det sammensatte fejlled skal være lig 0, skal man for hver af det to forventninger redegøre for, hvorfor de individuelt er lig nul.

Man skal overholde de ovenstående to betingelser. Alle andre faktorer, der ligger i u er ikek systematisk relateret til de tre forklarende variable. Hvis dette er gældende er error term lig nul.

Hvis approksimationsfejlen, , er ukorreleret med de tre x-variable og der derfor ingen systematik er, da er den approksimative fejl lig nul. Hermed bliver hele den forventede middelværdi lig nul.

De ting, der forsat er i fejlledet kan stadig være korreleret med fejlledet.

**W9.3-6 (L17)**

**W15.1 + sektion 1-2 i IV-note (L18)**

# Instrumental variabel (IV) estimation

Endogene forklarende variable

* Defineret som stokastiske variable, hvor .
* OLS er (asymptotisk) biased når x er endogen.
* Vi har altså, at den forventede værdi af fejlledet er forskellig fra nul.
* Hvordan kan endogenitet forekomme?
  + Udeladte forklarende variable (tit x-variable vi gerne vil medtage, der er uobserverbare, som derfor er indeholdt i fejlledet)
  + Målefejl i den forklarende variabel
  + Omvendt kausalitet

## Korrelation versus kausalitet

Vi kan ofte på baggrund af data finde variable, der er **korrelerede**, uden der nødvendigvis er en **kausal** sammenhæng. Korrelation er en statistisk sammenhæng, men denne siger ikke noget om en årsagssammenhæng.

Det er **sjældent** muligt at tillægge OLS-estimatet, , en kausal fortolkning, dvs. at en ændring i x medfører, at y ændres med , alt andet lige. Dette skyldes, at vi er nervøse for, at der er udeladte variable indeholdt i fejlledet, der er systematisk med x.

beskriver en korrelation, men medfører ikke nødvendigvis kausalitet.

## Instrumental variable (IVs)

Antager, at vi har en simpel regressionsmodel:

x mistænkes for at være endogen,

OLS er (asymptotisk) biased.

IV-tilgang: Find en instrumental variabel z som opfylder:

Udfordringen er at finde et hensigtsmæssigt instrument.

Vanskeligt i praksis med ikke-eksperimentelle data. Dvs. ikke et universelt værktøj, der kan bruges når OLS ikke kan. Der er stærke antagelser, hvorfor den ikke altid kan bruges i praksis.

Vi kan komme uden om endogenitet vha. IV-estimation.

De to betingelser for instrumentet er fundamentalt forskellige:

* Betingelse 1:
  + Instrumentet skal være korreleret med uobserverbare faktorer indeholdt i u
  + Den antagelse er baseret på teori og kan **ikke** testes, da vi aldrig kan observere u. Hvis vi skal retfærdiggøre, hvorfor dette er tilfældet og derfor skal man bruge gode argumenterer fra økonomisk teori.
* Betingelse 2:
  + Instrumentet skal korrelerer med den endogene variabel.
  + Den antagelse kan **testes** med data på z og x. Vi vil vise at kovariansen er både statistisk og økonomisk signifikant.

## IV-estimatioon for den simple regression

Vi har en simpel regressionsmodel:

x er endogen, hvorfor .

z er et relevant (henviser til betingelse 2.) og validt/gyldigt (står ortogonalt på det oprindelige fejlled) instrument:

IV-estimatoren af kan udledes vha. momentestimatoren:

Kovariansen mellem en stokastisk variabel og en skalar er definitorisk lig nul, hvorfor

.

Ovenstående formel dikterer, at den sande datagenererende proces for y giver en sand beta1, der er forholdet mellem de to kovarianser. Vi skal bruge vores relevante eksperiment, da vi skal sikre, at vores nævner er forskellig fra nul.

Vi har en simpel regressionsmodel:

**Population momenter:**

Relevante og valide instrumenter:

er **idenificeret** ved

**Data momenter:**

IV-estimatoren udledes derfor som ved at erstatte med empiriske momenter:

er den empiriske kovarians for z og y.

er den empiriske kovarians for z og x.

## IV-estimatorens egenskaber

* IV-estimatoren er **konsistent**, dvs. . Vores estimator konverger i sandsynlighed mod den sande værdi.
* IV-estimatoren er asymptotisk normalfordelt
* IV-estimatoren har pæne asymptotiske egenskaber, men der er følgende forbehold:
  + IV-estimatoren er ikke middelret (ikke god i små stikprøver). IV-estimatoren er ikke middelret, da middelrethed kræver, at den betingede forventning af fejleledet u givet x er lig nul. Dette kræver nul korrelation mellem ikke lineære funktioner af x ikke afhænger af z. Kig på MLR.1-4.
  + IV-estimatoren har ofte større varians end OLS.
  + ”Svage instrumenter, , kan føre til stor bias. Dvs. instrumenter, der har en svag forklaringsgrad, giver anledning til bias. Vi vil altså vise, at der er en økonomisk signifikans af variablen.
* Hvis x er eksogen, kan x bruges som instrument for x

OLS er et **specialtilfælde** af IV-estimatoren.

## Inferens med IV-estimatoren

Antager homoskedastiske fejlled:

Betinger på z, da det er alle eksogene variable.

Den asymptotiske varians er givet:

den kvadrerede korrelationskoefficient mellem x og z. Jo højere jo mindre varians på IV-estimatren. Svage instrumenter er forbundet med stor varians på IV-estimatoren. Dette er udtrykt gennem .

Variansen er konsistent estimeret ved:

NB:

OLS estimatren for den simple lineære regressionsmodel er mindre end IV-estimatoren. IV-estimatoren er forbundet med en større varians. Ved en perfekt korrelation mellem x og z, må de være samme variable og vi kan ikke bruge x variablen som instrument for sig selv, hvorfor OLS og IV variansen bliver ens.

t-test er asymptotisk normalfordelt, etc.

## OLS versus IV-estimation

* OLS estimator:

Størrelsen på den samlede bias er . Vi ønsker mere variation i x, da dette giver mindre bias.

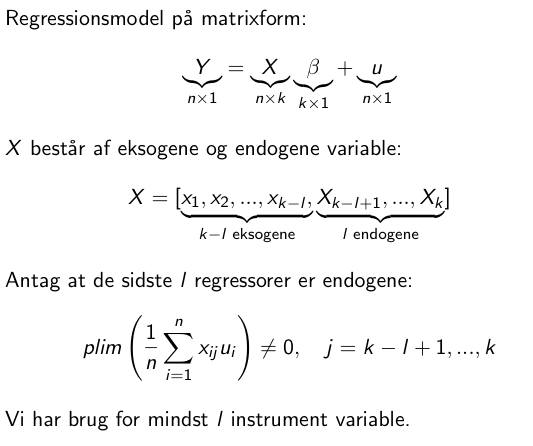
Konvergerer i ssh. Og IV-estimatoren er konsistent. Vi kan have en asymptotisk bias udtryk ved brøken . Vi ønsker stærke instrumenter, da alt andet lige vil gøre stor og bias er dermed lille.

Man vælger ud fra at finde ud af, hvilken bias man tror er værst. Begge brøker kan ikke findes. Man kan også implementere begge og være kritisk.

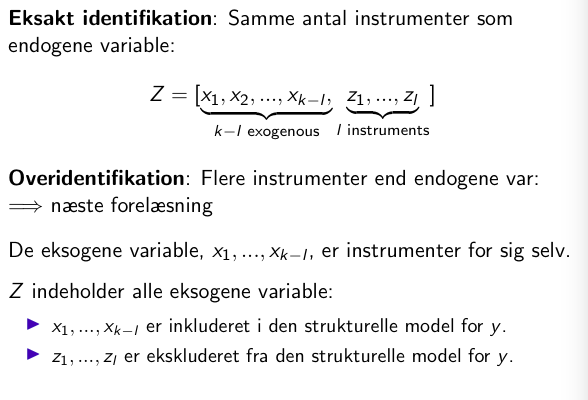
**W15.2-3, 15.5 + sektion 3 i IV-note – L19**

# Eksakt identifikation med multiple regressorer

## Multipel lineære regressionsmodel med endogene variable



Antager, at de er endogene, hvilket betyder at kovariansen mellem de forklarende variable og fejlledet af u er asymptotisk forskellig fra nul.



## IV antagelser

IV-estimatoren kan anvendes, hvis følgende antagelser gælder:

1. Instrumenterne er ukorreleret med fejlleddet:
2. Instrumenterne er korreleret med de endogene variable:

Minder om MLR.3.

1. Ingen perfekt multikolinaritet mellem instrumenterne:

Direkte pendant til MLR.3.

Disse betingelser illustrerer, hvorfor vores diskussion fra sidste gang kun gælder for den simple model. Da hvis vi har en model:

IV:

Må Nej, da instrumentetet er en regressor i sig selv. Dette er den fejl man kan få ved kun at kigge på de simple matricer fra sidste gang.

## IV-estimatoren i det eksakt identificerede tilfælde

Vi vil gerne udlede IV-estimatoren i det gennerelle tilfælde ud fra method of moments (MM). Tag udgangspunkt i teoretiske momenter

MM: Erstat teoretiske momenter med datamomenter:

IV-estimatoren i det eksakt identificerede tilfælde. OLS estimatoren er et specialtilfælde af IV-estimatoren.

Hvis X er eksogen:

IV-estimatoren er konsistent:

**Beviset for konsistens af IV-estimatoren:**

Plim til produktet af to stokastiske variable er produktet af deres to grænseværdier.

**Opsummering:**

Antallet af endogene variable = antallet af instrumenter

STATA’s ivregress-kommandor implementerer denne formel.

Hvis er endogen og instrumentet z kan bruges, skriver man:

**Ivregress 2sls y x1 (x2=z) x3**

Dette er ikke en statistisk model

2SLS kan generaliseres til det overidentificerede tilfælde, hvor: antallet af endogene variable er mindre end antallet af instrumenter.

# Two-stage least squares (2SLS): Det simple tilfælde

Simpel model:

x antages at være endogen:

Z er et relevant og validt instrument:

Trin 1: Estimér med OLS:

Trin 2: Estimér med OLS:

Vi har, at og begge er konststene estimatorer under samme antagelser.

Intuition:

* I Trin 1 opdeles x i en ”god” og en ”dårlig” del:
* Da gælder der, at
* Prædikteret er eksogen og bruges til at estimere

Skriv resten af slide 11!

**Trin 1:** Regresser de k variable i X på de k variable i Z:

De k hjælperegressioner kaldes ”first stage” regressionerne. Bruges KUN på endogene variable. E er fejlledet.

NB: Regressionerne for de eksogene variable er ”trivielle”. I den forstand, at hvis x er eksogen og vi kører en regression på alle variable af Z-matricen får vi en værdi 1 foran variablen selv. Vi bruger kun ”first stage” på endogene variable.

Beregn OLS-estimater og :

Hvor er en projektionsmatrix som er symmetrisk og idempotent .

**Trin 2:**  Regresser y på

Indsæt i OLS-formlen:

Indsæt udtrykket for

Brug regnereglen (ABC)’=C’B’A’ på første parentes:

Benyt regnereglen :

Ved at fjerne led får vi:

Dette udtryk er identisk med IV-estimatoren.

Bemærk at logning (1) også gælder i det overidentificerede tilfælde.

## Hvornår er IV-estimatoren den foretrukne estimator?

* OLS versus IV estimation:
  + Hvis x er **endogen**: IV konstistent, OLS biased
  + Hvis x er **eksogen**: Begge er konstistente, OLS er efficient, da OLS er assymptotisk efficient.
* Man bør altid sammeligne sine OLS og IV-estimater.
* Store forskelle kan være tegn på, at OLS er biased, hvorfor IV foretrækkes.
* Små forskelle kan være tegn på, at OLS ikke er biased og dermed den foretrukne estimator.
* Kan suppleres med et formelt test af eksogenitet (F20)

## Vigtige pointer fra IV-estimatoren

IV.estimator i det eksakt identificerede tilfælde:

Med mange endogene variable er det vigtigt at have mange ”stærke” instrumenter som har så lille indbyrdes korrelation som mulig. Dette skyldes, hvis de har høj indbyrdes variation, så kan vi rande ind i multikolinaritetsproblematik og det er svært at finde den partielle effekt.

IV-estimator kan implementeres ved 2SLS (=2x OLS)

**W15.4-15.6**

# Overidentifikation med multiple regressorer

Simpel regressionsmodel:

To relevante og valide instrumenter:

F.eks. mor og fars uddannelsesvalg. Hvis vi f.eks. estimerer afkast af uddannelse.

To mulige IV-estimatorer:

Vi kan kun anvende et instrument af gangen når vi har en endogen variabel.

## Sammenligning af OLS og IV

Når vi benytter IV-estimatoren for vi en lavere R^2 værdi end for OLS-estimatoren. Dette er dog ikke overraskende, at vi ser en lavere værdi, da OLS-estimatoren er udledt som en momentestimator, men kan også komme fra Maksimum Likelihood eller som et minimeringsproblem (den bedste rette linje, hvor vi minimerer vandrette afstand fra linjen til datapunkt). Da OLS minimerer summen af de kvadrerede residualer maksimerer den pr. definition R^2. IV-estimatoren maksimerer derimod ikke R^2, men opfylder momentbetingelser. Dermed har de to estimationsværktøjer to forskellige mål. IV sikrer Z-matricen står ortogonalt på vores residualer. Den forsøger dermed at bruge de rigtige momentbetingelser til at regne et estimat, der er konsistent.

## Two-stage Least Sqares (2SLS)

Eksakt identification (=) eller overidentification (>):

* Flere instrumenter end overidentificerede variable.

Trin 1: Beregn ud fra de eksogene variable i Z (Se eksempel på hjælperegression i slide):

Trin 2: Regresser :

Vi kan ikke benytte regneregel fra tidligere, da X og Z har forskellige dimensioner.

## Inferens

Ligner metode fra OLS. Variansen af IV-estimatoren med homoskedastiske fejl (MLR.5 er opfyldt):

Hvor og er IV residualen.

Variansen af IV-estimatoren med heteroskedastiske fejl (Dette er svarende til White’s formel):

**Ivregress** kan kombineres med **robust**-optionen for at implementere standardfejl som er robuste overfor heteroskedasticitet. Bruger kun ovenstående når n er stort. Kan ikke bruge 2SLS, da OLS standardfejl ignorerer at er en beregnet størrelse og ignorerer trin 1.

### Konklusion om overidentificeret IV-estimator

Den overidentificerede IV-estimator er mere efficient, da den har en lavere varians. Overidentifikation er relevant, da hvis vi har to stærke relevante og valide instrumenter har vores estimator en lavere varians og vores hældningsparameter er mere præcis.

Hvis vi skruer op for den indbyrdes korrelation mellem de to instrumenter og i grænsen, hvor de er perfekt korreleret får vi multikolinaritet og vi vil ikke længere foretrække den overidentificerede IV-estimator.

## Test af eksogenitet

Muligt at teste om x er endogen, når et relevant og validt instrument findes.

Test tager udgangspunkt i 2SLS-proceduren fra sidste gang:

Derfor, hvis:

Vi opstiller en lineær regressionsmodel:

instrumenter er til rådighed:

**Trin 1:**

**1. trins regressioner** for de endogene regressorer der estimeres ved OLS:

Residualer fra hjælperegression:

**Trin 2:**

Indsæt residualer i hovedmodel og opstiller dermed den udvidede model:

Test (anvend T-test for l=1 eller F-test for l>1) og , dvs. postulerer

* Hvis residualerne er signifikane, er en eller flere regressorer endogene.

Bemærk, at og ikke indgår i modellen.

Eksempel: Lønregression

* endogen (uddannelse)
* instrumenter (mors og fars uddannelse)

Konklusion: På baggrund af denne stikprøve kan uddannelse antages at være en eksogen variabel, hvor vi vil foretrække at bruge OLS-estimatoren, da den er mere efficient end IV i dette tilfælde. Men med denne stikprøves test er vi ikke langt fra at konkludere det modsatte, hvilket er hvad ivi ser i andre datasæt.

Med flere instrumenter end endogene variable er modellen **overidentificeret**.

* Dette giver mulighed for at teste restriktioner.
* Denne muligheder ikke til stede for eksakt identificerede modeller ().

Intuition bag overidentifikationstest (flere instrumenter end vi har brug for):

* Brug de første instrumenter til at beregne IV-residualer, .
* Undersøg om er korreleret med de g resterende instrumenter
* Hvis instrumenterne er valide, bør der ikke være en korrelation til de resterende instrumenter.

Hjælperegression:

. Interesseret i om der er 0 korrelation melle IV-estimator og resterede instrumenter. Dvs. ingen korrelation mellem Z og u.

Teststørrelsen er som følger

Eksempel: Lønregression

* c=3,84

Konklusion: Meget lav og en testværdi på 0,4. Vi har et instrument i overskud, hvorfor vi skal finde d. 95-fraktilen i en -fordeling. Her kan vi ikke afvise , da vi ikke er i den kritiske region. Dvs. mindst et instrument er eksogent.

**Eksakt identifikation** vil OLS-estimater være perfekt ukorreleret og vil derfor være eksakt lig nul. Se slide for udregning.

## 6-trins procedure for IV-estimation

1. Specificer relevant model:

Bestem, hvilke forklarende variable som potentielt er endogene (l variable).

1. Skriv færdig ud fra slide.

**W11.1 + HBN sektion 1-3**

# Stationære tidsserier

## Datatyper: Tværsnitsdata vs. Tidsserier

**Tværsnitsdata:** Observationer for mange forskellige enheder (individer, firmaer, kommuner, lande) på et givent tidspunkt.

**Tidsserier (tidsrækker)** er data, som er målt på en række forskellige tidspunkter (typisk med et fast interval). Her er der følgende empiriske eksempler.

* Årlige data (ledighed, BNP)
* Kvartalsdata (priser)
* Månedsdata (antallet af biler)
* Minutdata (aktiekurser)

Tidsserier er kendetegnet ved en naturlig ordning af observationerne (i modsætning til tværsnitsdata). En observation i dag kommer efter den i dag og før den i morgen – der er temporal sammenhæng i tidsseriedata.

## Tidsserier

Vi vil undersøge om OLS er et relevant værktøj i forhold til tidsseriemodeller. Hvornår giver OLS et korrekt svar, og hvornår kan dette værktøj ikke bruges.

En observation skrives som , hvor t refererer til det tidspunkt variablen er målt.

Tidsserien er givet ved:

Hvor er det første tidspunkt, vi har målt på, og er det sidste tidspunkt. En tidsserie er en række af observationer. Dvs. t observationer for t perioder. T måler sidste observation i vores data.

I vores data er der observationer.

Datasættet kan også angives som .

Vi gennemgår her kun eksempler, hvor intervallet mellem observationerne er konstant (f.eks. en måned). Vi skifter ikke mellem intervaller!

Observationerne er som regel **afhængige over tid**.

* En observation i en given periode er relateret til fremtidige og historiske observationer.

To observationer tæt på hinanden er ofte korrelerede (en høj observation følges ofte af en høj observation): .

* F.eks. arbejdsløsheden ændrer sig langsomt fra år til år
* Hvis man skal forudsige arbejdsløshedsprocenten et år frem, vil det derfor være en god ide at tage udgangspunkt i dette års arbejdsløshedsprocent.

Fordi vi har en naturlig ordning af data, kan vi tale om **prædeterminerede variable**, . Dvs. opfattes som kendt på tidspunktet t. De er bestemt før den periode vi kigger på. Dette er relevant for at modellerer, hvad der kommer til at ske fremadrettet.

## Stokastisk proces

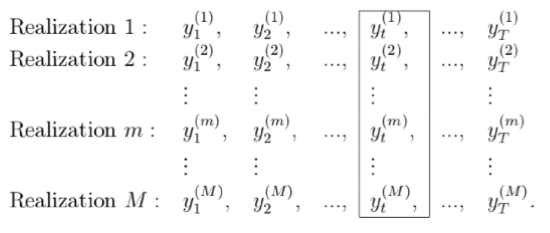
Når vi observerer en tidsserie, kan vi tænke på den som **en** realisation af en stokastisk proces.

* Vi har kun én stikprøve, men i en tidsseriekontekst er det mere abstrakt at tænke, hvordan en anden stikprøve vil se ud relativt til tværsnitsdata.
* En anden realisering fra vores stokastiske proces betyder, at vi skal trække historien endnu engang, og denne vil se meget forskellig ud fra det vi har i dag.

Vi har kun en realisation – og det er ikke så let at forestille sig, hvordan en anden realisation ser ud.

Dette er forskelligt fra tværsnitsdata. Her er det lettere at forestille sig, at vi udtager en anden stikprøve.

Vi forestiller os, at vi har M realisationer:



Vi vil gerne kende middelværdien

I princippet kan den udregnes som:

Men vi har kun en realisation. Vi observerer ikke store M gange, men kun én gang. Det er svært at lave udsagn når vi kun har én observation. Vi skal derfor ligge restriktive antagelser.

## Stationaritet

Vi kan ikke lave inferens, hvis vi kun har en realisation til hvert tidspunkt.

… med mindre fordelingen for er den samme.

Hvis observationerne er trukket fra samme fordeling, kalder vi det **stationaritet**.

Antager, at de underliggende fordelinger er de samme og vi har nu T observationer til at forklare den underliggende fordeling.

Definition 1: **Streng stationaritet**.

* En tidsrække er strengt stationær, hvis den fælles fordeling af s stokastiske variable er den samme.

Definition 2: **Svagt stationaritet**

En tidsrække er svagt stationær, hvos der gælder at

For alle t.

Det er afstanden mellem de to observationer, der betyder noget for .

## Tidsafhængighed

For tværsnitsdata antog vi, at fejlleddene var uafhængige og identiske fordelte (IID).

Derfor kan ”simple” versioner af Store Tals Lov og den Centrale Grænseværdisætning anvendes til at beskrive OLS-estimatorens asymptotiske egenskaber.

Med tidsrækker er det urealistisk at antage IID observationer. Dette er vores tilfældige stikprøve med homoskedastiske fejlled.

Vi kan erstatte ”identisk fordelt” med ”stationaritet” samt erstatte ”uafhængighed” med ”svag afhængighed”.

Defintion 3: **Svag afhængighed**

For en tidsrække er svagt afhængige, hvis og er approksimativt uafhængig, hvis .

Når vi har stationaritet og svag afhængighed gælder der, at

er en konsistent estimator for . Store T observationer, der kan bruges til f.eks. at beregne det empiriske gennemsnit og denne estimator vil være persistent med den sande middelværdi.

Hvordan måler man tidsafhængighed mellem observationer?

Autokorrelation:

Hvis der er en **høj**  grad af **tidsafhængighed**, kalder vi tidsserien for **persistent**. Dette er en måde at udtrykke, at tidsserien ikke er stationær.

Hvis autokorrelationen er positiv (negativ), vil store værdier af vores stokastiske variabel, , blive fulgt af store (små) værdier af .

Hvis processen er stationær, kan autokorrelationen også skirves som:

h fanger hvad er afstanden mellem de to observationer vi har i tankerne.

Estimatoren for autokorrelationsfunktionen er givet ved:

Autokorrelationsfunktionen kan også bestemmes ved OLS estimation i følgende regressiosmodel:

Hvis den sande korrelation så gælder, at den asymptotiske fordeling   er normalfordelt med varians:

## Statistisk tidsseriemodel (regressionsmodel):

Model hvor vi vil forklare variation i vores ud fra de k forklarende variable:

Denne er statisk, da der kun er en umiddelbar effekt fra en forklarende variabel i periode t i den periode. Dette er ikke realistisk, hvis en variabel i en given periode afhænger af andre dynamiske specifikationer.

Hvis har en **stationær fordeling** og **er svagt afhængigt**, kan OLS anvendes. Dette SKAL antages før end, at OLS kan anvendes til at estimere parameterestimater og give dem en evt. kausal fortolkning.

**Stationaritet:** (streng og svag). De fælles fordelinger er de samme.

**Svag uafhængighed:** i stedet for uafhængighed (i tværsnit).

Mange økonomiske tidsserier er dog kendetegnet ved ikke-stationaritet og høj tidsafhængighed pga. trends, niveauskift, etc.

Mange økonomiske tidsserier er dog kendetegnet ved ikke-stationaritet og høj tidsafhængighed pga. trends, niveauskift, etc.

**W10 + W11.2-11.5 + HBN sektion 4-6**

# Ikke stationære tidsserier

Meget persistente tidsserier er ofte ikke-stationære. Test for stationaritet introduceres i Økonometri II.

## Detrending

Trends medfører, at middelværdien ikke er konstant.

Dette er ikke foreneligt med definitionen af stationaritet.

En løsning: Anvend **afvigelser fra trenden**.

I praksis tager man sin model:

Parametrene kan beregnes vha. OLS. Residualer fanger afvigelser.

Afvigelser fra trenden beregnes som:

Er den transformerede tidsserie stationær og svagt afhængig?

## Differender

En anden ofte anvendt løsning: Anvend **differenser**

Differenser beregnes som to efterfølgende perioder. Giver procentvis ændring fra år til år:

Variablen er integreret af 1. orden, hvis er ikke-stationære, hvorimod er stationær.

* I dette tilfælde betegnes som en -variabel (integreret af 1. orden).
* Stationære tidsserier betegnes som -variable (stationære).

Når variable er ikke stationære er de -variabel og hvis de er stationære er de -variable.

Er den tranformerede tidsserie stationær og svagt afhængig?

Stationær og svagt afhængig. Fjerner bekymring om ikke stationaritet.

## Kointegration

Tidssrierne for forbrug og indkomst ser ud til, at ”følges ad”.

En tredje metode: Anvend en **lineær kombination** af flere ikke-stationære tidsrækker.

Hvis den lineære transformation er stationær, viser det, at variablene **kointegererer.**

Er opsparingsraten, , en stationær tidsserie?

Dette er en simpel relation, hvis man tager . Det er den måske, afhængigheden aftager over tid. Se slide!

Hvis ja, er der tegn på en kointegrationsrelation ml. og .

Kointegrationsanalyse behandles i Adv. Macroeconometrics.

## Statistisk model

Variablene er målt på det samme tidspunkt. Modellerer forhold mellem y variabel i periode t for at måle den statiske sammenhæng, hvordan afhænger y i periode t af forklarende faktorer x i samme periode. Der er ingen dynamik:

Parameteren angiver effekten af på .

indeholder alle de forklarende variabel, og er en vektor for alle forklarende variable i periode t.

Vi antager vi har et pænt fejlled, dvs. fejlled har ingen samvariation med x, hvorfor vi ved at en enhedsændring fører til en kausal effekt på y der er fanget af betakoefficienten.

## Autoregressiv model

Den simple autoregressive model er givet ved:

Den afhængige variabel afhænger kun af sig selv. Pametren udtrykker, hvor stor tidsafhængighed, der er for y. Hvis er y kun lig vores fejlled. Hvis de er meget høj ses det intuitivt, at y i periode t afhænger meget af y i tidligere perioder.

Modellen kaldes en første ordens autoregressiv model, **AR(1).**

Generel **AR(P)**-model:

## Autoregressive distributed lag model

Eksempel på ADL-model:

Dynamisk modeæ, hvor historiske variationer på afhængige og forklaret variabel indgår. Dvs. model med laggede variable af y og x-variable.

Modellen er god til at beskrive processer med dynamiske tilpasninger, og kan udvides med fæere æags.

ADL-model kan skelne ml. effekter på kort og lang sigt.

Hvidan påvirkes y, når øges med én enhed?

Er nødt til at kigge på og etc:

Hvis , har kun en **transitiv** effekt på :

En permanent ændring i har en permanent effekt på .

Den forventede værdi af

Hvis og er stationære, gælder , etc.

Derfor er langsigtseffekten af en permanent ændring i :

Langsigtseffekten er blot summen af alle kortsigtseffekter.

## Antagelser for den lineære regressionsmodel:

**Antagelse 1:** Modellen skal være lineær i parametrene. (Svarer til MLR.1)

**Antagelse 2:** (Svarer til MLR.4)

Udelukker en systematisk sammenhæng mellem fejlled og forklarende variable. Udelukker f.eks. udeladte variable og omvendt kausalitet. Verører kun størrelser i periode t.

**(ingen feedback fra**  til eller udeladte variable)

**Antagelse 3:** Ingen perfekt multikolinaritet i . Forklarende variable, der har variation.

**Antagelse 4:**  har en stationær fordeling og er svagt afhængig.

**Antagelse 5:** Stren eksogenitet:

## OLS-estimatoren:

Momenestimator:

Under antagelse er 1-4 OLS-estimatoren **konsistent.** Assymptotisk konvergerer estimatoren mod sin sande værdi.

Husk en andne version af store tals lov pga. tidsserier.

Under antagelse 1-5 er OLS-estimatoren **middelret:**

## Middelrethed:

Strengt eksogenitet er sjældent opfyldt for tidsserier.

Hvis modellen indeholder den laggede afhængige variabel, vil OLS som regel ikke være middelret.

Eksempel: Autoregressiv model (AR(1)):

Antagelse 5 (streng eksogenitet) er ikke opfyldt, fordi

Simulationseksperiment:

Simulerer en tidsserier AR(1)-model, hvor man kan beregne OLS estimator, der kan findes flere gange. Givet specifikationen er den sande , og vi ser på slide det giver en skæv fordeling med en venstre hale, der trækker gennemsnittet ned og giver negativ bias. Vi underestimerer den sande for små stikprøver. Viser OLS-estimatoren ikke er middelret.

## OLS-estimatoren asymptotiske fordeling

Hvis antagelserne 1-4 er opfykdt, og fejlleddet er:

* **Homoskedastsik:**
* **Serielt ukorreleret:**  for

Er OLS-estimatoren asymptotisk normalfordelt:

Hvis vi erstatter variansen med den estimerede varians, får vi:

Hvor vi simmerer over tidsperioder. Antal tidsperioder i vores data skal være stor før end vi kan bruge de assymptotiske approksimationer.

Hvis der er **heteroskedasticitet,** anvendes den **robuste**  estimator af variansen:

hvor .

Dette er det samme som tværsnitsdata.

## Autokorrelation i fejlleddet:

Serielt ukorreleret:

Alternativt: Antag at modellen er dynamisk **komplet**. Hvor vi har taget højde for alt relevant historisk information - bør ikke være noget autokorreleret i fejlledet:

Tidligere værdier af y og x indeholder ikke mere information, end det der er i .

Eksempel på autokorrelret fejlled:

Hvis jeg har autokorrelerede fejllled, så er den korrete secifikation som funktion af den laggede y, og og den laggede forklarende variabel samt et fejlled , der ikke er autokorreleret.

Modellen kan i setdet formuleres som:

Hvilke konsekvenser har **autokorrelation** i fejlledet?

* OLS er generelt stadig **konsistent**.
* OLS er generelt stadig **asymptotisk normalfordelt**.
* Variansen skal korrigeres (jf. W8)
* Der findes estimationer af kovarianser, som er robuste overfor bååde heteroskedasticitet og seriel korrelation (HAC).
* I STATA kan man bruge: **regress VCe(hac3) y x, robust**

## Vigtige pointer:

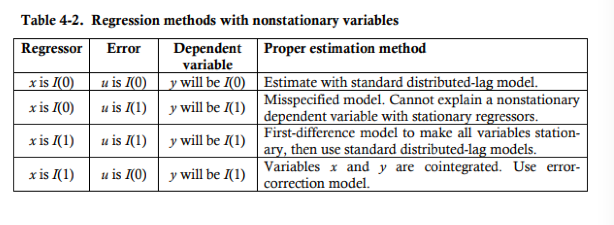
Tidsserier kan gøres stationære ved at lave transformationer (fx detrending, differenser, eller kointegrationsrelationer).

OLS er konsistent under anteglse af svag fhængighed, stationaritet og eksogenitet.

Streng eksogenitet er nødvendig for middelrethed.

OLS er asymptotisk normalfordelt.

## Regressionsmetoder med ikke stationære variable



her bruges OLS.

**HBN sektion 7-9**

# Misspecifikationstest

## Misspecifikationstest 1 – Ingen autokorrelation i fejlleddet

Test for **ingen autokorrelation** i fejlleddet. Tester om den er dynamisk komplet.

Kan undersøges ved en **Breush-Godfrey** misspecifikationstest.

Hjælperegression:

Forholdet mellem residualerne i periode t og hvorvidt de afhænger af residualer fra forrige periode. afgør i hvilken grad residualer i dag afhænger af residualer i går.

* Hypotese: (Ingen seriel korrelation i residualerne, og dermed ingen tidsafhængighed)
* Teststørrelse: (Lav vidner om ingen systematik mellem residualer i går og i dag.
* Under er teststørrelsen asymptotisk -fordelt (1 frihedsgrad, da vi tester en parameter og dermed en restriktion).

Bemærk: er uobserveret. Sættes oftest lig 0.

I STATA: anvend fx **estat bgodfrey, lags(1)**. Hvis man er i tvivl om, hvad STATA beregner er det bedre at manuelt implementere dette.

## Misspecifikationstest 2

Test for **ingen heteroskedasticitet** i fejlleddet. Om vi kan antage MLR.5 eller ej.

Kan undersøges ved et **White/Breush-Pagan** test.

Hjælperegression:

Ser om der er lineær sammenhæng mellem kvadrerede residualer og de kvadrerede forklarende variable. Kunne også gøres på ikke lineær form, f.eks. krydsprodukt.

* Hypotese:

(De kvadrerede residualer ikke samvarierrer med vores forklarende variable)

* Teststørrelse:
* Under er teststørrelsen asymptotisk fordelt. Vi har 2k restriktioner og derved 2k frihedsgrader.

I STATA: **erstat hettest …indsæt variable…, iid**

## Misspecifikationstest 3

Test for **velspecificeret** funktionel form.

Kan undersøges ved **RESET.**

Hjælperegression:

Undersøger om indgår signifikant samvariation med residualet eller ej. Dette angiver . Ellers er der noget ikke linaritet som y fanger.

Hypotese:

Teststørrelse:

Under er teststørrelsen asymptotisk -fordelt.

I STATA: **estat ovtest**

Bemærk: STAT udfører test pba. følgende model:

Substantielt er der ikke stor forskel. STATA er bare mere generel. Hvis man bruger denne kommando i STATA skal man bare angive den korrekte hjælperegression mv.

**W13.1-W13.4**

# Gentagne tværsnitsdata

## Gentagne tværsnitsdata i modsætning til tværsnitsdata

**Tværsnitsdata:**

Data på en afhængig variabel, der varerer over nogle enheder i. Vi har data for en given tidsperiode i t.

Observationer af enheder, i=1,…,n, i en given periode, t:

Data er en tilfældig stikprøve fra populationen. Det er iid-data. Trukket uafhængigt fra identiske fordelinger (ikke hvis homoskedastiske fejl).

**Gentagne tværsnitsdata (T=2):**

* Periode-1 tværsnit: Her har vi trukket observationer tilfældigt fra populationen.
* Observerer samme variable. Den afhængige variabel har fodtegn 2, hvilket betyder den er målt i periode to tværsnitsdata.

Udgangspunktet er, at der ikke er overlap i individer mellem perioder. Vi udtrækker tilfældige stikprøver med unikke individer.

Fordele ved gentagne tværsnitsdata:

* Flere observationer (dermed mindre standardfejl)
* Mere Variation
* Kan undersøge flere hypoteser

## Gentagne tværsnitsdata

Vi kan kombinere data fra to tidsperioder .

Metode 1: Opstil fællesmodel og estimér

Metode 2: Opstil periode-specifikke modeller

Her opstilles periodespecifikke modeller. En for periode 1 og en fra periode 2. På baggrund af den første estimeres et sæt af parametre For den anden estimeres parametre, der er specifikke for periode 2.

Metode 3: En mellemløsning. Nogle parametre varierer på tværs af tidsperioderne. Antal estimerede parametre er s, hvor

S er større end antal parametre fra model 1 ( men mindre end antal parametre i model 3 ().

Regressionsmodel kan tillade, at parametrene ændrer sig over tid.

Denne tilgang er velkendt fra W7.

Lad d2 være en tidsdummy defineret som:

Simpel model som tillader periode-specifik konstantled:

Vi tillader at modellen kan parallelforskydes, så data fra de to perioder har forskellig skæring.

Kan introducere periode-specifikke hældningsparametre vha. interaktionsled.

## Test for strukturelle ændringer

Er der strukturelle parameterændringer/brud, hvis vi vil tillade dette, bruger vi efterfølgende model.

Hvordan kan vi teste for strukturelle parameterændringer?

General model:

Dette er den funktionelle form vi opstillede ved Chow-testet.

Hypotese .

Teststørrelse: (robust) F test.

Vi kan teste en hypotese om vores parametre er konstante.

## Politikanalyse med gentagne tværsnitsdata

Spm. Påvirker forbrændingsanlæg i huspriserne i Boston?

Hypotese: Huspriserne falder alt andet lige, hvis et forbrændingsanlæg bygges i lokalområdet.

Hvis muligt, brug information **før** og **efter** politikændring.

Data: Salgspriser og karakteristika for huse i forskellig afstand til forbrændingsanlægget.

Tværsnit: 1978 og 1981

Data indsamlet før og efter beslutningen om at opføre forbrændingsanlæg (konstruktion påbegyndt i 1981):

Naiv metode: Anvend kun 1981-data til at estimere:

hvor rprice er prisen på huset i 1978-USD og nearinc er en dummy for, om huset ligger tæt på forbrændingsanlægger.

OLS af 1981.data:

Testes på et 5% signifikansniveau ved vi, at vi ligger så langt ude i halen, at vi afviser og finder, at der er en signifikant prisforskel mellem huse, der ligger tæt på og langt fra.

Er det et godt estimat af effekten af forbrændingsanlægget?

Dette er nok lidt problematisk, da politikkere ikke altid har borgernes tanker i højsæde. Har man en mistanke om, at det ligges i et område med nogle specifikke karakteristika, så er det ikke tilfældig, hvor anlægget ligger, og vil også afspejle initiale karakteristika som ikke bare er ligge tæt på eller langt fra.

Problematisk, anlægget placeres i områder med lave huspriser i 1978.

OLS af 1978-data:

.

Denen er igen signifikant. Dette viser, at initialt er der en prisforskel mellem huse. Derfor understøtte det en hypotese om, at anlægget blev placeret i et område, hvor huse generelt havde lavere værdi.

Hvordan kan vi estimere effekten af forbrændingsanlægget?

**Bedre metode:** Difference-in-difference (dif-in-dif)

Spørgsmålet: Er husprisforskellen mellem områder tæt på og langt fra forbrændingsanlægget blevet større i 1981 end i 1978?

Dif-in dif estimator:

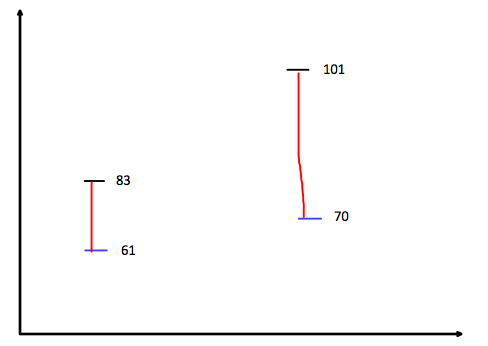
Effekten på huspriser fra forbrændingsanlægger er minus 12.000 dollars.

Antagelser bag dette resultat er, at vi i sagens natur kan udelukke, at der har været andre interventioner, der ville kunne drive forskellen i huspriser.

Det ville være relevant at undersøge, om beliggenhederne fulgte samme trends eller forskellige trends frem mod denne ændring vi ser på. Hvis de havde haft forskellige trends kunne dif-and-df skyldes forskellen i prisudviklingen og dermed ikke et skift. Man skal kunne afvise, at der er sket noget andet systematisk på tværs af de to beliggenheder.

Dif-in-dif tager højde for:

* Huspriserne kan ændre sig fra 1978 til 1981.
* Huspriserne kan initialt være forskellige i forskellige områder.



## Politikanalyse med gentagne tværsnitsdata

Dif-in.dif regressionsmdoel:

Prisudvikling fra 978 til 1981. Fanger om der har været makroøkonomiske lndringer, der har påvirket huspriser.

Prisforskellen før interventionen. Hvad er den initiale prisforskel.

Dif-in-dif i huspriserne som skyldes anlægget. Dette er OLS-etsimatet.

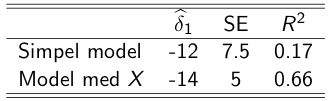
Hypotese:

Tilføj ekstra kontrolvariable:

er ikke længere simple forskelle i gennemsnit. Der er en ækvivalens mellem den simple regressionsmodel og dif-in-dif.

Hypotese:

Resultater:



På baggrund af den simple model kan vi ikke afvise, at den sande effekt er lig nul på et 5% signifikansniveau. Vi fortolker stadig -14 som et dif-in-dif estimat, hvor vi igen afviser til fordel for alternativhypotesen. Vi konkluderer, at anlægget har en negativ effekt på huspriserne.

## Dif-in-dif: Gennemsnit vs. Regression

1. Gennemsnit:

Rækkefølgen af leddene kan ændres, da rækkefølgen er arbitrær.

Vi kan tænke dif-in-dif som forskellen mellem priser på tværs af perioder, som prisudviklingen på tværs af tid på den belligenhed tæt på relativt til den lang fra. Dvs. prisudviklingen på tværs af belligenheder.

Standardfejl er mere besværlige at beregne

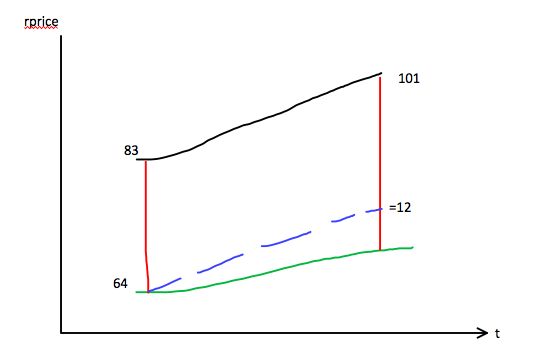
1. Regression:

Kan tilføje kontrolvariable (kan forbedre præcision)

Kan let beregne standardfejl og opstille hypotesetest.

Ulempen her er, at vi mister den pæne intuitive størrelse som når vi beregner gennesmnsit.

## Grafisk illustration ad dif-in-dif



## Naturlige eksperimenter

Med inspiration fra rigtige eksperimenter fra science, anvender økonomer ofte ”naturlige” eksperimenter, hvor en gruppe er ”behandlet” og en anden ikke er (kontrol).

”Behandlings” gruppe: Huse nær forbrændingsanlægget

”Kontrolgruppe”: Huse langt fra forbrændingsanlægget

Analyse: Sammenlign grupperne før og efter

Analogt til ”behandling” og ”placebo” i medicinske studier.

# Paneldata for T=2

Paneldata: Samme n individer observeres i periode 1 og 2.

* Periode 1 data:

Observationer på afhængige variable i periode 1.

* Periode 2:

Perfekt overlap ift. hvilke enheder, der indgår i periode 1 og 2.

* Samlet set: 2n observationer af n individer
* Bemærk: Der kan være mange år imellem t=1 og t=2.

Også kendt som longitudinal data.

## Model med uobserveret heterogenitet

Simple model:

er skæringen for periode 2. I den konkrete model antager vi effekten af x på y er den samme uanset, hvilken periode vi er i ().

Tidsdummien modellerer makroeffekter.

Fejlleddet har to komponenter: .

Dette fejlled er uobserverbart, men fejlleddet er uobserverbart heterogenitet, der er individspecikke, men som ikke varierer over tid. F.eks. ved afkast af uddannelse, og medfødte evner ikke kan observeres men påvirker løn. Dette er individspecifik, men er ikke noget der ændres over tid (generelt).

Uobserveret heterogenitet (uobserveret ”fixed effect”):

* Varierer ikke over tid
* Individspecifikt

Idiosynkratisk fejlled

* Varierer tilfældigt på tværs af individer og tidsperioder (varierer over i dimensionen og t dimensionen)
* Fungerer som vores sædvanlige fejlled

## Antageler for model med uobserveret heterogenitet

* Tilfældig stikprøve (ingen korrelation mellem individ i og j)
* Fejlled er defineret som .

De fra tværsnitsdimensionen er tilfældigt udvalgt, men er de samme over perioder.

* Streng eksogenitet:
* Medfører at det idiosynkratiske fejlled er ukorreleret med:
  + Forklarende variable
  + Observerbare individ-specifikke effekt
* Ingen sammenhængen mellem i en given periode og noget, der findes i x.
* **Bemærk:** Ingen antagelser vedrørende

## Korreleret uobserveret heterogeitet

Antag at

I dette tilfælde vil OLS være biased og inkonsistent.

Med tværsnitsdata kan dette problem ikke løses uden ydereligere antagelser (f.eks. IV-metoden): Dvs. hvis vi kan finde et relevant validt instrument, kan dette afhjælpes.

Med gentagne observationer på de samme enheder er der dog andre muligheder.

**Ny metode:** Transformer modellen således at:

* Parameteren kan identificeres
* Den uobserverbare heterogenitet er elimineret

Denne har til formål at elleminere og muliggøre, at stadig kan findes vha. OLS, da det fortsat er en lineær regressionsmodel.

En mulig tilgang: **First-Difference (FD) estimatoren.**

## First-difference estimation:

Model:

Vi har fortsat et fejlled bestående af et idiosynkratisk fejlled og et individspecifikt.

Periode 2:

Periode 1:

Førstedifferenser:

Vi er gået fra paneldata med to observationer pr. enhed til kun at have et tværsnit af førstedifferencer. Vi har dog fået fjernet der var kilden til endogenitetsproblemet.

Data består nu af et tværsnit af førstedifferenser.

Differensen fjerner som var kilden til endogenitet.

OLS er konsistent, hvis .

## Fordele og ulemper ved paneldata

**Fordele:**

Paneldata gør det muligt at fjerne uobserverbare individspecifikke størrelser som ellers giver bias for OLS.

**Ulemper:**

Som regel vanskeligt eller svært at indsamle paneldata. Interviewe samme personer over år.

Ikke muligt at estimere effekten af x-variable som ikke varierer over tid.

## Vigtige pointer

* Gentagne tværsnitsdata gør det muligt at undersøge flere hypoteser. Fx politikevalueringer og parameterstabilitet.
* Politikevaluering kan, hvis man kan finde en kontrolgruppe, lave ved dif-in-dif estimation.
* Metoden er let at implementere som regressionsmodel.
* Paneldata gør det muligt at fjerne endogenitetsproblemer som skyldes ikke-observerede individspecifikke størrelser.
* Generel spørgsmål: Er det mest troværdigt at beregne estimater ved brug af variation over tid for de samme individer eller variation på tværs af individer i den same tidsperiode?

**W13.5+14.1+14.2**

# Paneldata

Kan tænkes som en kombination af tværsnitsdata og tidsseriedata, da vi har enheder vi har indsamlet tilfældigt i populationen og disse følges over tid. Ofte er N meget stor relativt til T.

Dette er vigtigt at pointere ud fra, at hvis T er meget stor skal man vurdere om der er stationaritet og svag tidsafhængighed. Dette kan man ignorere, hvis tidsdimensionen er meget kort.

## Simpel panel data regressionsmodel for T>2

Regressionsmodellen:

for og .

Vi antager en simpel lineær regressionsmodel. Vi har et fejlled, der afhænger af både tværsnitsdimensionen og tidsdimensionen.

Under betingelsen:

er OLS estimatoren konsistent. Den betingede værdi af fejlledet givet vores forklarende variabel er lig nul.

I en pooled OLS ignorerer vi faktum af, at vi har paneldata og benytter det som tværsnitsdata. Dette kendes som **pooled OLS** estimator. Denne er konsistent.

## Panel data regressionsmodel med uobserveret heterogenitet

Vi betragter modellen:

hvor er et fejlled med et individspecifikt led.

1. **Pooled OLS**  er konsistent, hvis , dvs. at både den individspecifikke effekt og det idiosynkratiske fejlled er ukorreleret med vores forklarende variabel.
2. Hvis pooled OLS er konsistent, findes der en mere efficient GLS estimator. Dette er den såkaldte **Random Effects (RE)**  estimator. Vi kan finde en anden estimator med en lavere varians. RE er mere efficient, da hvis vores fejlled består af en individspecifik del, vil den serielle korrelation i det samlede fejlled, vil denne være forskellig fra nul grundet den individspecifikke effekt. OLS er kun konsistent men ikke efficient.   
   RE skal vi ikke kunne implementere i STATA.
3. I mange tilfælde er det muligt, at er korreleret med .   
   Løsningen: Transformer modellen for t eliminere og benyt følgende relevante transformationer:
   1. **First Differences (FD)** estimator.
   2. **Fixed Effects (FE)** estimator.

## FD estimator

Vi betragter modellen:

for og .

Løsning er at endogenitetskilden kan elimineres vha. førstedifferenser:

hvor . Der er ikke en periodespecifik skæring, hvorfor skæringen definitorisk er lig nul. Vi har elimineret og indgår stadig multiplikativt.

FD estimatoren er konsistent, hvis , dvs. vi kan udelukke sammenhængen mellem u i førstedifferenser og x i førstedifferenser.

Antallet af observationer efter transformationen FD: .Vi mister altså en tidsdimension grundet førstedifferens.

Vi fortolker i overensstemmelse med den oprindelige model. Kritik er, at den eliminerer alle forkalrende variable, der ikke varierer over tid.

## FE estimator

Vi betragter modellen:

for og .

Når vi impelenterer FE estimatoren er det med udgangspunkt i tidsgennemsnit. Vi tager gennemsnittet regressionsmodel på tværs af T. Dvs. for hver individ i vores data beregner vi et individs gennemsnitlige y over perioden:

Betragt within-transformed (afvigelser indenfor gennemsnit) model:

Denne transformation sikrer er elimineret og som er lineære, hvorfor vi kan bruge OLS.

FE estimatoren er konsistent, hvis (svarer til MLR.4).

FE kræver **streng eksogenitet** stærkere antagelse end for FD. Vi har ikke alene inforation på tidspunkt t, men afhænger også af de idiosynkratiske fejlled i alle perioder. Dvs. vi kræver, at der ingen forhold er mellem de idiosynkratiske fejlled og x i alle tænkelige perioder. FD har kun antagelse om periode T og T-1, hvorfor mindre restriktiv.

Gevinst ved within-tranformationen er at det ikke koster også nogle observationer

Antal observationer efter within-transformationen:

## RE

Vi betragter modellen:

for og .

Hvis pooled OLS er konsistent, e RE estimatoren mere efficient.

**Intuition:**  for t, s I modellen givet .

Efficient GLS estimation er opnået ved at benytte OLS:

hvor fanger den serielle afhængighed over tid (se W for detaljer).

RE estimatoren er konsistent, hvis

RE kræver **streng eksogenitet.** Fordi RE estimatoren kun anvender en brøkdel af gennemsnit, så gælder det stadig, at der ingen systematisk sammenhæng mellem det transformerede fejlled og vores x må være i en hvilken som helst periode.

## Inferens

Vi bliver nødt til at lave antagelser angående den serielle korrelation i det idiosynkratiske fejlled, . Dette kommer først på senere kurser.

Nærlæs appendix **13.A** og **14.b.**

Standard inferens er valid for FD estimatoren når (dette skal gælde før end, at vi kan bruge vores almindelige OLS formler):

og .

Dette implicerer, at er meget seriel korreleret.

Standard inferens er valid for FE estimatoren når:

Dette implicerer at er ikke seriel korreleret.

Med ydereligere antagelser for og , virker standard inferens også for RE.

Skriv resten af 9!!

## Argumenter til at foretrække en tilgang frem for den anden

FD og FE er det samme for T=2. De vil føre til numerisk samme estimater i denne situation. Dette kan vises let.

Hvis FE estimatoren er konsistent så er FD (konsistent under mindre restrktive antagelser) også. FD estimatoren er at foretrække, hvis seriel afhængighed i det idiosynkratiske fejlled finder sted over tid. Modsat, hvis man ikke forventer dette. Hvis man skal træffe et valg skal det afhænge af, hvordan man forventer fejlledet udvikler sig over tid. Der er dog ikke et åbentlyst argument for at vælge det ene frem for det andet.

Pooled OLS estimator er konsistent under strengere antagelser, men RE estimatoren (mindre std. Fejl) er mere efficient.

Med tilgængelige data kan man benytte alle estimatorerne.

Hvilken en skulle man foretrække? Her kan man tænke over følgende:

* Hvilke underliggende antagelser ses mest realistiske?
* Er de empiriske estimater nogenlunde ens eller meget forskellige?
* Hvilke estimatorer er mest robuste? FD virker til at være mere robust, da den ikke kræver streng eksogenitet, som FE er følsom overfor.
* Sammenlign og diskutér!

### Huskeliste til at vælge mellem:

* Hvis benyt FD eller FE. Der er endogenitetsproblemer, hvis kovariansen er forskellig fra nul. Da vi har flere observationer over tid på samme enhed kan man lave en transofrmation, så individspecifikke del af fejlledet forsvinder.
* Hvis benyt pooled OSL eller RE.
* Dermed skal parametrene blive beregnet fra variationer imellem enheder på tværs af tid (tværsnitsdata) eller imellem tid på tværs af enheder. Vi er følsomme for, at individer er forskellige systematisk, og at dette er uobserverbart. En tværsnitsverden snakker vi om udeladte variable, da dette giver problemer ift. vores økonometriske analyse og vores kausale fortolkning. I paneldata gør vi det modsatte, vi bruger FD og FE hvor vi kigger på variation over tid på de samme enheder. Dvs. det er tidsændringerne for den samme enhed vi bruger til at beregne effekten af x på y. Denne type analyse er typisk mindre følsom overfor, at vi sammenligner pærer og bananer dvs. individspecifikke uobserverbare størrelser. Paneldata er ofte at foretrække men er ikke altid tilgængeligt! Hvad er det for noget variation i vores data vil bruge til at beregne estimator. Hvilken variation vi tror giver retvisende estimater med en kausal fortolkning.
* Husk at hver estimator bygger på specifikke antagelser.
* Vigtigt at diskuterer, hvilken der er bedst i et givent tilfælde.

Udeladte variable giver anledning til endogenitet og bias for OLS estimatoren. Dette har vi prøvet at afhjælpe vha. paneldata. Dvs. vi forsøger at få efficiente estimater.