

Øvelsestime 7

Heteroskedasticitet

Agenda

- 1) Introduktion til PS7
- 2) Gruppearbejde: Diskussion
- 3) Opsamling
- 4) Gruppearbejde: Stata
- 5) Opsamling

Tobias Gabel Christiansen
tgc@econ.ku.dk
19-04-2021

KØBENHAVNS UNIVERSITET



0) Planen for semesteret

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

 Her er vi :-)

1) Introduktion til PS7

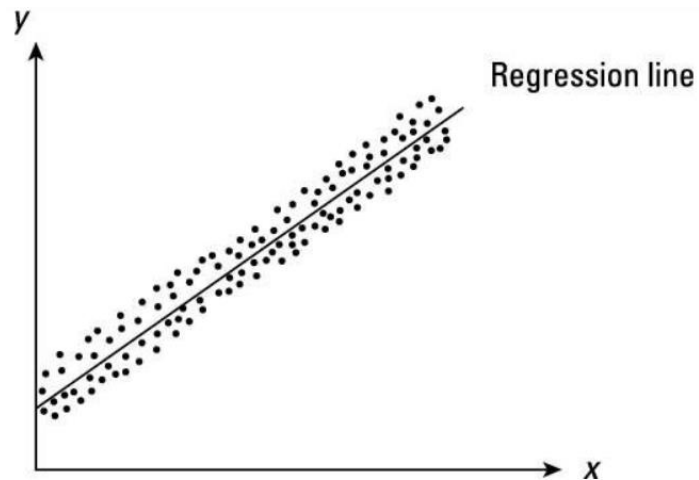
- I PS7 har vi fokus på situationer hvor MLR.5 er brudt (dvs. situationer med heteroskedasticitet).
- Heteroskedasticitet betyder at fejlledeets varians varierer på tværs af observationer. Med andre ord er vores observationer ikke længere i.i.d (independent and identically distributed) men i.n.i.d (independent and not identically distributed).
- Lader vi $u = [u_1 \ u_2 \ \dots \ u_n]'$, hvor $E[u] = 0$ er kovariansmatrixen er givet ved:

$$\text{var}(u|X) = E(uu'|X) = \begin{pmatrix} \sigma_1^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_n^2 \end{pmatrix} \neq \begin{pmatrix} \sigma^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

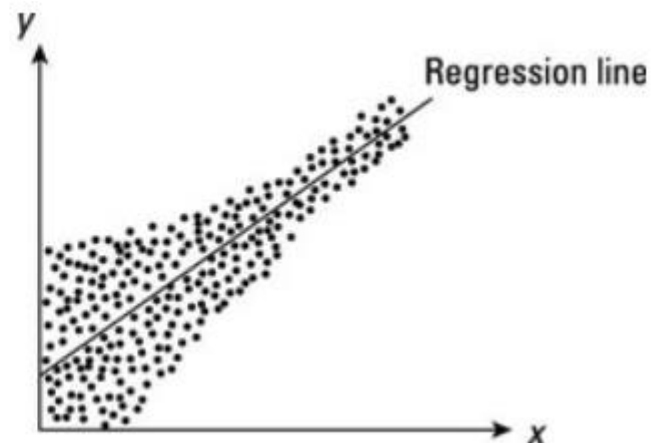
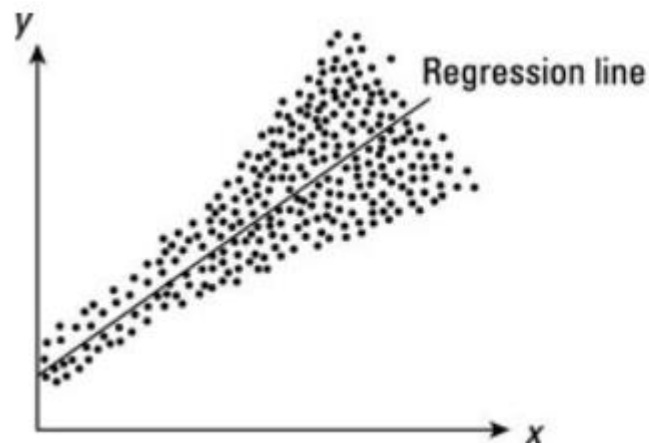
- Da vores observationer udelukkende er kendetegnet ud fra vores kontrolvariablene ($X = x_i$) tolkes det generelt som at fejlledeets varians varierer med X og skrives på formen $\text{var}(u|X) = \sigma^2 h(X)$.

1) Introduktion til PS7

Eksempel på homogenitet:



Eksempler på heteroskedasticitet:



1) Introduktion til PS7

Hvis MLR.1-4 stadig er overholdt er konsekvenserne af heteroskedasticitet at:

- i) OLS estimatoren ikke er "BLUE" men blot "LUE" (dvs. ikke længere mest efficiente estimator)
- ii) Vigtigere er det at den formel vi bruger til at beregne estimatorens varians ($\text{var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$) ikke længere er valid. Det betyder at vores standardfejl ikke er valide, og dermed også at F- og t-test ikke er valide når de beregnes ud fra disse.

Bemærk: OLS estimatoren er altså stadig unbiased og konsistent (kræver kun MLR.1-4)

1) Introduktion til PS7

Der er flere måder at håndtere heteroskedasticitet på:

1. Den simpleste (og mest brugte i praksis) er at ignorere at OLS estimatoren ikke længere er efficient og blot korrigere vores beregning af estimatorens standardfejl så denne er valid. Med andre ord bruges robuste standardfejl (se forelæsning 13).
2. Alternativ kan vi forsøge at lave en simpel logaritmisk transformation af vores model og se om det hjælper på problemet. Hvis det gør, kan vi estimere den nye model med OLS som vi plejer. Bemærk: denne transformation ændrer også på vores fortolkning af parametrene.
3. Sidst men ikke mindst, kan vi bruge Weighted Least Squares (WLS) estimatoren. Bemærk: denne transformation ændrer ikke på vores fortolkning af parametrene.

I dag ser vi lidt nærmere på alle tre muligheder

1) Introduktion til PS7

- Vi arbejder videre med modellen fra PS6 hvor vi forsøgte at estimere lejlighedspriser i København
- Den foretrukne model var givet ved:

$$\log(\text{price}_i) = \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \beta_1 m2_i + \beta_2 m2_i^2 + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + \epsilon_i$$

- I PS7 ændrer vi modellen til:

$$\text{price}_i = \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \beta_1 m2_i + \beta_2 m2_i^2 + \beta_3 \text{rooms}_i + \beta_4 \text{toilets}_i + u_i$$

- Vi tager altså ikke længere logaritmen til priserne. Dette ændrer dels:
 - i) Hvordan vi tolker vores parametre (ikke længere relative ændringer i prisen)
 - ii) Desuden kan logaritmiske transformationer gøre fordelinger mere symmetriske og dermed reducere heteroskedasticitet (ser vi nærmere på til sidst i dag)

2) Gruppearbejde - Diskussion

- To-do 25 minutter:
 1. Suppose that assumption MLR.5 **does not** hold for model (2). What does this mean for the properties of the OLS estimator?
 2. Suppose that apartment size, $m2_i$, affects the variance of the error term and is thus a relevant scale variable driving the heteroskedasticity. Provide an economic argument for why the variance of the error terms may depend on apartment size. Explain how it is possible for the error terms to satisfy MLR.4, but not MLR.5.
 3. Write down an expression for the conditional variance of u_i for the case where the variance of the error term is exactly proportional to $m2_i$. How can the WLS estimator be implemented in this case?

OBS: Brug for hjælp? Skriv Jeres gruppenummer i Padlet (link i chatten)

3) Gruppearbejde - Opsamling

1. Suppose that assumption MLR.5 **does not** hold for model (2). What does this mean for the properties of the OLS estimator?

3) Gruppearbejde - Opsamling

1. Suppose that assumption MLR.5 **does not** hold for model (2). What does this mean for the properties of the OLS estimator?
- OLS estimatoren er stadig unbiased og konsistent hvis MLR.5 ikke overholdt. Disse egenskaber kræver kun at MLR.1-4 er overholdt.
 - Hvis MLR.5 ikke holder så er OLS estimatoren dog ikke længere den bedste lineære estimator (BLUE). Dvs. der findes andre lineære estimatorer som har lavere varians og dermed er mere præcise.
 - Vigtigere er det at den formel vi bruger til at beregne estimatorens varians ($\text{var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}$) ikke længere er valid. Det betyder at vores standardfejl og dermed også F- og t-test ikke er valide når de beregnes ud fra disse (statistikkerne er ikke længere hhv. F- og t-fordelt under H_0).

3) Gruppearbejde - Opsamling

2. Suppose that apartment size, $m2_i$, affects the variance of the error term and is thus a relevant scale variable driving the heteroskedasticity. Provide an economic argument for why the variance of the error terms may depend on apartment size. Explain how it is possible for the error terms to satisfy MLR.4, but not MLR.5.

3) Gruppearbejde - Opsamling

2. Suppose that apartment size, $m2_i$, affects the variance of the error term and is thus a relevant scale variable driving the heteroskedasticity. Provide an economic argument for why the variance of the error terms may depend on apartment size. Explain how it is possible for the error terms to satisfy MLR.4, but not MLR.5.
- Der kan være flere årsager til at fejleddets varians (og dermed også variationen i lejlighedspriserne) afhænger af størrelsen på boligen
 - Fx kan man forestille sig at større lejligheder er udsat for større usikkerhed når prisen bestemmes, fordi efterspørgslen efter store (og som oftest dyre) lejligheder er mere volatil.
 - Med andre ord bliver der foretaget færre handler af store lejligheder end af gennemsnitlige/små lejligheder da markedet for disse er mindre, og dermed er det svære at sige hvad den rette ligevægtspris er.

3) Gruppearbejde - Opsamling

2. Suppose that apartment size, $m2_i$, affects the variance of the error term and is thus a relevant scale variable driving the heteroskedasticity. Provide an economic argument for why the variance of the error terms may depend on apartment size. Explain how it is possible for the error terms to satisfy MLR.4, but not MLR.5.
- At MLR.4 ($E(u|X) = 0$) er overholdt mens MLR.5 ($var(u|X) = \sigma^2$) er brudt betyder at fejlledet i gennemsnit er nul, men at variansen variere på tværs af observationer.
 - Med andre ord er MLR.4 og MLR.5 vidt forskellige antagelser:
 - Rent sandsynlighedsteoretisk, er det muligt at stokastiske variable (fx u og X) er mean independent, samtidigt med at deres varianser afhænger af hinanden (eller omvendt).

3) Gruppearbejde - Opsamling

3. Write down an expression for the conditional variance of u_i for the case where the variance of the error term is exactly proportional to m^2_i . How can the WLS estimator be implemented in this case?

3) Gruppearbejde - Opsamling

3. Write down an expression for the conditional variance of u_i for the case where the variance of the error term is exactly proportional to $m2_i$. How can the WLS estimator be implemented in this case?
- Hvis variansen er ligefrem proportional med $m2_i$ så gælder det at:

$$\text{var}(u_i|X) = \sigma^2 m2_i$$

- Vi antager altså at vi kender formen på heteroskedasticiteten ($h(x_i) = m2_i$). Bemærk at σ^2 stadig er en ukendt konstant.
- WLS transformerer modellen så variansen på fejllædet er konstant (homoskedastisk) ved at dividere igennem med $\sqrt{m2_i}$:

$$\frac{\text{price}_i}{\sqrt{m2_i}} = \frac{\beta_0}{\sqrt{m2_i}} + \delta_1 \frac{KbhK_i}{\sqrt{m2_i}} + \delta_2 \frac{KbhN_i}{\sqrt{m2_i}} + \delta_3 \frac{KbhV_i}{\sqrt{m2_i}} + \beta_1 \frac{m2_i}{\sqrt{m2_i}} + \beta_2 \frac{m2_i^2}{\sqrt{m2_i}} + \beta_3 \frac{\text{rooms}_i}{\sqrt{m2_i}} + \beta_4 \frac{\text{toilets}_i}{\sqrt{m2_i}} + \frac{u_i}{\sqrt{m2_i}}$$

3) Gruppearbejde - Opsamling

$$\frac{price_i}{\sqrt{m2_i}} = \frac{\beta_0}{\sqrt{m2_i}} + \delta_1 \frac{KbhK_i}{\sqrt{m2_i}} + \delta_2 \frac{KbhN_i}{\sqrt{m2_i}} + \delta_3 \frac{KbhV_i}{\sqrt{m2_i}} + \beta_1 \frac{m2_i}{\sqrt{m2_i}} + \beta_2 \frac{m2_i^2}{\sqrt{m2_i}} + \beta_3 \frac{rooms_i}{\sqrt{m2_i}} + \beta_4 \frac{toilets_i}{\sqrt{m2_i}} + \frac{u_i}{\sqrt{m2_i}}$$

- Denne transformation bryder ikke med MLR.1-4
 - i) Modellen er stadig lineær i parametre
 - ii) Sample er stadig random
 - iii) Generelt stadig ingen perfekt multikollinearitet
 - iv) $E\left(\frac{u_i}{\sqrt{m2_i}} \middle| X\right) = \frac{1}{\sqrt{m2_i}} E(u_i | X) = 0$
- Transformationen ændrer desuden ikke på parametrene som kan fortolkes som før. Fx er effekten af ét ekstra rum (dvs. en stigning i $rooms_i$ og altså ikke én ekstra enhed af $\frac{rooms_i}{\sqrt{m2_i}}$) på prisen (dvs. $price_i$ og ikke $\frac{price_i}{\sqrt{m2_i}}$) givet ved β_3 (alt andet lige)
- OLS og WLS estimatorerne er forskellige estimators for de samme parametre.
- Bemærk dog at modellen ikke længere har et konstantled (vigtigt at huske når WLS implementeres manuelt)

3) Gruppearbejde - Opsamling

$$\frac{price_i}{\sqrt{m2_i}} = \frac{\beta_0}{\sqrt{m2_i}} + \delta_1 \frac{KbhK_i}{\sqrt{m2_i}} + \delta_2 \frac{KbhN_i}{\sqrt{m2_i}} + \delta_3 \frac{KbhV_i}{\sqrt{m2_i}} + \beta_1 \frac{m2_i}{\sqrt{m2_i}} + \beta_2 \frac{m2_i^2}{\sqrt{m2_i}} + \beta_3 \frac{rooms_i}{\sqrt{m2_i}} + \beta_4 \frac{toilets_i}{\sqrt{m2_i}} + \frac{u_i}{\sqrt{m2_i}}$$

- Transformationen korrigerer for heteroskedasticitet da:

$$var\left(\frac{u_i}{\sqrt{m2_i}} \middle| X\right) = \frac{1}{m2_i} var(u_i|X) = \frac{1}{m2_i} \sigma^2 m2_i = \sigma^2$$

- Hvis vi har specificeret formen for heteroskedicitet rigtigt ($h(x_i) = m2_i$) kan vi med WLS opnå en estimator med homoskedastiske fejllid som er BLUE!
- Bemærk: Hvis vi specificerer formen for heteroskedicitet forkert (dvs. $h(x_i) = m2_i$ er forkert) er WLS estimatoren stadig unbiased og konsistent (som OLS-estimatoren) men lider stadig af heteroskedasticitet. Med andre ord lider WLS estimatoren i det tilfælde af de samme mangler som OLS når MLR.5 er brudt.
- Et alternativ til antage formen for heteroskedacitet $h(x)$ er at estimere denne. Med andre ord bruger vi vores data til først at estimere $h(\widehat{x})$, for derefter at bruge denne størrelse på præcis samme måde som ovenfor. I kan læse mere om hvordan dette gøres og hvilke egenskaber estimatoren har i Wooldridge kapitel 8.

3) Gruppearbejde - Opsamling

Sidebemærkning:

- Grunden til at det hedder WLS er at vi vægter hver observation med $\frac{1}{h(x_i)}$ når vi minimere sum of squared residuals:

$$\min_{\hat{\beta}} \sum_i^n \left(\frac{\hat{u}_i}{\sqrt{h(x_i)}} \right)^2 = \min_{\hat{\beta}} \sum_i^n \frac{\hat{u}_i^2}{h(x_i)}$$

- Da variansen er ligefrem proportional med $h(x)$ [$var(u_i|X) = \sigma^2 h(x_i)$], svarer det til at vi giver mindre vægt til observationer med stor usikkerhed og omvendt. Dette er grunden til at WLS er mere efficient end OLS, da sidstnævnte giver samme vægt til alle observationer ($h(x_i) = 1$).

4) Gruppearbejde - Stata

- To-do indtil 12:30:
- Arbejd på Stataøvelserne

Hint:

- For at gøre det lettere at læse vores resultater skalerer vi $price_i$ til millioner kroner og skalerer $m2_i^2$ med 1000

```
replace price=price/1000000  
gen sqm2=m2*m2/1000
```

OBS: Brug for hjælp? Skriv Jeres gruppenummer i Padlet (link i chatten)

5) Opsamling - Stata

1. Estimate model (2) by OLS using CPHAPTS.dta.

- Ved at dividere price med 1 mio. skalerer vi alle koefficienter med 1 mio:

$$\frac{y}{k} = \frac{\beta_0}{k} + \frac{\beta_1}{k} x_1 + \frac{u}{k} = \beta_0^s + \beta_1^s x_1 + u^s$$

- Vi gjorde ikke dette i PS6 da vi brugte $\log(\text{price}_i)$ og dermed fokuserede på relative ændringer (niveau for prisen har ingen effekt og skaleringer af forklarende variable sætter sig blot i konstantleddet). Resultaterne forbliver uændret. Husk blot at tolke effekten på prisen i mio. kroner fremfor kroner
- Da koefficienten på sqm2 var lille i forvejen, dividere vi denne variabel med 1000 for at gøre koefficienten tilsvarende større:

$$y = \beta_0 + \beta_1 k \frac{x_1}{k} + u = \beta_0 + \beta_1^s x_1^s + u$$

- Husk at tolke koefficienten på $\frac{\text{sqm2}}{1000}$ som en ændring i denne størrelse og ikke sqm2 .

```
replace price=price/1000000
gen sqm2=m2*m2/1000
gen KbhK=1 if location=="KBH K"
replace KbhK=0 if KbhK==.
```

```
gen KbhN=1 if location=="KBH N"
replace KbhN=0 if KbhN==.
```

```
gen KbhV=1 if location=="KBH V"
replace KbhV=0 if KbhV==.
```

```
regress price KbhK KbhN KbhV m2 sqm2 rooms toilets
```

Source	SS	df	MS	Number of obs	=	988
Model	653.50041	7	93.3572014	F(7, 980)	=	289.21
Residual	316.341052	980	.322796992	Prob > F	=	0.0000
				R-squared	=	0.6738
				Adj R-squared	=	0.6715
Total	969.841463	987	.982615464	Root MSE	=	.56815

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
KbhK	.1924627	.0479291	4.02	0.000	.0984072	.2865182
KbhN	-.2133712	.0489475	-4.36	0.000	-.3094252	-.1173173
KbhV	-.06464	.0562868	-1.15	0.251	-.1750966	.0458166
m2	.0371188	.002064	17.98	0.000	.0330683	.0411693
sqm2	-.0669611	.0077956	-8.59	0.000	-.082259	-.0516631
rooms	-.0547265	.0285585	-1.92	0.056	-.1107694	.0013163
toilets	.064069	.0884127	0.72	0.469	-.109431	.237569
_cons	-.1933774	.1372517	-1.41	0.159	-.4627184	.0759635

5) Opsamling - Stata

1. Estimate model (2) by OLS using CPHAPTS.dta.

- Estimaterne har samme fortegn som estimerne fra PS6
- Estimaterne fortolkes nu som *absolutte* ændringer i den forventede salgspris

```
replace price=price/1000000
gen sqm2=m2*m2/1000
gen KbhK=1 if location=="KBH K"
replace KbhK=0 if KbhK==.
```

```
gen KbhN=1 if location=="KBH N"
replace KbhN=0 if KbhN==.
```

```
gen KbhV=1 if location=="KBH V"
replace KbhV=0 if KbhV==.
```

```
regress price KbhK KbhN KbhV m2 sqm2 rooms toilets
```

Source	SS	df	MS	Number of obs	=	988
Model	653.50041	7	93.3572014	F(7, 980)	=	289.21
Residual	316.341052	980	.322796992	Prob > F	=	0.0000
				R-squared	=	0.6738
				Adj R-squared	=	0.6715
Total	969.841463	987	.982615464	Root MSE	=	.56815

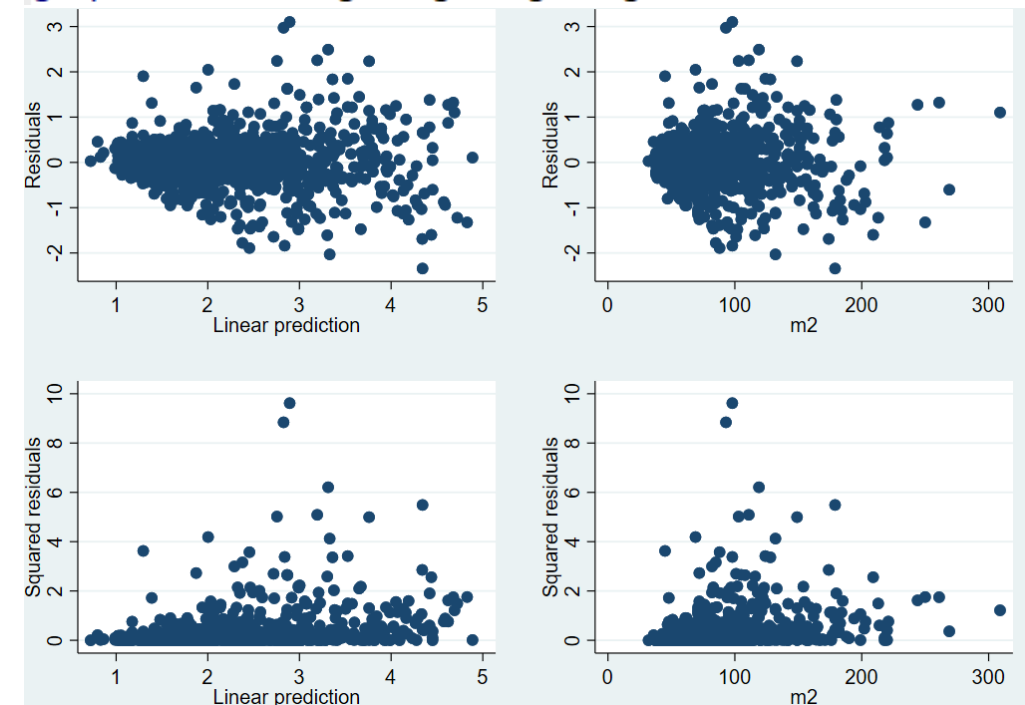
price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
KbhK	.1924627	.0479291	4.02	0.000	.0984072	.2865182
KbhN	-.2133712	.0489475	-4.36	0.000	-.3094252	-.1173173
KbhV	-.06464	.0562868	-1.15	0.251	-.1750966	.0458166
m2	.0371188	.002064	17.98	0.000	.0330683	.0411693
sqm2	-.0669611	.0077956	-8.59	0.000	-.082259	-.0516631
rooms	-.0547265	.0285585	-1.92	0.056	-.1107694	.0013163
toilets	.064069	.0884127	0.72	0.469	-.109431	.237569
_cons	-.1933774	.1372517	-1.41	0.159	-.4627184	.0759635

5) Opsamling - Stata

2. Make scatter plots of $\hat{u}_i$ and $\hat{u}_i^2$ against $\widehat{price}_i$ and $m2$ as a graphical test of heteroskedasticity. You should produce 4 plots in total. Do the

- Ved at plotte residualerne ($\hat{u}_i$) mod $\widehat{price}_i$ kan vi danne os et indtryk af om variansen for residualerne er en funktion af vores forkarende variable (husk $\widehat{price}_i$ er en lineær funktion af alle de forklarende variable). Hvis variansen for $\hat{u}_i$ varierer med de forklarende variable indikerer det at variansen for u_i også gør (=heteroskedasticitet).
- Ved at plotte residualerne mod $m2_i$ zoomer vi ind og ser om variansen af residualerne er en funktion af størrelsen på lejlighederne.
- Graferne giver en indikation af heteroskedasticitet, men der er ikke noget klart mønster.

```
regress price KbhK KbhN KbhV m2 sqm2 rooms toilets
predict uhat, residuals
predict yhat, xb
gen uhat2=uhat*uhat
scatter uhat yhat, name(fig1, replace)
scatter uhat m2, , name(fig2, replace)
scatter uhat2 yhat, name(fig3, replace)
scatter uhat2 m2, , name(fig4, replace)
graph combine fig1 fig2 fig3 fig4
```



5) Opsamling - Stata

3. Perform the Breusch-Pagan test for heteroskedasticity by regressing $\hat{u}_i^2$ against the right hand side variables of model (2). Write up the null and
- Breusch-Pagan testen er en formel måde at teste om der er tegn på heteroskedasticitet.
 - Testen bygger på at variansen for fejllædet er konstant under antagelsen om homoskedasticitet:

$$\text{var}(u|X) = E(u^2|X) = \sigma^2$$

- Med andre ord er $E(u^2|X)$ ikke en funktion af vores forklarende variable under homoskedasticitet. Da vi aldrig observerer u_i kan vi ikke teste dette direkte. Vi kan derimod erstatte u_i^2 med vores estimerede residualer fra vores model $\hat{u}_i^2$. Vi estimerer derfor følgende model:

$$\hat{u}_i^2 = \phi_0 + \phi_1 KbhK_i + \phi_2 KbhN_i + \phi_3 KbhV_i + \phi_4 m2_i + \phi_5 m2_i^2 + \phi_6 rooms_i + \phi_7 toilets_i + \eta_i$$

- Vi kan ud fra dette teste om $\hat{u}_i^2$ er en funktion af X gennem følgende hypotese:

$$\begin{aligned} H_0: \phi_1 &= \dots = \phi_7 = 0 \\ H_A: \exists i: \phi_i &\neq 0, \quad i = 1, \dots, 7 \end{aligned}$$

- Dette kan testes enten gennem en F-test eller en LM-test. Hvis vi forkaster nulhypotesen så forkaster vi antagelsen om homoskedasticitet.
- **Bemærk:** Vi tester om MLR.5 er brudt i den oprindelige model. I modellen hvor den afhængige variabel er $\hat{u}_i^2$ (auxiliary modellen) antages det ofte at MLR.5 overholdt. LM-testen (samt F- og t-test) er der for vallide med konventionelle standardfejl under denne antagelse.

5) Opsamling - Stata

3. Perform the Breusch-Pagan test for heteroskedasticity by regressing $\hat{u}_i^2$ against the right hand side variables of model (2). Write up the null and

regress uhat2 KbhK KbhN KbhV m2 sqm2 rooms toilets

Source	SS	df	MS	Number of obs	=	988
				F(7, 980)	=	13.93
Model	52.0806422	7	7.44009174	Prob > F	=	0.0000
Residual	523.376636	980	.534057792	R-squared	=	0.0905
				Adj R-squared	=	0.0840
Total	575.457278	987	.583036756	Root MSE	=	.73079

uhat2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
KbhK	.1156022	.0616494	1.88	0.061	-.0053779	.2365822
KbhN	-.0118628	.0629593	-0.19	0.851	-.1354133	.1116877
KbhV	.0281995	.0723996	0.39	0.697	-.1138765	.1702756
m2	.0121819	.0026549	4.59	0.000	.0069719	.0173918
sqm2	-.0222252	.0100271	-2.22	0.027	-.0419024	-.0025481
rooms	-.0770872	.0367337	-2.10	0.036	-.149173	-.0050014
toilets	.1473982	.1137219	1.30	0.195	-.0757682	.3705646
_cons	-.4885814	.1765416	-2.77	0.006	-.8350244	-.1421384

Regressionen af $\hat{u}_i^2$ på X refereres til som en auxiliary regression.

5) Opsamling - Stata

3. Perform the Breusch-Pagan test for heteroskedasticity by regressing $\hat{u}_i^2$ against the right hand side variables of model (2). Write up the null and

LM-test

- LM-testen er givet ved: $LM = nR_{\hat{u}^2}^2$ hvor n angiver antal observationer og $R_{\hat{u}^2}^2$ er de forklarende variables forklaringsgrad i regression med $\hat{u}_i^2$ på X .
- Lm-testen følger asymptotisk en χ_k^2 -fordeling (hvor k angiver antallet af restriktioner) under MLR.1-5 og H_0 .

```
regress uhat2 KbhK KbhN KbhV m2 sqm2 rooms toilets
gen BP=e(N)*e(r2)
gen c=invchi2tail(7,0.05)
su BP c
```

Variable	Obs	Mean	Std. Dev.	Min	Max
LM	988	89.41702	0	89.41702	89.41702
c	988	14.06714	0	14.06714	14.06714

- Da LM-statistikken er langt over den kritiske værdi for et 5% signifikansniveau forkaster vi nulhypotesen om homoskedasticitet.
- Vi finder altså empirisk belæg for heteroskedasticitet.

5) Opsamling - Stata

3. Perform the Breusch-Pagan test for heteroskedasticity by regressing $\hat{u}_i^2$ against the right hand side variables of model (2). Write up the null and

F-test

- F-testen implementeres gennem vores auxiliary regression af $\hat{u}_i^2$ på vores forklarende variable. Vi ved at F-testen asymptotisk følger en F-fordeling under MLR.1-5 og H_0 .

- Vi finder en F-værdi på 13.93:

```
regress uhat2 KbhK KbhN KbhV m2 sqm2 rooms toilets  
test KbhK KbhN KbhV m2 sqm2 rooms toilets
```

```
F( 7, 980) = 13.93  
Prob > F = 0.0000
```

- Vi forkaster derfor nulhypotesen om homoskedasticitet på et 5% signifikansniveau til fordel for at der er tegn på heteroskedasticitet.

5) Opsamling - Stata

4. Perform a specific Breusch-Pagan test for heteroskedasticity assuming that apartment size is the most likely cause of heteroskedasticity. Per-

- Vi bliver nu bedt om at undersøge om variansen af u_i er en funktion af $m2_i$.
- Dette gøres ved præcis samme fremgangsmåde som før, men da vi nu fokuserer på én af de forklarende variable kan vi benytte en t-test. Formelt estimerer vi altså følgende model:

$$\hat{u}_i^2 = \phi_0 + \phi_1 m2_i + \eta_i$$

- Hypotesen er derfor givet ved:

$$H_0: \phi_1 = 0$$

$$H_A: \phi_1 \neq 0$$

- Under MLR.1-5 og H_0 følger t-statistikken asymptotisk en t-fordeling. Vi beregner $t = 9.03$ hvilket er langt over den kritiske værdi for konventionelle signifikansniveauer. Vi forkaster derfor nulhypotesen, og finder en signifikant og positivt sammenhæng mellem $\hat{u}_i^2$ og $m2_i$. Det indikerer at variansen på fejllidet vokser i størrelsen på lejligheden.

`regress uhat2 m2`

uhat2	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
m2	.0057413	.000636	9.03	0.000	.0044932	.0069893
_cons	-.1616995	.058268	-2.78	0.006	-.276043	-.047356

5) Opsamling - Stata

5. Suppose that the variance of the error term in model (2) is:

$$\text{Var}(u_i|X) = \sigma^2 h(X) = \sigma^2 m2_i$$

where $h(X) = m2_i$. Does this type of heteroskedasticity correspond to your previous finding? Estimate model (2) using the Weighted Least

- Opg. 4 indikerer at vi ikke kan forkaste at variansen på fejllidet og $m2_i$ er ligefrem proportionale (lineær sammenhæng). Antagelsen af $\text{var}(u_i|X) = \sigma^2 m2_i$ passer ind med dette.
- Vi kan implementere WLS under antagelse at $h(X) = m2_i$ ved at specificere vægten $(\frac{1}{h(x)})$ til sidst i kommandoen:

```
regress price KbhK KbhN KbhV m2 sqm2 rooms toilets [aweight=1/m2]
```

5) Opsamling - Stata

- Både OLS og WLS estimatoren er unbiased og konsistent under MLR.1-4.
- Det giver derfor god mening at estimererne er tæt på hinanden.
- Vi ser at standardfejlene er tæt på at være ens (hverken generelt større eller mindre på tværs af metoderne).
- Hvis der er heteroskedasticitet og vi har specificeret denne rigtig er WLS estimatoren efficient (teoretisk burde standardfejl være mindre - i praksis afhænger det også af stikprøven).
- Bemærk dog også at måden vi beregner standardfejl på i OLS er forkert (korrigerer ikke for heteroskedasticitet).
- Det er derfor ikke så meningsfuldt at sammenligne på tværs af metoderne. Vi kunne med fordel sammenligne WLS standardfejl med OLS hvor det bruges robuste standardfejl.

	(1)	(2)
VARIABLES	OLS	WLS
KbhK	0.192*** (0.0479)	0.207*** (0.0434)
KbhN	-0.213*** (0.0489)	-0.180*** (0.0414)
KbhV	-0.0646 (0.0563)	-0.0457 (0.0503)
m2	0.0371*** (0.00206)	0.0359*** (0.00224)
sqm2	-0.0670*** (0.00780)	-0.0647*** (0.00977)
rooms	-0.0547* (0.0286)	-0.0209 (0.0279)
toilets	0.0641 (0.0884)	0.0478 (0.107)
Constant	-0.193 (0.137)	-0.197 (0.150)
Observations	988	988
R-squared	0.674	0.662

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling - Stata

6. Calculate the WLS estimates manually by dividing all the variables in model (2) by $\sqrt{h(X)} = \sqrt{m2_i}$.

- For at estimere modellen med WLS manuelt skal vi dividere alle vores variable med $\sqrt{m2_i}$
- Dette gøres nemmest gennem et loop:

```
gen w=1/sqrt(m2)
foreach x in price KbhK KbhN KbhV m2 sqm2 rooms toilets {
    gen `x' star=`x'*w
}
```

- Når vi har alle vores transformererede variable kan vi estimere den transformererede model:

```
reg pricestar KbhKstar KbhNstar KbhVstar m2star sqm2star roomsstar toiletsstar w, nocon
```

- Bemærk vi skal huske at ekskludere konstantleddet (,nocon) da vores transformererede konstantled ($\frac{1}{\sqrt{m2_i}}$) indgår i stedet.

5) Opsamling - Stata

6. Calculate the WLS estimates manually by dividing all the variables in model (2) by $\sqrt{h(X)} = \sqrt{m2_i}$.

```
reg pricestar KbhKstar KbhNstar KbhVstar m2star sqm2star roomsstar toiletsstar w, nocon
```

Source	SS	df	MS	Number of obs	=	988
Model	60.9124188	8	7.61405235	F(8, 980)	=	2204.72
Residual	3.3844551	980	.003453526	Prob > F	=	0.0000
				R-squared	=	0.9474
				Adj R-squared	=	0.9469
Total	64.2968739	988	.065077808	Root MSE	=	.05877

pricestar	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
KbhKstar	.2068564	.0434015	4.77	0.000	.1216858	.292027
KbhNstar	-.1802255	.0413728	-4.36	0.000	-.261415	-.0990361
KbhVstar	-.045675	.050306	-0.91	0.364	-.1443949	.0530449
m2star	.0358836	.0022423	16.00	0.000	.0314833	.0402839
sqm2star	-.064713	.0097687	-6.62	0.000	-.0838829	-.0455431
roomsstar	-.0209116	.0279276	-0.75	0.454	-.0757164	.0338932
toiletsstar	.0477806	.1068408	0.45	0.655	-.1618824	.2574437
w	-.1970393	.1503233	-1.31	0.190	-.4920319	.0979533

- Koefficienter og standardfejl er identiske med det vi fandt i opg. 5.
- Bemærk dog at R^2 og F -testen for hele modellen er forskellige (skyldes bl.a. at modellen ikke har et konstantled så den almindelige måde at beregne R^2 og F -test er misvisende). Disse bør derfor ikke bruges. Benyt i stedet dem Stata rapporterer når WLS beregnes automatisk.

5) Opsamling - Stata

7. The OLS estimator is unbiased under MLR.1– MLR.4, but without MLR.5, standard inference is not valid. In the presence of heteroskedas-

- Indtil nu har vi beregnet OLS estimatorens varians ud fra formlen:

$$\text{var}(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

- Hvor $\hat{\sigma}^2 = \frac{\sum_i^n \hat{u}_i^2}{n-(k+1)}$. Udledningen af denne byggede på at $\text{var}(u|X) = \sigma^2 I_n$. Hvis denne antagelse er brudt er ovenstående formel ikke længere korrekt. I stedet kan vi beregne variansen ud fra følgende formel:

$$\widehat{\text{var}}(\hat{\beta}) = n(X'X)^{-1}X'\hat{\Sigma}X(X'X)^{-1}$$

- Hvor $\hat{\Sigma} = \frac{1}{n} \sum_i^n \hat{u}_i^2 x_i x_i'$. Denne måde at beregne standardfejl er valid under arbitrære former for heteroskedasticitet samt homoskedasticitet. De robuste standardfejl kan både være større og mindre end klassiske standardfejl (afhænger af formen for heteroskedasticiteten).
- Vi kan også bruge disse robuste standardfejl til at danne robuste t- og F-statistikker.
- Hvorfor bruger vi ikke altid robuste standardfejl?
 - i) I praksis er det ofte tilfældet at man altid rapporterer robuste standardfejl
 - ii) Robuste standardfejl er dog kun asymptotiske valide ($\hat{\Sigma}$ er konsistent men biased). Ved små samples kan robuste standardfejl derfor være misvisende.
 - iii) Pga. dette er det også bedre at bruge de klassiske standardfejl hvis vi er sikre på at MLR.5 er overholdt.

5) Opsamling - Stata

7. The OLS estimator is unbiased under MLR.1– MLR.4, but without MLR.5, standard inference is not valid. In the presence of heteroskedas-

- Vi implementerer robuste standardfejl i Stata ved at skrive ", robust" efter regressionen:

```
regress price KbhK KbhN KbhV m2 sqm2 rooms toilets, robust
```

- Estimererne forbliver uændrede. Det er vores standardfejl der er ændret. I dette tilfælde er de overordnet set en smule større.
- De t-statistikker der automatisk rapporteres er nu robuste og valide under MLR.1-4 + $n \rightarrow \infty$.
- Vi kan også beregne valide F-statistikker ved som sædvanlig at skrive fx

```
test rooms toilets
```

Lige efter regressionen foretaget med robuste standardfejl

	(1)	(2)
VARIABLES	Klassiske	Robuste
KbhK	0.192*** (0.0479)	0.192*** (0.0529)
KbhN	-0.213*** (0.0489)	-0.213*** (0.0412)
KbhV	-0.0646 (0.0563)	-0.0646 (0.0568)
m2	0.0371*** (0.00206)	0.0371*** (0.00283)
sqm2	-0.0670*** (0.00780)	-0.0670*** (0.0118)
rooms	-0.0547* (0.0286)	-0.0547 (0.0335)
toilets	0.0641 (0.0884)	0.0641 (0.133)
Constant	-0.193 (0.137)	-0.193 (0.183)
Observations	988	988
R-squared	0.674	0.674

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

5) Opsamling - Stata

7. The OLS estimator is unbiased under MLR.1– MLR.4, but without MLR.5, standard inference is not valid. In the presence of heteroskedas-
- Først tester vi følgende hypotese for parameteren på $KbhV_i$:

$$H_0: \delta_3 = 0$$

$$H_A: \delta_3 \neq 0$$

- Med robuste standardfejl ved vi at t-statistikken asymptotisk følger en normalfordeling under MLR.1-4 og H_0 . Beregner vi t-statistikken ud fra tabellen fra forrige slide finder vi at denne er -1,14 hvilket er væsentligt under den kritiske værdi på et 5% signifikansniveau (-1,96), så vi kan ikke forkaste at de forventede priser på Østerbro og Vesterbro er ens (alt andet lige).
- Vi tester derefter om:

$$H_0: \beta_3 = \beta_4 = 0$$

$$H_A: H_0 \text{ ikke sand}$$

- Vi bruger en robust F-test til dette og finder at:

```
test rooms toilets
```

```
F( 2, 980) = 1.70
```

```
Prob > F = 0.1828
```

- Så vi kan ikke forkaste nulhypotesen om at antal værelser og antal toiletter ikke har en effekt på den forventede pris (alt andet lige).

5) Opsamling - Stata

8. Calculate the Breusch-Pagan test for the original model (1), where the dependent variable is $\log(prices)$. Are there signs of heteroskedasticity?

- Vi går nu tilbage til den fortrukne model fra PS6:

$$\log(price_i) = \beta_0 + \delta_1 KbhK_i + \delta_2 KbhN_i + \delta_3 KbhV_i + \beta_1 m2_i + \beta_2 m2_i^2 + \beta_3 rooms_i + \beta_4 toilets_i + \epsilon_i$$

- Vi foretager en BP af denne model:

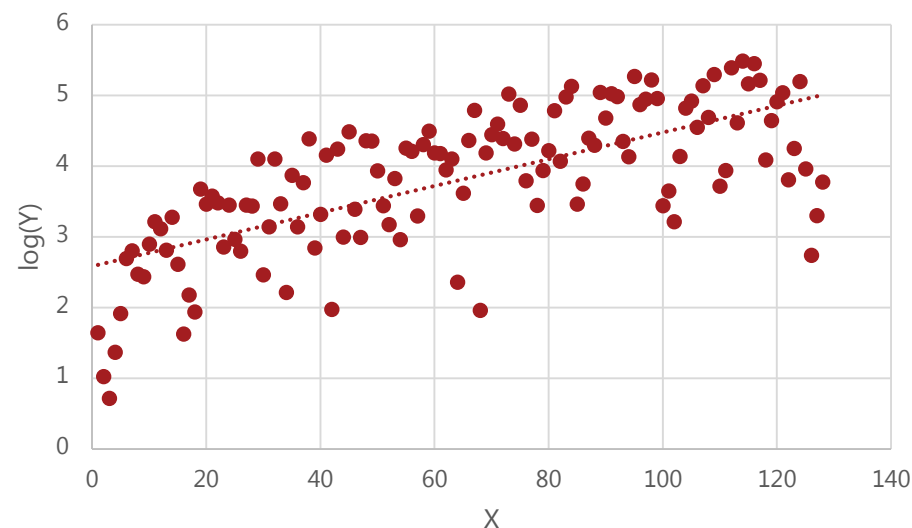
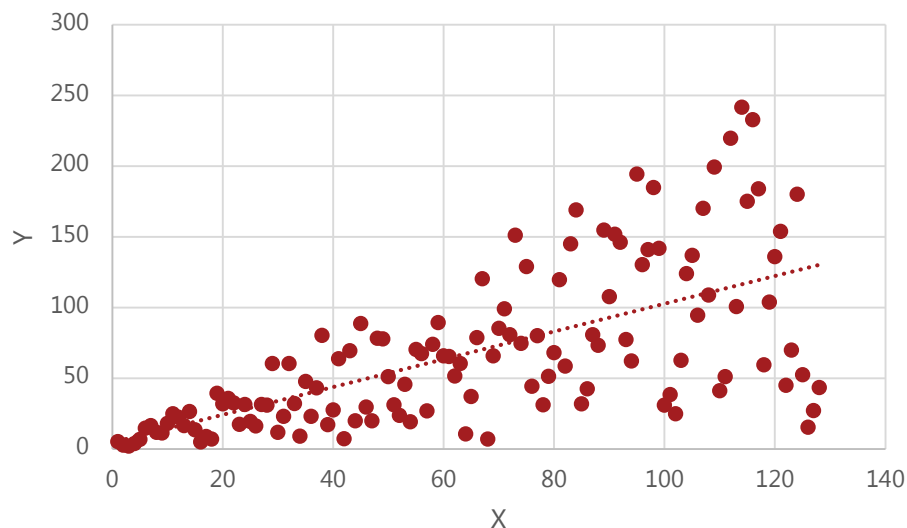
```
regress logprice KbhK KbhN KbhV m2 sqm2 rooms toilets
predict uhat, residuals
gen uhat2=uhat*uhat
regress uhat2 KbhK KbhN KbhV m2 sqm2 rooms toilets
gen BP=e(N)*e(r2)
gen c=invchi2tail(7,0.05)
su BP c
```

Variable	Obs	Mean	Std. Dev.	Min	Max
BP	988	4.01891	0	4.01891	4.01891
c	988	14.06714	0	14.06714	14.06714

- Vi kan ikke længere forkaste H_0 på et 5% signifikansniveau, og finder derfor empirisk belæg for homoskedasticitet.

5) Opsamling - Stata

- Log-transformationer kan altså hjælpe på problemer med heteroskedasticitet. Det ændrer dog også på vores fortolkning af parametrene (i modsætning til WLS og robuste standardfejl)
- Log-transformationer hjælper typisk når variansen på vores afhængige variabel vokser med størrelsen på de forklarende variable.
- Ved at tage logaritmen trækker vi store værdier tættere sammen:



- Omvendt kan logaritme transformationer forværre problemet hvis man har mange værdier mellem 0 og 1 (værdier eksploderer i negativ retning)

Tak for i dag

- Næste gang går vi igennem OO2
- Derefter skal vi se nærmere på Instrumental Variable