Multipel lineær regression.

Regression analyse (engelsk *regress* betyder at falde fra en højde ind mod midten) dækker over et bredt spektrum fra den simple lineære regression til den flervariable analyse også kaldet den multipel regression. I den multipel regression, hvor der indgår flere uafhængige variabler er målet at estimere hældningskoefficienterne til de uafhængige variabler, sådan at man kan forudsige værdien på den afhængige variabel. En passende model til den multiple regression skal bestemmes, og den kan både være et polynomium, logaritmisk osv. men i dette tilfælde bliver det en lineær model, deraf navnet multipel lineær regression.

I vejledningen anvendes der CAS værktøjet Excel - dataanalyse eventuelt med brug af Megastat. Multipel regression er i princippet lige så let at behandle som simpel lineær regression - men tolkningen af Excel output er en videnskab for sig selv. Dette dokument indeholder en guide i hvordan dette gøres.*

** Maple kan også anvendes til multipel lineær regression.*

Introduktion til multipel regression.

$$Model: \quad y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

Den estimerede model og dermed de estimerede koefficienter $\hat{\beta}_i$ skrives som vist herunder. Indeks k angiver, at der er op til k uafhængige variabler, og i er indekset hvor der gælder at $0 \leq i \leq k$.

$$Y = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

Data er givet i excel-ark. Regressionsanalysen anvendes og man får et output.

Hypotese. F-test

$$H_o: \ \beta_1 = \beta_2 = .. = \beta_k \ \ i \geq 1$$

$$H_1: \ mindst\ én\ af\ \beta_i \neq 0$$

Forkast nulhypotesen hvis -

$$F_{tabel} < F_{obs}$$

Hypotese. t-test

$$H_o: \ \beta_i = 0 \ \ i \geq 1$$

$$H_1: \ \beta_i \neq 0$$

Forkast nulhypotesen hvis -

$$t = \frac{b_i - \beta_i}{s_b} > t_{tabel}$$

Hold øje med justeret R^2. Årsagen til at denne værdi bruges er, at denne tager hensyn til antal variabler, og at determinationskoefficienten R^2 vil gå mod 1, hvis de uafhængige antal variabler går mod uendelig, eller antal observationer pr søjle er lig med antal uafhængige variabler - som kaldes *overfitting*

Hvis en af variablerne i t-testen bliver accepteret af nulhypotesen, fjernes denne. Husk aldrig at fjerne mere end én variabel ad gangen. Så starter man forfra igen.

Når alle hældningskoefficienter er signifikante til den multipel lineære regressionsmodel, så udføre man den sidste test for at undgå multikollinearitet. Multikollinearitet betyder at 2 variabler kan stjæle forklaring fra hinanden, dette skal undgås.

Excel udarbejder en korrelationsmatrice. Hypotesen korrelationstest,

$$H_o: \ \rho_i = 0 \ \ i \geq 1$$

$$H_1: \ \rho_i \neq 0$$

Accepter nulhypotesen hvis -

$$t = \frac{r}{\sqrt{1 - r^2}} \cdot \sqrt{n - (k - 1)} < |t_{tabel}|$$

God regel at sætte absolut værdien på korrelationskoefficienten til at være mindre end 0.3 (Robusthed).

Når alle test er gennemført, bør ligningen til den multipel lineære regressionsmodel være fundet.

På Samsø har ferieudlejning fået den opgave at prissætte sommerferiehuse for 2019. Følgende data forelægger fra 2018. Se datafil, hvor Obs svarer til søjle A.

1		Sommerferie 2019				
2	Obs	pris	kvm	sovep	kyst	komfort
3	1	3930	65	5	400	2
4	2	4340	68	8	500	3
5	3	5240	75	8	200	5
6	4	4421	62	5	200	4
7	5	4421	65	6		

Målet er at opbygge en forklarende model "pris pr. uge", hvor der indgår et passende antal variabler. Værktøjet til dette hedder den multipel lineær regressionsmodel. I dette dokument vil der blive gennemgået, hvilken værktøj eleven skal kunne beherske og tolke på. Desuden vil der være tekstboks til undervisere, som forklare teorien lidt dybere bag den multipel lineær regressionsanalyse.

For at kunne forklare prisen pr uge til et givent sommerhus på Samsø, hvor følgende variabler kommer i betragtning {kvm, sovep, kyst og komfort}, skal der testes for samvariation også kaldet F-test, dvs. om det vil være muligt at finde en årsagssammenhæng. Der testes for signifikante hældninger t-test, og til sidst en test for multikollinearitet, da det skal undgås at flere uafhængige variabler afhænger af hinanden.

Først laves en fuldstændig analyse, hvor der ikke udelukkes nogen variabler. I Excel går man ind i dataanalyse, og

> Inden man starter selve analysen, er det vigtigt at gøre sig klart om de uafhængige variabler kan forklare den afhængige variabel. I datafilen er der 4 variabler (søjle C, D, G og F), der skal forklare den ugentlige pris på et sommerhus. Der skal være en forsigtighed i behandling af data, da et sommerhus med mange kvm også vil have flere soverum og derfor også plads til flere personer, dette kan medfører at der er kollinearitet.
>
> Men i luksusklassen af sommerhuse kan der også være mange kvm og få soverum, så man skal passe på. Alt dette skal man tage i sin betragtning, før analysen begynder.

videre til regression - klik på denne. Sidst i dokumentet er der en vejledning i at installere dataanalyse.

<u>Modelforudsætninger:</u>

1. $Model: \quad y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$
2. $Sum\ af\ alle\ restled\ er\ nul. \quad E(\varepsilon) = 0$
3. $\varepsilon_i\ er\ normalfordelt \sim N(0, \sigma^2)\ (visuelt)$
4. $\sigma_\varepsilon^2\ er\ \ konstant\ (visuelt)$
5. $Uafhængighed\ \ mellem\ \ \varepsilon_i\ og\ \ \varepsilon_{i+1}\ (visuelt)$
6. $Ingen\ Multikollinearitet$

> Baggrunden til disse betingelser findes i, at den afhængige variabel Y også opfører sig stokastisk, og det er forventelig at dets værdier er normalfordelt med konstant varians.

Ligningen til den multipel regressionsmodel skrives som.

$$Y = \hat{\beta}_o + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

De estimerede hældningskoefficienter har symbolet - $\widehat{\beta_\iota}$. Der vil nu komme en gennemgang af hvert punkt, så man kan se hvordan analysen fungere. Det er ikke en forudsætning, at eleven kan gå i dybden på alle disse seks punkter.

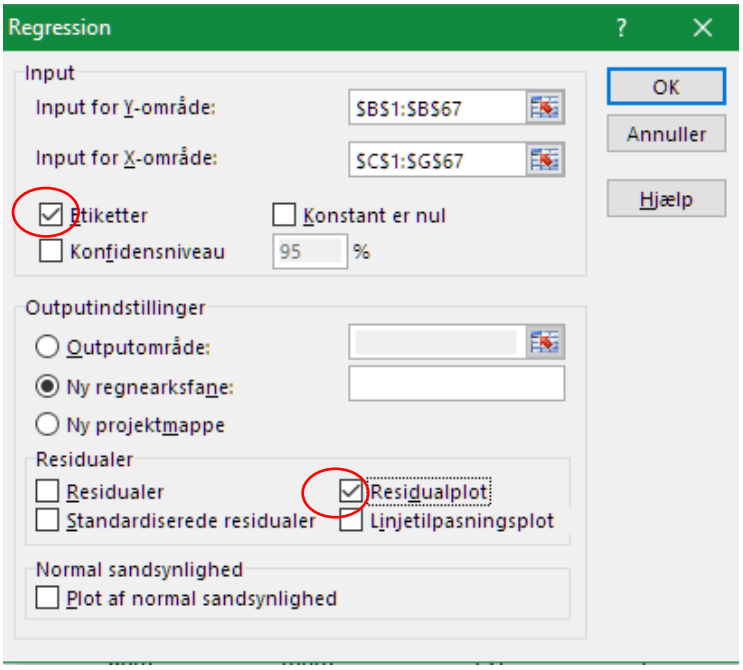
Forklaring til de seks punkter til modelforudsætninger for multipel lineær regression.

1. Hypotesetest for samvariation. F-test - estimerede hældninger (undtagen $\beta_o - skæring\ med\ y\ akse$) testes, hvor alternativ hypotesen er at mindst en estimeret hældning er signifikant,
2. Hypotesetest for signifikante hældninger. Dette kaldes en t-test, dvs hvor den estimerede hældning har en 95% konfidensinterval, der ikke indeholder nul.
3. Middeltallet for residualled (restleddene) skal være nul, og disse restled skal være normalfordelt omkring nullinjen.
4. Restled ε er afstanden fra observationspunkt til modellen. Krav er at restleddene skal vise varianshomogenitet, dvs konstant varians.
5. Uafhængighed mellem restleddene, det betyder at der ikke må tegnes et mønstre omkring nullinjen Variansen på restled skal være konstant er let at se visuelt, ved at der ikke må være et mønstre omkring nullinjen.
6. Multikollinearitet, undersøges normalt med programmet "korrelationsanalyse" i Excel, men der er kun få programmer der også tester om man skal fjerne nogle af variablerne (Megastat). En tommelfinger regel siger at hvis værdien numerisk er større end 0,25 - 0,30 så bør en af variablerne fjernes (Pearsons formel).

Her vises hvordan man kommer gennem disse seks punkter. Det er vigtigt at have megen tålmodighed og huske på, at der er tilfælde, hvor det kan gå meget let, hvorimod andre gange skal man gennem alle punkter adskillige gange.

Guiden på side 2 viser, hvad man skal holde øje med - og hvor der er mulige faldgruber.

Punkt 1. Start med at vælge data henholdsvis y-værdier og x-værdier. Husk at sætte flueben på "Etiketter" og Residualplot.



Klik på OK, og man får en ANAVA {ANalyse Af VArians}. Her er en gennemgang af alle modelforudsætninger.

Regressionsstatistik	
Multipel R	0,844
R-kvadreret	0,712
Justeret R-kvadreret	0,692
Standardfejl	881,975
Observationer	63

Det første man skal være opmærksom på er den lave "Signifikans F" værdi. Hvis denne værdi er mindre end 0.05 betyder det, at nogle af de uafhængige variabler der kan bruges i vores model. Men samtidig fortæller *t-test* at variablen *kyst* ikke er signifikant og dermed skal kyst fjernes da p-værdien for denne er større end 0.05 (0.63)

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	4	111609766,1	27902442	35,86984	4,48971E-15
Residual	58	45117058,13	777880,3		
I alt	62	156726824,2			

	Koefficienter	*Standardfejl*	*t-stat*	*p-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	-693,674	501,643	-1,383	0,172	-1697,821	310,473
kvm	25,593	4,627	5,531	0,000	16,330	34,855
sovep	295,288	95,458	3,093	0,003	104,208	486,369
kyst	-0,062	0,127	-0,484	0,630	-0,316	0,193
komfort	360,183	100,214	3,594	0,001	159,582	560,783

På næste side er ANAVA´en forklaret i detaljer.

Dette er det første Excel-output man møder, og den første opgave er at undersøge om der er nogle hældningskoefficienter er der signifikante. Hvis der nu skulle ske det, at *Signifikans F* er større end 0.05, vil man være i den situation at testen stopper her, da nul-hypotesen accepteres - se F-test for samvariation.

Hypotesetest for samvariation - F-test.

$$H_o\colon\ \beta_1 = \beta_2 = \cdots = \beta_k = 0\ \ ingen\ forklarende\ variabler$$

$$H_1\colon\ ikke\ alle\ \beta'er\ er\ nul\ \ \ Der\ er\ forklarende\ variabler.$$

Output fra Excel viser, at "Signifikans F værdien" er lig med $4.89\ 10^{-15} \ll 0.05$ hvilket betyder at nulhypotesen forkastes, og at der er nogen variabler der vil kunne forklare prisen på sommerhuset.

Punkt 2. Vi skal nu finde frem til, hvilken variabler der kan bruges.

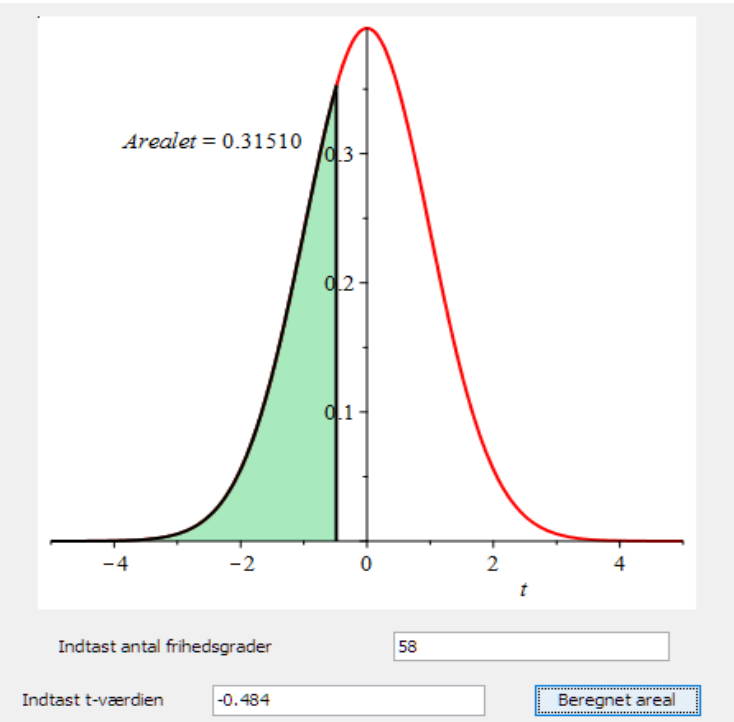
Hypotesetest for hældningskoefficienter - t-test.

Her kommer første betingelse se side 2 - hældninger der krydser 0 i konfidensintervallet på 95% skal fjernes. Årsagen er at disse estimerede hældningskoefficienter ikke er signifikante til vores model. Husk det er forbudt at fjerne flere uafhængige variabler ad gangen, trods det at der kan være flere variable, der ikke er signifikante.

Eksempel på brug af *t-test* når variablen *kyst* forkastes. Der laves en individuel t-test for hvert hældningskoefficient, og målet er at disse skal være signifikante. Umiddelbart kan der af Excel output ses at variablen *kyst* acceptere H_o.Beregning:

$$t = \frac{\hat{\beta}_{kyst} - \beta_{kyst}}{s_{kyst}} = \frac{-0.062 - 0}{0.127} = -0.484$$

Der er et krav på at *p-værdien* < 0.05 for at forkaste nulhypotesen. *P-værdien* bestemmes her. Da alle test er 2-sidet, skal arealet herunder også multipliceres med 2.



Da det er en 2-sidet test vil *p-værdien* være det dobbelte af det viste areal, dvs $2 \cdot 0.315 = 0.63$. Dette ses også i tabellen. Yderlig passer den lave værdi fra "*Signifikans F*" ikke med den høje værdi fra *"kyst"*, dette er også en af grunden til at fjerne *"kyst"*.

Samtidig notere man;

$$Multipel\ R = 0{,}844 \quad og \quad Justeret\ R^2 = 0{,}692$$

Huskeregel: Jo flere variabler jo højere værdi på "*Multipel R*", som derfor går mod værdien 1. Filosofisk kendes dette fra, at hvis man har nok af søforklaringer kan alt forklares, og dette skal vi undgå her. Målet er at hæve "$Justeret\ R^2$" værdi og samtidig undgå at "*Multipel R*" ændres meget, når der fjernes mulige forklarende variabler. Det er svært. Nu fjerne vi variablen kyst og ser hvad der sker med hældnings KI og på "*Multipel R*" sammenholdt med " $Justeret\ R^2$ værdi.

Regressionsstatistik	
Multipel R	0,8432
R-kvadreret	0,7110
Justeret R-kvadreret	0,6963
Standardfejl	876,2349
Observationer	63

Avanceret tolkning. Når der fjernes eller tilføres variabler, bør man holde øje med SK (Summen af fejl leddenes Kvadrater). Værdien på SK bør enten holdes konstant eller endnu bedre falde i størrelsen, når der testes på de uafhængige variabler. Hvis værdien er høj, kan det også betyde at der er outliniers.

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	3	111427357,2	37142452	48,37595	6,60137E-16
Residual	59	45299467,02	767787,6		
I alt	62	156726824,2			

	Koefficienter	*Standardfejl*	*t-stat*	*p-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	-685,550	498,099	-1,376	0,174	-1682,244	311,144
kvm	25,280	4,552	5,553	0,000	16,171	34,389
sovep	280,968	90,171	3,116	0,003	100,535	461,401
komfort	382,478	88,433	4,325	0,000	205,524	559,431

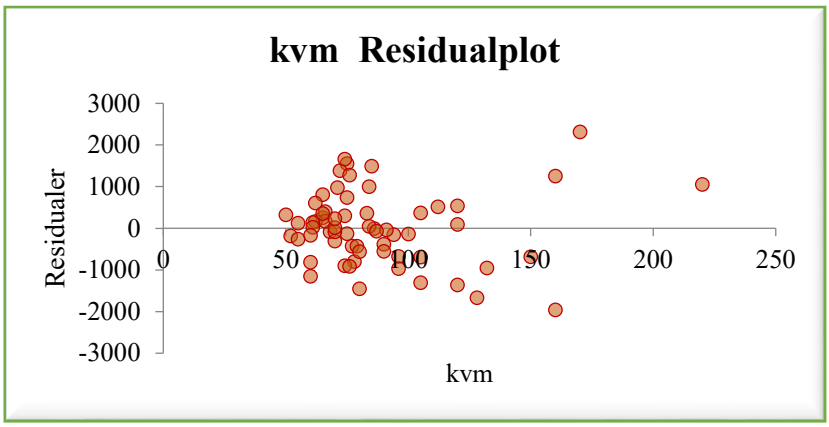
F-testen og de individuelle t-test kan nu accepteres som signifikante, det vil sige, at nul-hypotesen forkastes og følgende variabler vil være forklarende. Bemærk også at når "*kyst*" fjernes, så får "*komfort*" mere forklaringskraft - kortere hældnings interval.

Men hvilken betydning har dette fået på -

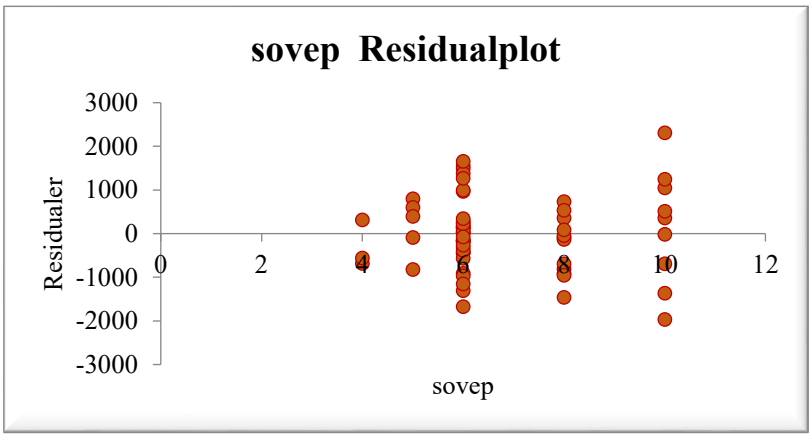
$$Multipel\ R = 0{,}843\ (før\ 0{,}844) \quad og \quad Justeret\ R^2 = 0{,}696\ (før\ 0{,}692)$$

Det er meget lille ændring, hvilket betyder at der er nogen variabler der tager forklaringskraft fra andre – punkt 6.

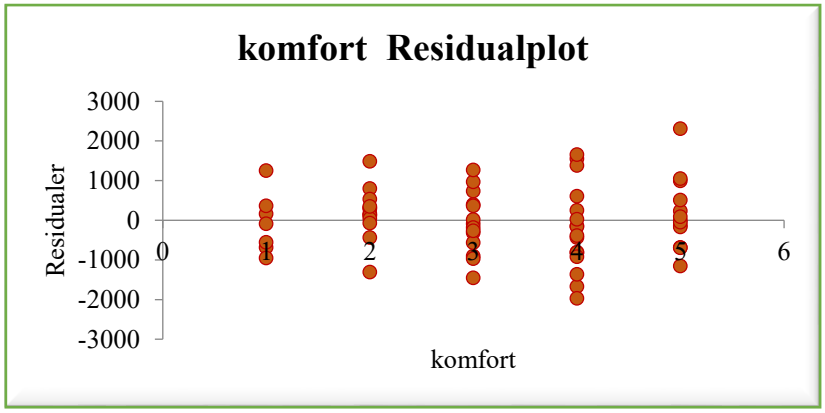
Punkt 3-5. Graferne fra residual plottene er vist på næste side, hvor disse trin visuelt undersøges om de er brugbare eller ej. Man kan også gå til trin 6 og dernæst tilbage til trinene 3 - 5. Det skal bemærkes, at trinene fra 3 - 5 normalt testes i et CAS værktøj, men det kræver en langt større viden end vi skal bruge her.



Punkterne 3 til 5 som er modelforudsætningerne, skal visuelt undersøges på disse 3 residualplots. Det bliver en vurderingssag om man acceptere forudsætningerne. På dette residualplot for kvm. kan man noterer sig, at der er nogle sommerhuse, der har mange kvadratmeter, og dernæst stille sig det spørgsmål om dette er repræsentativt.



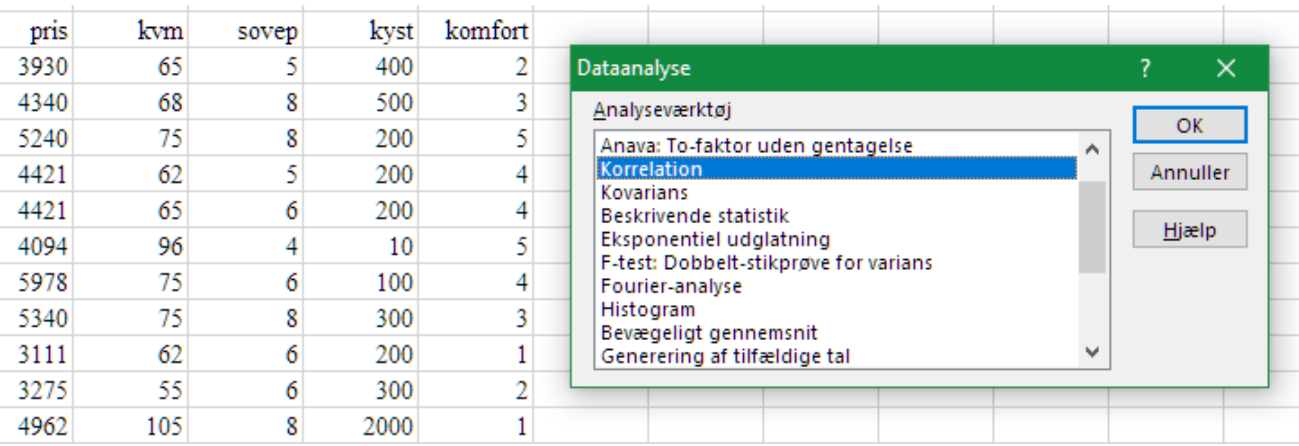
Grafen for residualplottet til sovp. har muligvis et problem omkring konstant varians. Igen er det de store sommerhuse, der giver de høje afvigelser.



Grafen til residualplottet for komfort, er i dette tilfælde det plot, der faktisk opfylder alle de 3 punkter.

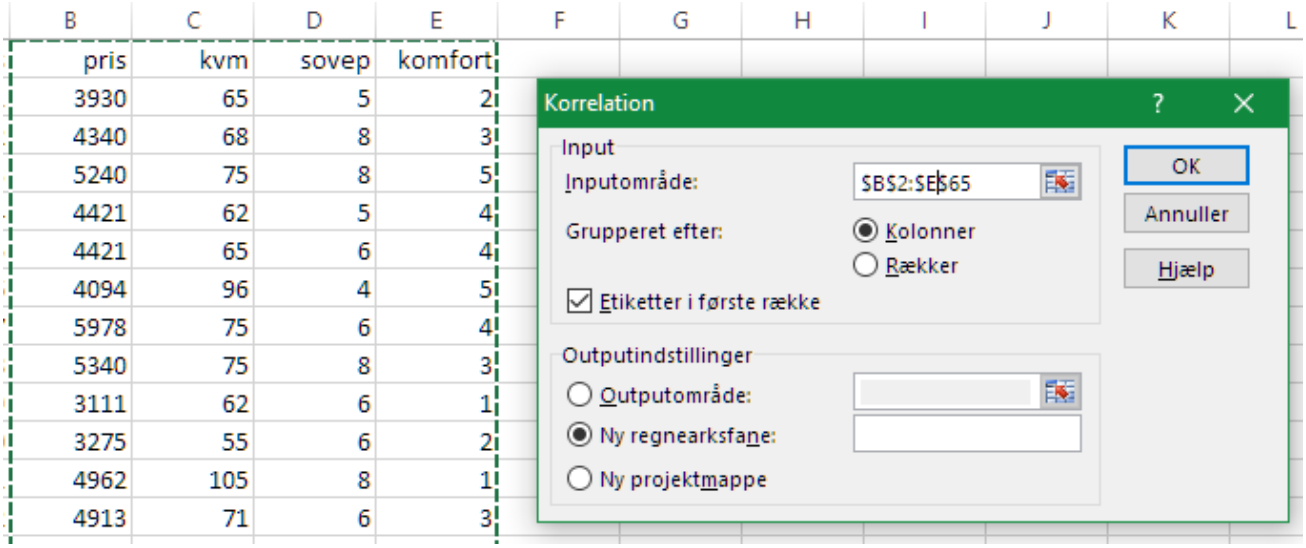
Det bliver en vurderingssag, hvornår man skal acceptere model forudsætningerne, men det er sjældent, at plottene er perfekte. Alligevel ses der ikke nogen umiddelbare brud på betingelserne. Husk - det er meget synligt på graferne, hvis de ikke kan opfylde betingelserne.

Punkt 6. Den sidste modelforudsætning er testen for ” *multikollinearitet”*. Bestemmes via den såkaldte korrelationsmatricen. Man går ind i Excel og vælger alle data inklusiv data benævnelse.



pris	kvm	sovep	kyst	komfort
3930	65	5	400	2
4340	68	8	500	3
5240	75	8	200	5
4421	62	5	200	4
4421	65	6	200	4
4094	96	4	10	5
5978	75	6	100	4
5340	75	8	300	3
3111	62	6	200	1
3275	55	6	300	2
4962	105	8	2000	1

Husk at fjerne kyst-data



pris	kvm	sovep	komfort
3930	65	5	2
4340	68	8	3
5240	75	8	5
4421	62	5	4
4421	65	6	4
4094	96	4	5
5978	75	6	4
5340	75	8	3
3111	62	6	1
3275	55	6	2
4962	105	8	1
4913	71	6	3

Excel giver dette resultat af korrelationstesten. Selve tabellen kaldes korrelationsmatricen.

	pris	*kvm*	*sovep*	*komfort*
pris	1			
kvm	0,726851	1		
sovep	0,697863	0,641836	1	
komfort	0,434789	0,09479	0,248338	1

Når værdierne i korrelationsmatricen aflæses, skal man starte med søjlen *”pris”*. Prisen er selvfølgelig korreleret med pris, og giver naturligt også værdien 1. Men det ses også at *”pris”* er pænt korreleret med de uafhængige variabler *kvm, sovep* og *komfort.* De andre søjler skal nu sammenlignes med rækkerne, og man ser en rimelig stærk korrelation mellem *”sovp”* og *”kvm”*. Men hvor langt skal man ned i værdien på korrelationskoefficienten, før det ikke er signifikant at 2 variabler er korreleret? Her skal man bruge Pearsons korrelationsformel.

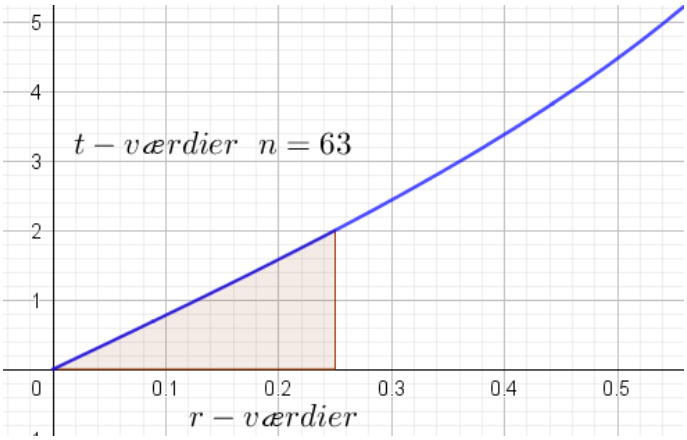
Pearson korrelationsformel.

$$t(n,r) = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-3} = \frac{0.25}{\sqrt{1-0.25^2}} \cdot \sqrt{63-3} = 2$$

$$t \cong 2 \;\; svare \;\; til\; test\; på\; signifikansniveau\; 5\% \;\; mod\; korrelation.$$

$$3 - tallet\; er\; frihedsgrader\; fra\; ANAVA'en.$$

Med andre ord hvis vi ser korrelationskoefficienter større end 0.25, bør man overveje hvilke variabler der skal fjernes. Graf til *t(n,r)*.



Grafen viser t-værdien som funktion af korrelationskoefficienten r, når *n = 63*. Bemærk hvis stikprøvestørrelsen n bliver mindre, så kan r tillade sig at blive større, hvis betingelsen for t-værdien maksimalt må være 2.

<u>Bemærkning til korrelationsanalysen</u>. Hvis man har installeret Megastat (licens) i Excel, vil korrelationsanalysen være meget lettere at tolke på. Dette skyldes,at Pearsons korrelationstest er indbygget i Megastat. Testen er 2-sidet og er vist her

$$H_o: \qquad \rho = 0 \quad ingen\;\; korrelation$$

$$H_1: \qquad \rho \neq 0 \quad korrelation$$

Excel output hvor Megastat er anvendt.

Korrelation Matrix

	pris	*kvm*	*sovep*	*kyst*	*komfort*
pris	1,000				
kvm	,727	1,000			
sovep	,698	,642	1,000		
kyst	,142	,341	,348	1,000	
komfort	,435	,095	,248	-,337	1,000

63 sample size

± ,248	critical value of r .05 (to-sidet)
	critical value of r .01 (to-sidet)

Vi ser at sovep. og kvm. er korreleret med værdien 0,641836, dette ses også i Excel output uden brug af Megastat. Jeg vælger, at fjerne sovep, da denne værdi er mindst korreleret med pris i forhold til kvm. Bemærk kritiske værdier på signifikansniveau 0.05 er numerisk 0.248.

Dermed kan vi gå til det endelige resultat - pris er forklaret med kvm og komfort.

Resultatet

Regressionsstatistik	
Multipel R	0,814
R-kvadreret	0,663
Justeret R-kvadreret	0,652
Standardfejl	937,674
Observationer	63

ANAVA

	fg	*SK*	*MK*	*F*	*Signifikans F*
Regression	2	103972893,7	51986447	59,1271	6,50646E-15
Residual	60	52753930,52	879232,2		
I alt	62	156726824,2			

	Koefficienter	*Standardfejl*	*t-stat*	*p-værdi*	*Nedre 95%*	*Øvre 95%*
Skæring	202,491	437,150	0,463	0,645	-671,939	1076,920
kvm	34,375	3,738	9,195	0,000	26,897	41,852
komfort	450,156	91,734	4,907	0,000	266,660	633,653

Nu er alle betingelser opfyldt, og det er specielt antal sovepladser der har haft en lille betydning på $Multipel\ R$ og $Justeret\ R^2$, da begge er faldet lidt i værdien. Dette er prisen "matematisk" for at opfylde modelforudsætningerne.

Faktisk er det indlysende, at antal soveværelser og kvm må være korreleret.

$$Multipel\ R = 0{,}814\ (før\ 0{,}843)\ \ og\ \ \ Justeret\ R^2 = 0{,}652\ (før\ 0{,}696)$$

Konklusion. Bedste model til at forklare prisen på sommerhus bliver da,

$$pris\ pr\ uge = 34.4 \cdot kvm + 450.2 \cdot komfort + 202.5$$

Eksempel. En kunde spørger efter et nyere sommerhus på 95 kvm og med middel komfort på 3. Hvilken pris kan der forventes pr uge for dette sommerhus?

$$pris\ pr\ uge = 34.4 \cdot 95 + 450.2 \cdot 3 + 202.5\ \cong kr\ 4820\ pr\ uge$$

Hvis man på et sådant sommerhus kan finde en lavere pris pr uge, er det meget vel et godt tilbud.

MÅRUP ØSTERSTRAND NORD

DKK 4.955,00

113 - Musvitvej 2

Personantal: 4 Lille hyggeligt velholdt feriehus på en dejlig lukket grund med terrasse som er delvist overdækket. Tæt på god badestrand.

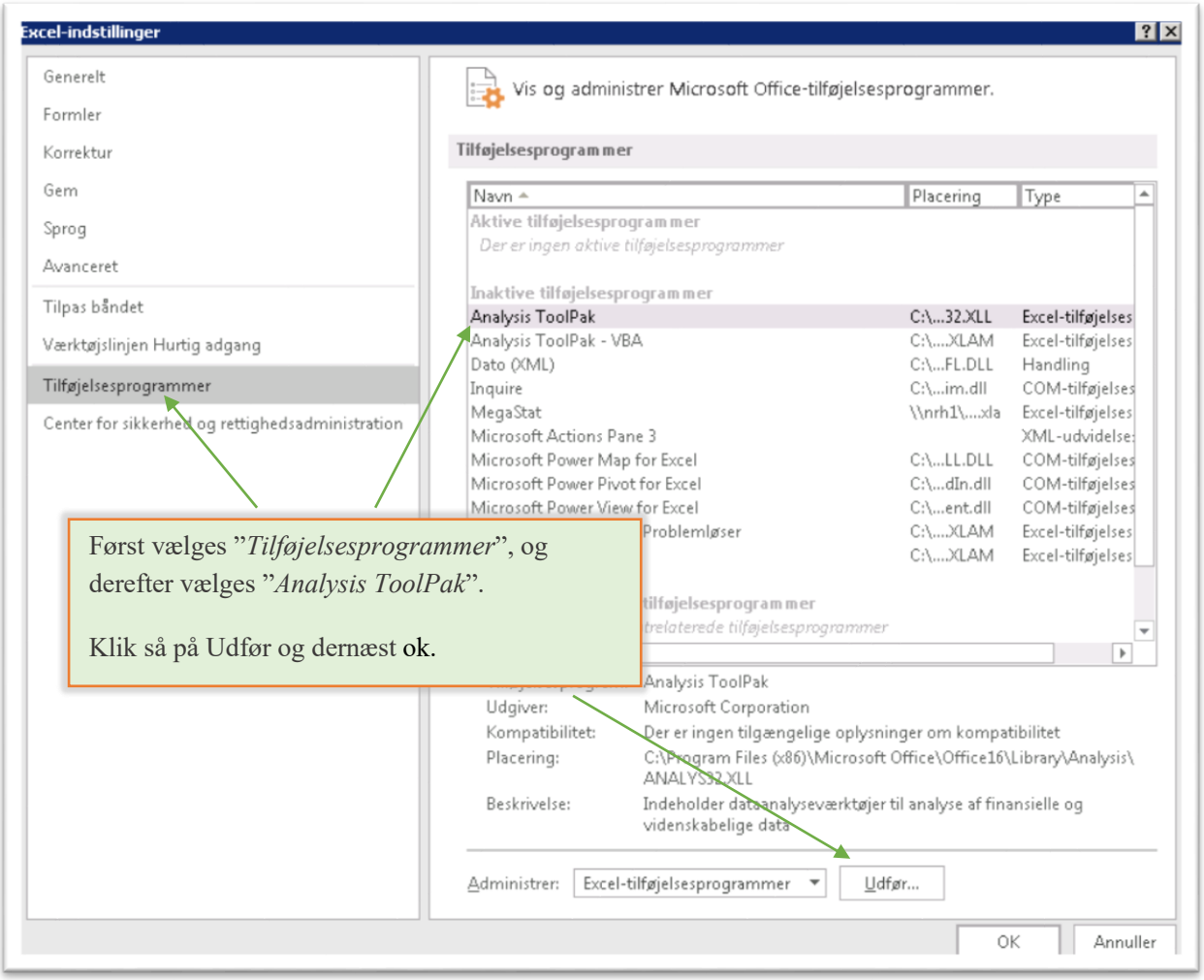
Vejledning til brug af dataanalyse i Excel.

Der vil være flere mulige CAS-værktøjer der kan lave multipel regressions analyse. Det umiddelbare letteste program er *Dataanalyse* fra Excel. De fleste elever vil have adgang til dette lille værktøj.

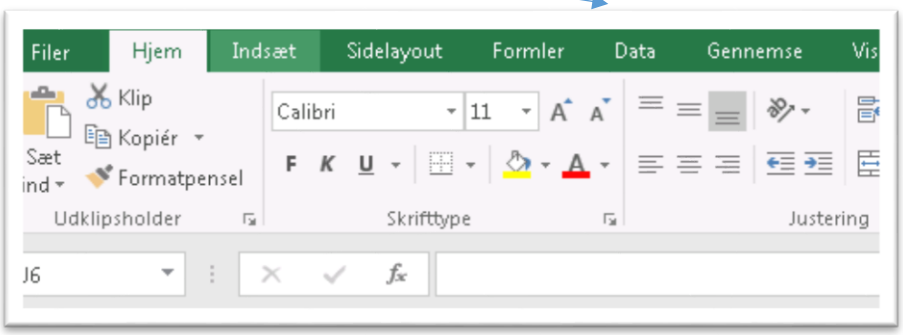
Vejledning til hvordan man installerer "*Dataanalyse*" fra Excel.

Trin 1. Start med at åbne Excel – vælg dernæst øverste fane til venstre "filer". Gå derefter ned til indstillinger.

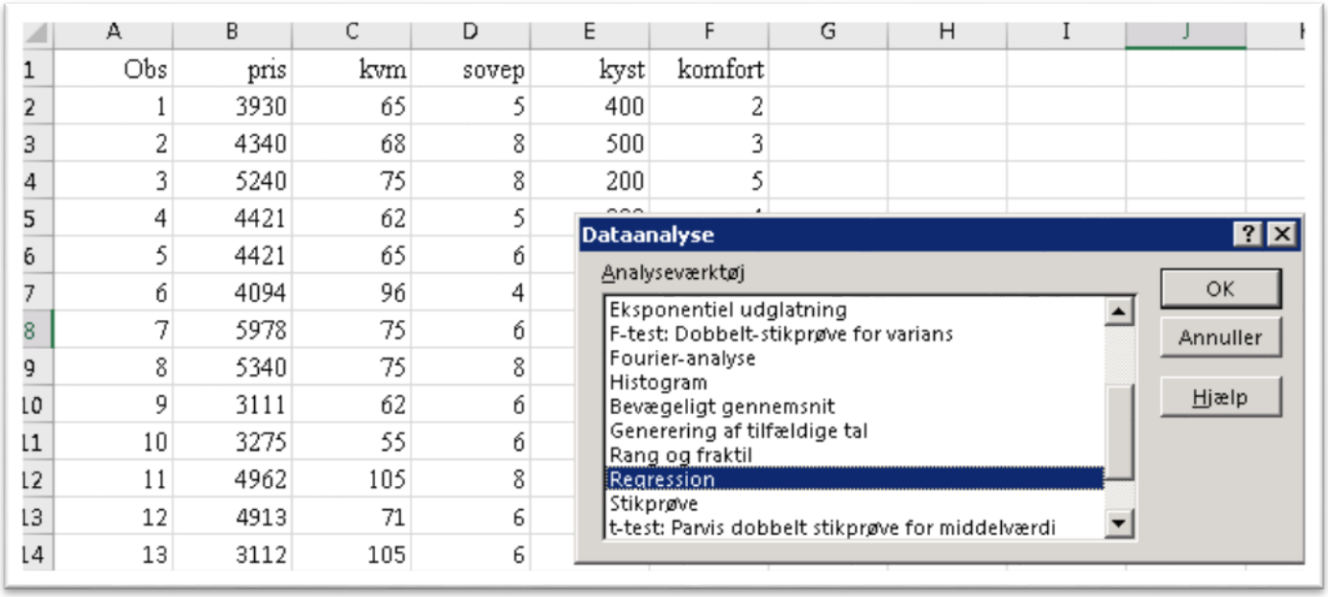
Trin 2.I vinduet Excel indstillinger ses dette -



Hvordan finder jeg *Dataanalyse*? Klik på fanen Data, og yderst til højre ses dataanalysen



Trin 3. Klik på dataanalyse ude til højre og der åbnes et nyt vindue.



	A	B	C	D	E	F
1	Obs	pris	kvm	sovep	kyst	komfort
2	1	3930	65	5	400	2
3	2	4340	68	8	500	3
4	3	5240	75	8	200	5
5	4	4421	62	5		
6	5	4421	65	6		
7	6	4094	96	4		
8	7	5978	75	6		
9	8	5340	75	8		
10	9	3111	62	6		
11	10	3275	55	6		
12	11	4962	105	8		
13	12	4913	71	6		
14	13	3112	105	6		

Indtil nu ligner det den almindelige metode ved simpel lineær regressions analyse, men ved multipel regression er der den ”lille” forskel, at her skal man vælge ”*x-område*” ved at tage alle søjler med – fra C til F søjle. Husk at ”*y-område*” er B søjlen i dette tilfælde.