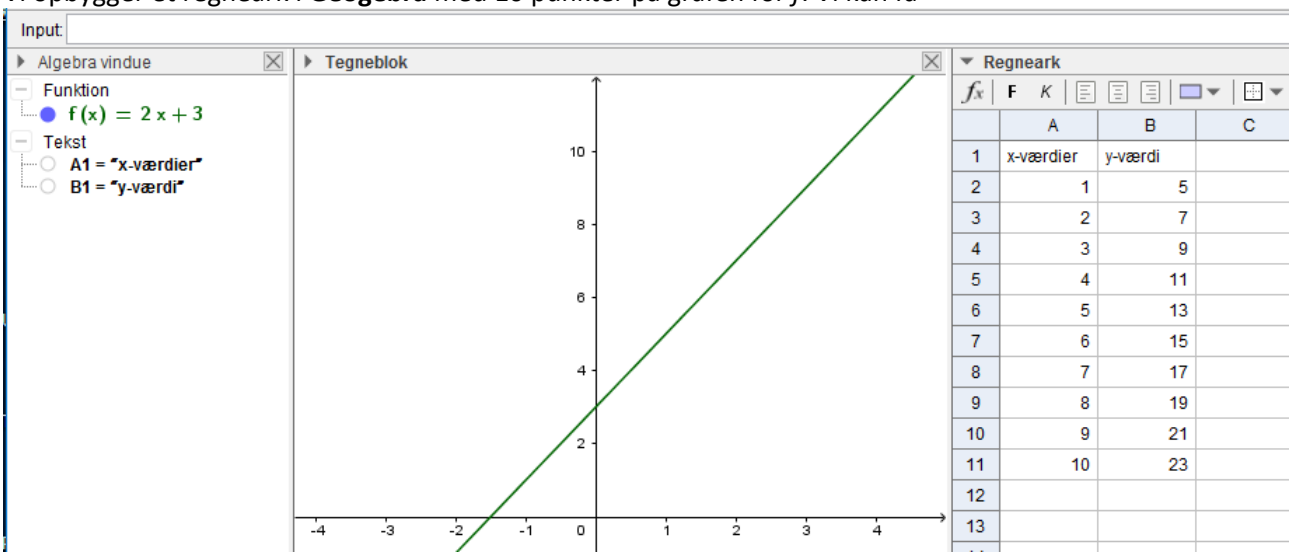
Projekt 8.4 Geogebra-eksperiment med residualspredningen: Bedste estimator for varians for residualer ved lineær regression

Eksperimentet i dette papir ligger også som video på https://youtu.be/9xDcGeWaP7E.

Problemstilling: Varians for residualer ved lineær regression – hvilken værdi er den bedste estimator?
Dette projekt gennemføres med værktøjet Geogebra. Du kan her finde en værktøjsuafhængig beskrivelse af, hvordan projektet kan gennemføres med andre værktøjer.

Vi regner på et datasæt med 10 talpar.

Del 1: Vi vælger her, at udgangspunktet er en lineær sammenhæng $f(x) = 2x + 3$.
Vi opbygger et regneark i **Geogebra** med 10 punkter på grafen for *f*. Vi kan få

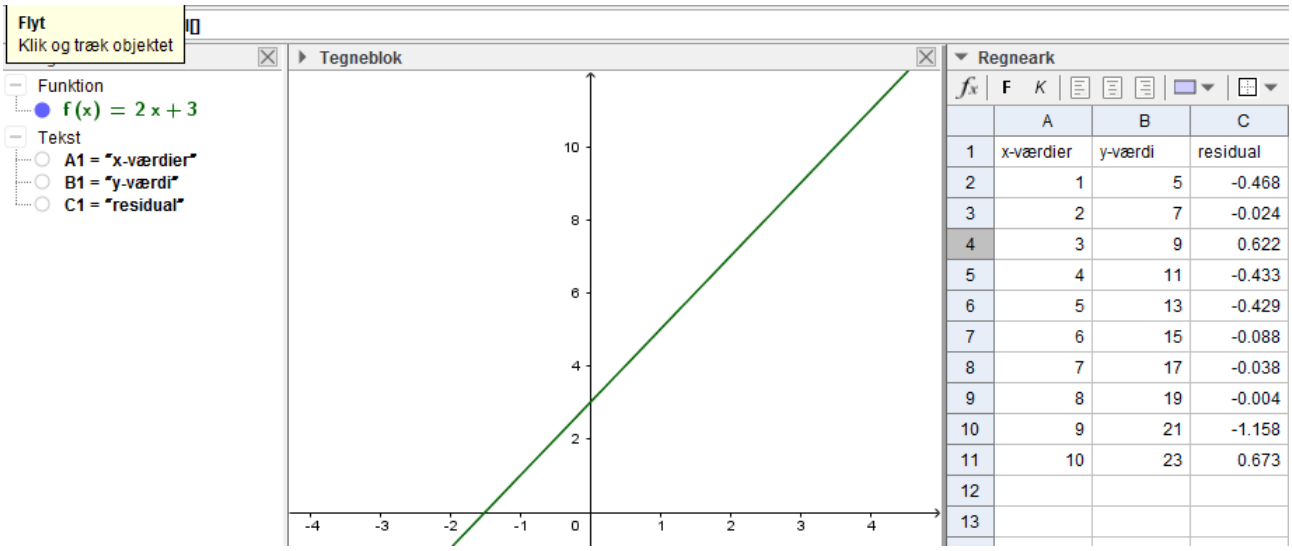


	A	B	C
1	x-værdier	y-værdi	
2	1	5	
3	2	7	
4	3	9	
5	4	11	
6	5	13	
7	6	15	
8	7	17	
9	8	19	
10	9	21	
11	10	23	
12			
13			

Desuden antager vi, at vi har residualer e_i, der alle er normalfordelte med $N\left(0, \frac{1}{2}\right)$.

Dvs. den "sande" varians er $\frac{1}{4}$. Det er denne vi i det følgende vil antage, vi ikke kender, og som vi vil estimere.

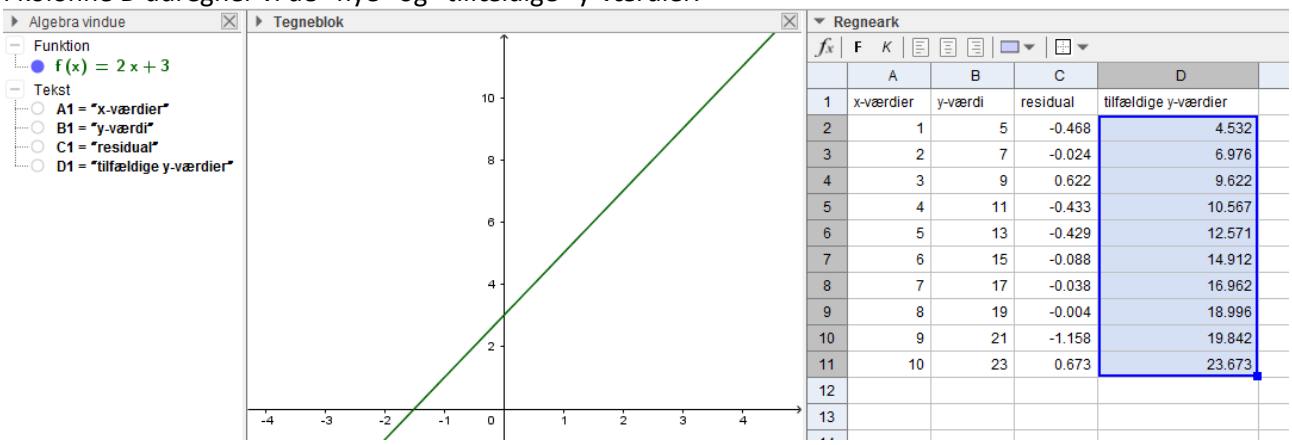
I kolonnen "C" simulerer vi 10 residualer med kommandoen TilfældigNormal(<middelværdi>,<spredning>).
Vi kan få



	A	B	C
1	x-værdier	y-værdi	residual
2	1	5	-0.468
3	2	7	-0.024
4	3	9	0.622
5	4	11	-0.433
6	5	13	-0.429
7	6	15	-0.088
8	7	17	-0.038
9	8	19	-0.004
10	9	21	-1.158
11	10	23	0.673
12			
13			

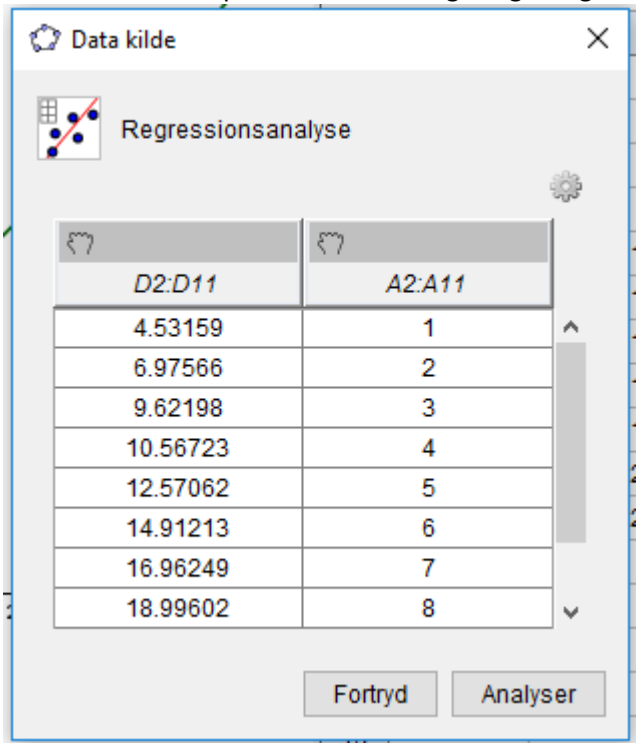
Når vi taster Crtl R, så får vi 10 nye residualer.
Del 2: Vi vil nu simulere "nye" og "tilfældige" y-værdier, hvor den bag ved liggende lineære model er *f*.

I kolonne D udregner vi de ”nye” og ”tilfældige” y-værdier.



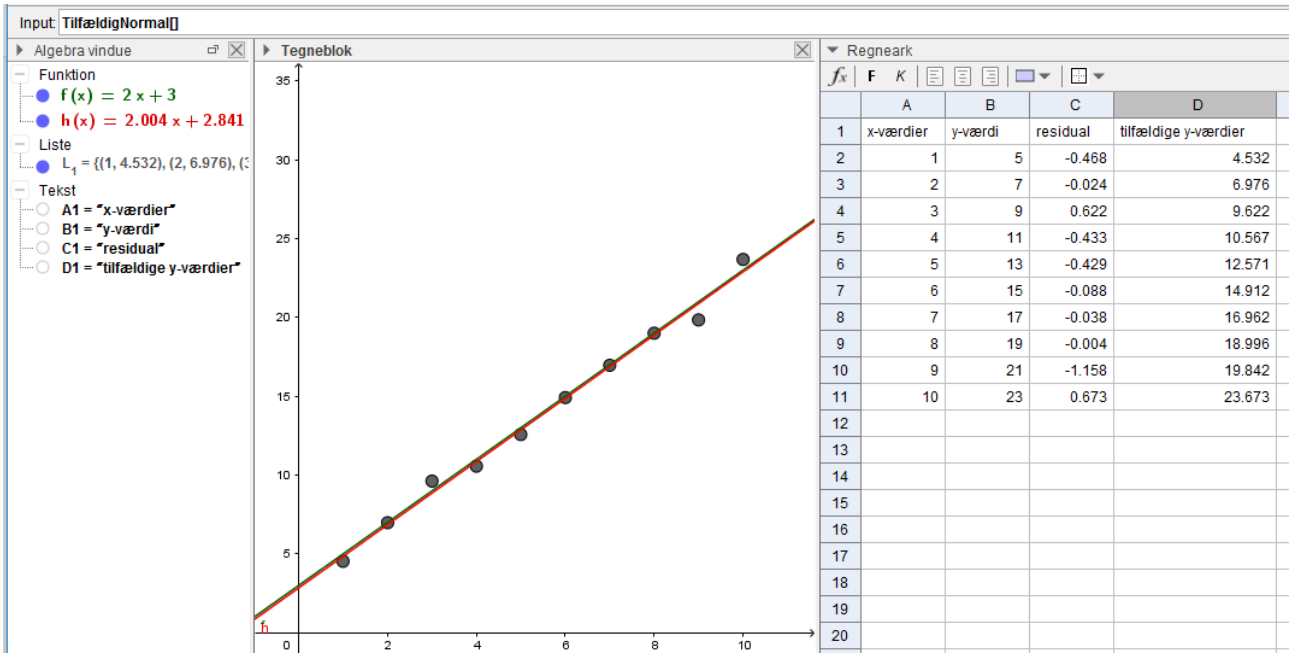
	A	B	C	D
1	x-værdier	y-værdi	residual	tilfældige y-værdier
2	1	5	-0.468	4.532
3	2	7	-0.024	6.976
4	3	9	0.622	9.622
5	4	11	-0.433	10.567
6	5	13	-0.429	12.571
7	6	15	-0.088	14.912
8	7	17	-0.038	16.962
9	8	19	-0.004	18.996
10	9	21	-1.158	19.842
11	10	23	0.673	23.673
12				
13				

Del 3: Vi udfører nu lineær regression på punkterne (<x-værdier>,<tilfældige y-værdier>).
Vi markerer de respektive kolonner og vælger regressionsværktøjet. Måske får vi



Marker kolonne A og tryk på ”hånden”, derefter marker kolonne D og tryk på hånden.
Klik derefter på Analyser.

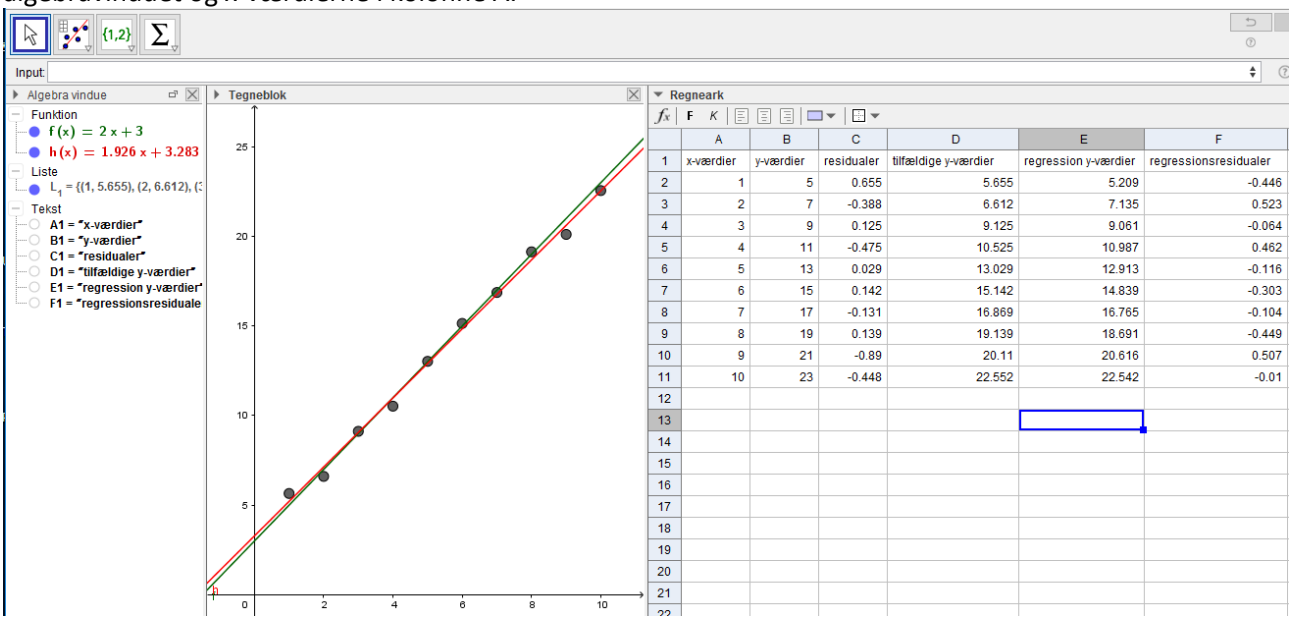
Hvis vi vælger ”Lineær”, og derefter ”Kopier til tegneblok”, så kan vi få:



	A	B	C	D
1	x-værdier	y-værdi	residual	tilfældige y-værdier
2	1	5	-0.468	4.532
3	2	7	-0.024	6.976
4	3	9	0.622	9.622
5	4	11	-0.433	10.567
6	5	13	-0.429	12.571
7	6	15	-0.088	14.912
8	7	17	-0.038	16.962
9	8	19	-0.004	18.996
10	9	21	-1.158	19.842
11	10	23	0.673	23.673

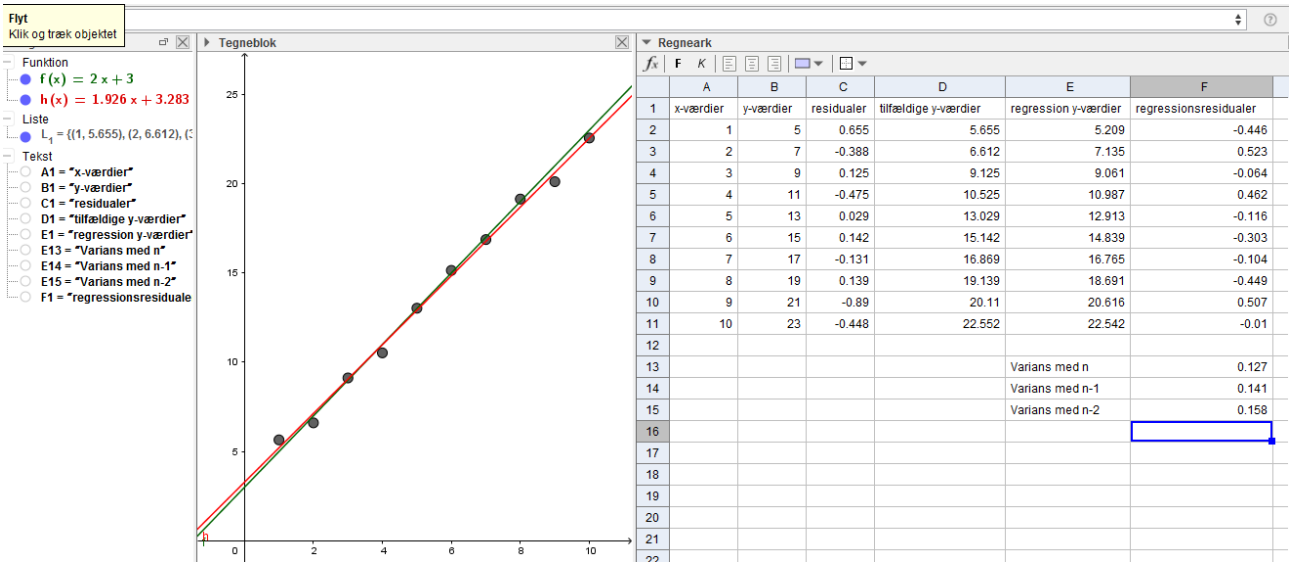
Klikker vi på Crtl R, så laver vi hver gang en ny simulering.

Del 4: For hver simulation vil vi udregne regressionsmodellens y-værdier ud fra regressionsforskriften i algebravinduet og x-værdierne i kolonne A.



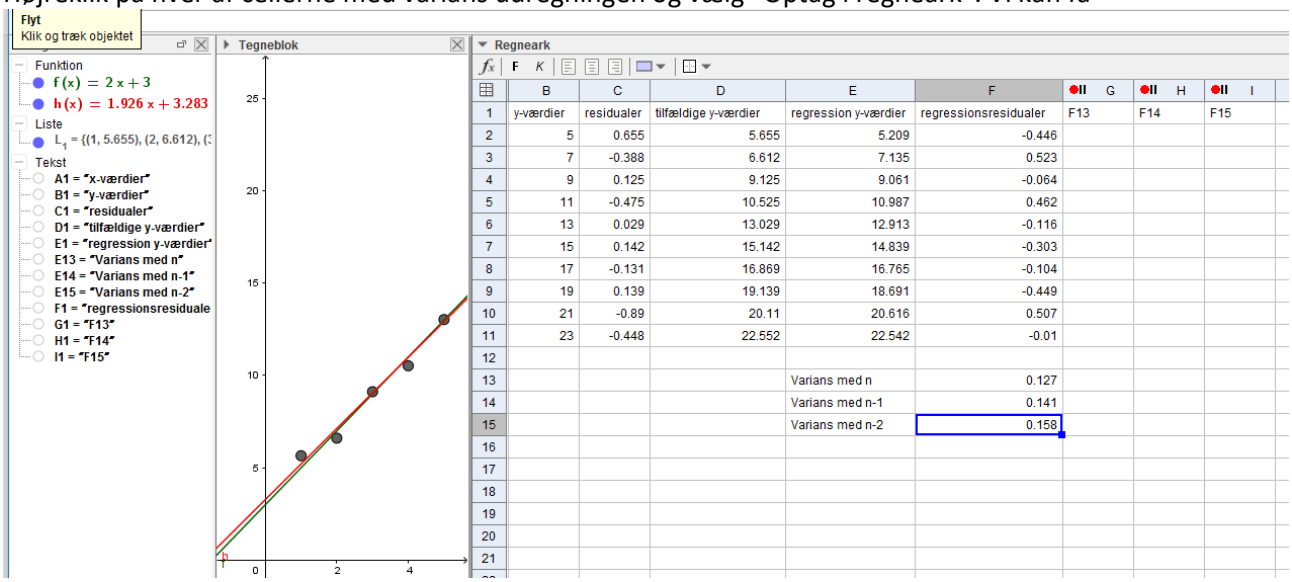
	A	B	C	D	E	F
1	x-værdier	y-værdier	residualer	tilfældige y-værdier	regression y-værdier	regressionsresidualer
2	1	5	0.655	5.655	5.209	-0.446
3	2	7	-0.388	6.612	7.135	0.523
4	3	9	0.125	9.125	9.061	-0.064
5	4	11	-0.475	10.525	10.987	0.462
6	5	13	0.029	13.029	12.913	-0.116
7	6	15	0.142	15.142	14.839	-0.303
8	7	17	-0.131	16.869	16.765	-0.104
9	8	19	0.139	19.139	18.691	-0.449
10	9	21	-0.89	20.11	20.616	0.507
11	10	23	-0.448	22.552	22.542	-0.01

Vi kan nu udregne residualerne for regressionsmodellen i kolonne F, og derefter residualernes varians i tre udgaver. En version, hvor vi dividerer kvadratsummen med *n*, *n-1* og *n-2*. Vi kan bruge kommandoen Varians(F2:F11).



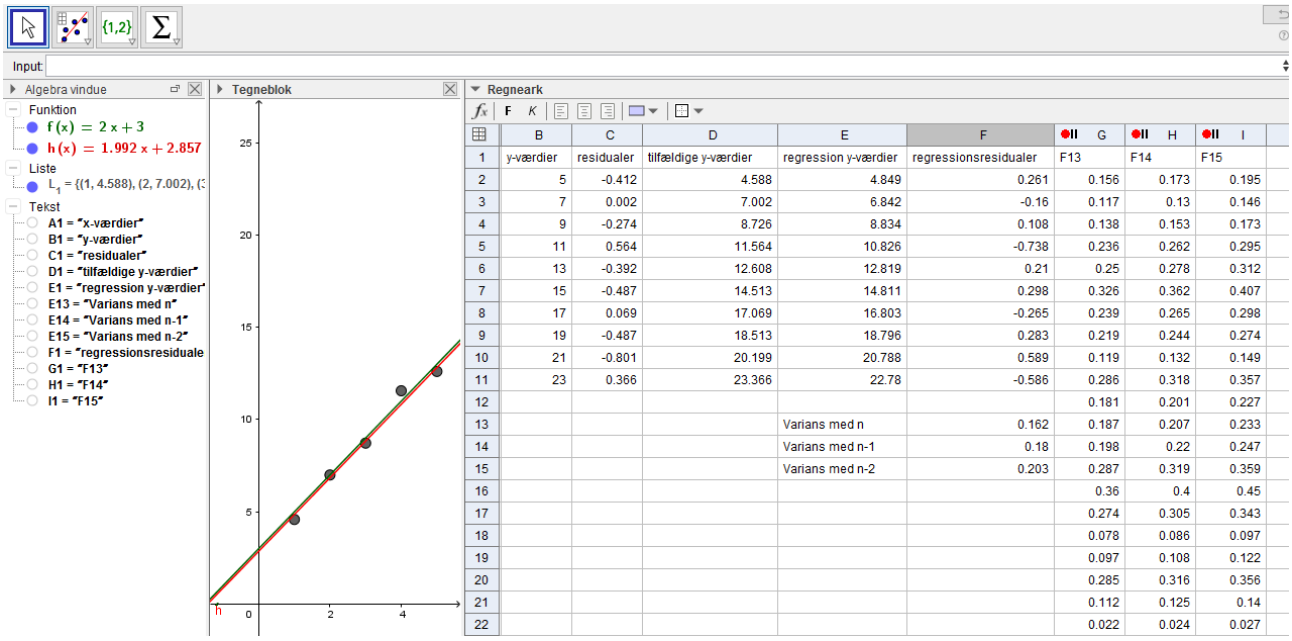
	A	B	C	D	E	F
1	x-værdier	y-værdier	residualer	tilfældige y-værdier	regression y-værdier	regressionsresidualer
2	1	5	0.655	5.655	5.209	-0.446
3	2	7	-0.388	6.612	7.135	0.523
4	3	9	0.125	9.125	9.061	-0.064
5	4	11	-0.475	10.525	10.987	0.462
6	5	13	0.029	13.029	12.913	-0.116
7	6	15	0.142	15.142	14.839	-0.303
8	7	17	-0.131	16.869	16.765	-0.104
9	8	19	0.139	19.139	18.691	-0.449
10	9	21	-0.89	20.11	20.616	0.507
11	10	23	-0.448	22.552	22.542	-0.01
12						
13					Varians med n	0.127
14					Varians med n-1	0.141
15					Varians med n-2	0.158
16						
17						
18						
19						
20						
21						

Del 5: Vi vil nu lave 1000 simuleringer af de regressionen, så vi får 1000 værdier for hver af varianserne. Højreklik på hver af cellerne med varians udregningen og vælg "Optag i regneark". Vi kan få



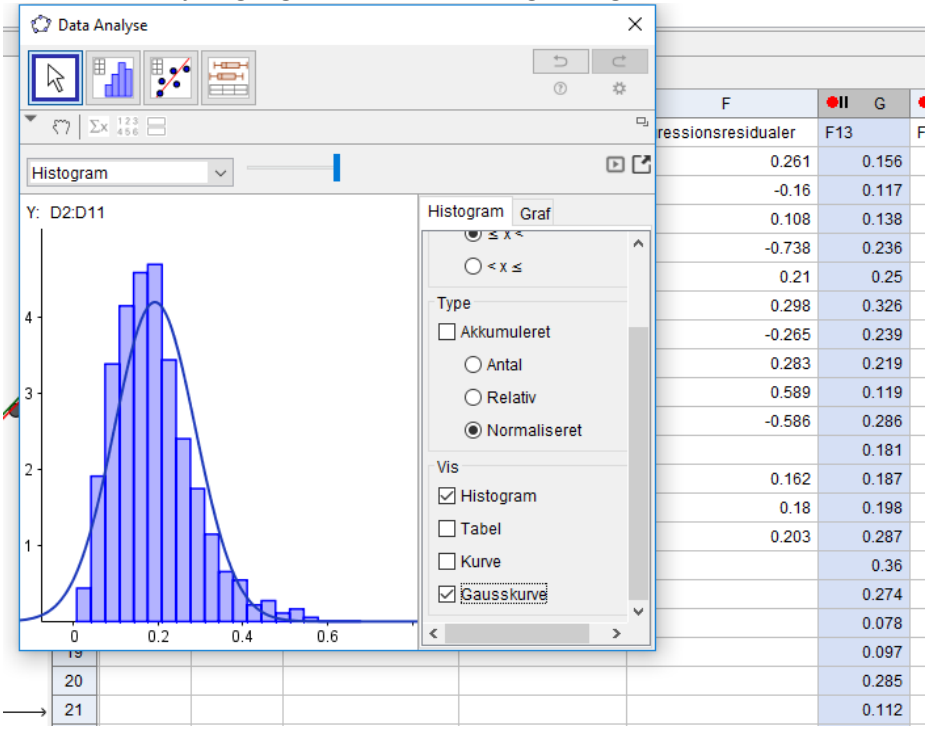
	B	C	D	E	F	G	H	I
1	y-værdier	residualer	tilfældige y-værdier	regression y-værdier	regressionsresidualer	F13	F14	F15
2	5	0.655	5.655	5.209	-0.446			
3	7	-0.388	6.612	7.135	0.523			
4	9	0.125	9.125	9.061	-0.064			
5	11	-0.475	10.525	10.987	0.462			
6	13	0.029	13.029	12.913	-0.116			
7	15	0.142	15.142	14.839	-0.303			
8	17	-0.131	16.869	16.765	-0.104			
9	19	0.139	19.139	18.691	-0.449			
10	21	-0.89	20.11	20.616	0.507			
11	23	-0.448	22.552	22.542	-0.01			
12								
13				Varians med n	0.127			
14				Varians med n-1	0.141			
15				Varians med n-2	0.158			
16								
17								
18								
19								
20								
21								

Tryk på Crtl R mange gange, og vi kan få

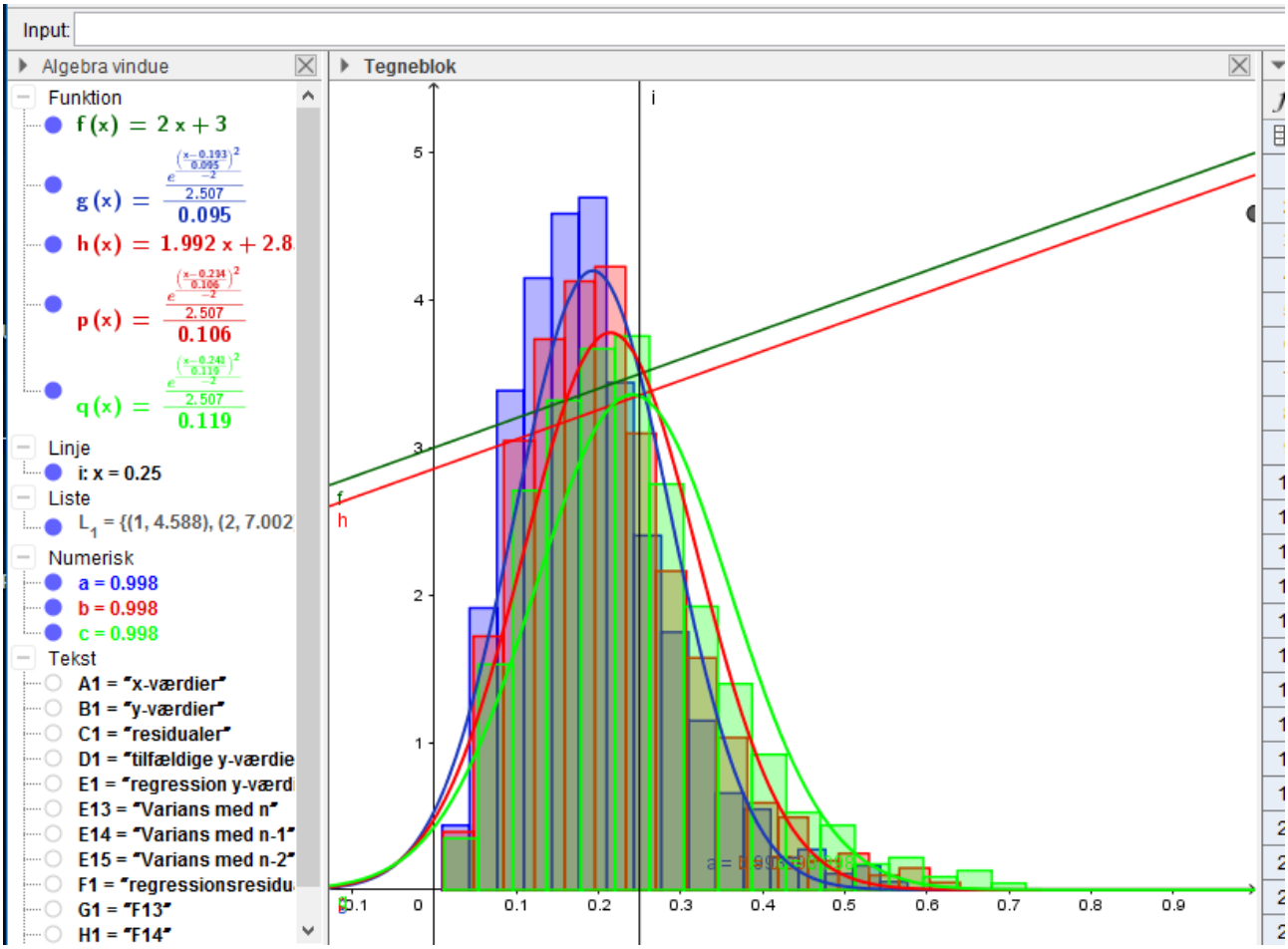


	B	C	D	E	F	G	H	I
1	y-værdier	residualer	tilfældige y-værdier	regression y-værdier	regressionsresiduale	F13	F14	F15
2	5	-0.412	4.588	4.849	0.261	0.156	0.173	0.195
3	7	0.002	7.002	6.842	-0.16	0.117	0.13	0.146
4	9	-0.274	8.726	8.834	0.108	0.138	0.153	0.173
5	11	0.564	11.564	10.826	-0.738	0.236	0.262	0.295
6	13	-0.392	12.608	12.819	0.21	0.25	0.278	0.312
7	15	-0.487	14.513	14.811	0.298	0.326	0.362	0.407
8	17	0.069	17.069	16.803	-0.265	0.239	0.265	0.298
9	19	-0.487	18.513	18.796	0.283	0.219	0.244	0.274
10	21	-0.801	20.199	20.788	0.589	0.119	0.132	0.149
11	23	0.366	23.366	22.78	-0.586	0.286	0.318	0.357
12						0.181	0.201	0.227
13				Varians med n	0.162	0.187	0.207	0.233
14				Varians med n-1	0.18	0.198	0.22	0.247
15				Varians med n-2	0.203	0.287	0.319	0.359
16						0.36	0.4	0.45
17						0.274	0.305	0.343
18						0.078	0.086	0.097
19						0.097	0.108	0.122
20						0.285	0.316	0.356
21						0.112	0.125	0.14
22						0.022	0.024	0.027

Del 6: Vi undersøge fordelingen af hver af varianserne.
Marker hver søjle og tegn et normeret histogram og eventuelt en normalfordelingskurve. Vi kan få



Vi vælger at kopiere alle histogrammer til tegneblokken, sammen med den "sande" varians. Samlet kan vi få:



Konklusion: Vores simuleringer viser, at residualvariansen, hvor vi dividerer med $n-2$ er den bedste estimator for den "sande" residualvarians.