

Hjemmeopgave

Basal statistik for sundhedsvidenskabelige forskere, efterår 2021

Udleveret xx 2021, afleveres senest ved øvelserne i uge xx (- 2021).

Et gammelt studie af Galton er beskrevet i artiklen: James A Hanley (2004): Transmuting Women into Men, The American Statistician, 58:3, 237-243.

har undersøgt sammenhængen mellem højden af forældre og afkom.

Disse data er også beskrevet i R:

<https://vincentarelbundock.github.io/Rdatasets/doc/mosaicData/Galton.html>

På hjemmesiden

<https://publicifsv.sund.ku.dk/~sr/BasicStatistics/hjemmeopgave.html>

ligger oplysninger om 7 variable for hver af 934 børn:

idnr: Løbenummer for personen

family: Familiens løbenummer

father: Faderens højde i tommer

mother: Moderens højde i tommer

childNum:hvad er det mon?

gender: Barnets køn (1: male eller 2: female)

childHeight: Barnets højde i tommer

Bemærk: I denne besvarelse benyttes alle observationer.

Først skal vi lige indlæse vores materiale, og derefter vi redefinerer vi variabelen **gender** som **sex** for at kunne lave **gender** til en karaktervariabel. Dette gøres ved først blot at ændre variabelnavnet fra **gender** til **sex** i Variable View og dernæst at benytte

Transform/Recode into Different Variables, hvor vi sætter **sex** over i Input Variable --> Output Variable, skriver **gender** i Name under Output Variable og trykker Change. Herefter definerer man værdierne under Old and New Values.

Først sættes dog flueben ved **Output variables are strings**, og herefter skrives 2 i Value under Old Value og female i Value under New Value, samt Add. Herefter skrives tilsvarende 1 i Value under Old Value og male i Value under New Value, samt Add.

Inden vi går i gang med spørgsmålene, skal vi også lige checke rimeligheden af vores variable, f.eks. ved at udregne summary statistics for hvert køn. Dette gøres efter min mening mest overskueligt ved at benytte **Data/Split File**, hvor man sætter flueben ved **Compare Groups** og sætter **gender** over i **Groups Based on**. Derefter kan man benytte **Analyze/Descriptive Statistics/Descriptives** med diverse variable i **Variable(s)**. Herved får vi outputtet:

Descriptive Statistics					
gender		N	Minimum	Maximum	Mean
female	father	453	62,0	78,5	69,260
	mother	453	58,0	70,5	64,149
	childNum	453	1	15	5,00
	childHeight	453	56,0	70,5	64,104
	Valid N (listwise)	453			
male	father	481	62,0	78,5	69,138
	mother	481	58,0	70,5	64,033
	childNum	481	1	10	2,26
	childHeight	481	60,0	79,0	69,234
	Valid N (listwise)	481			

Vi ser her, at materialet består af 453 piger og 481 drenge, og at summary statistics for de 4 højde-variable ser rimeligt "ens" ud, så vi ikke forventer skrivefejl. Det er dog lidt svært at forholde sig til selve højderne, da de er i tommer.

Vi går nu i gang med selve spørgsmålene:

1. *Konverter alle højder fra tommer til cm.*

Her har man selvfølgelig brug for en omregningsfaktor, og det findes nemt på google, at 1 tomme (inch) svarer til 2.54 cm, så vi definerer nye variable med **Transform/Compute** efter nedenstående definitioner (idet vi samtidig forkorter variabelnavnene):

```
fht=father*2.54
mht=mother*2.54
```

height=childHeight*2.54

2. Se på (dele af) originalmaterialet

<http://www.medicine.mcgill.ca/epidemiology/hanley/galton/notebook/index.html>
og vurder hvad variablen childNum mon betyder.

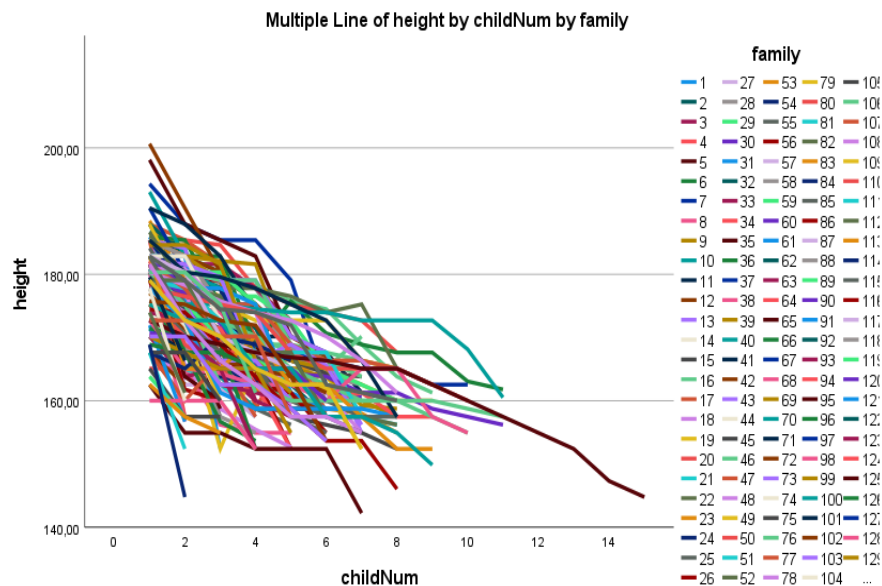
Fra den angivne hjemmeside (side 2), som vist her:

	Father	Mother	Son's order of height	Daughter's order of height
1	15.5	7.0	13.2	9.2, 4.0, 9.0
2	15.5	6.5	13.5, 12.5	8.5, 5.5
3	15.0	6.0	11.0	8.0
4	15.0	4.0	10.5, 7.5	7.0, 4.5, 3.0
5	15.0	4.5	12.0, 9.0, 7.0	6.5, 2.5, 2.5
6	14.0	5.0		4.5
7	14.0	5.0	16.5, 14.0, 13.0, 13.0	10.5, 4.0
8	14.0	6.5		10.5, 7.0, 6.0
9	14.0	6.0		8.0
10	14.0	5.5		5.5
11	14.0	2.0	14.0, 10.0	10.0, 7.0, 6.0, 3.5, 3.0
12	14.0	1.0		5.0
13	13.0	7.0	11.0	2.0
14	13.0	7.0	5.0, 7.0	
15	13.0	6.5	11.0, 10.5	6.7
16	13.0	6.5	12.0, 10.5, 10.5, 10.5, 10.5	5.7, 6.5, 4.5, 3.5
17	13.0	4.5	14.0, 13.0, 10.5, 2.5	6.5, 2.5
18	13.0	4.0		6.0, 4.5, 4.0
19	13.0	3.0		2.7
20	12.7	4.0	13.5, 13.0, 12.7	10.0, 6.5, 2.5, 2.0, 6.0
21	12.0	5.0	13.0	2.5, 2.0
22	12.0	4.0	12.0, 11.0	7.0
23	12.0	5.0	14.0, 10.5, 9.5	6.0, 5.5, 5.0, 4.0
24	12.0	5.5		5.0

kan man udlede, at childNum står for rækkefølgen af børnene fra den højeste til den laveste.

Vi kan underbygge dette ved at lave en figur. Vi benytter Graph/Chart Builder/Line og vælger det til højre.

Sæt nu height på Y-aksen, childNum på X-aksen og family i Set color, hvorved vi får figuren:



På denne figur ser vi konsekvent aftagende (eller i hvert fald ikke-stigende) højde med `childNum` og får således bekræftet, at der er tale om en nummerering efter højde.

(a) *Hvad kan vi bruge det til?*

Det kan vi vist ikke rigtigt bruge til noget.... Det er en information, man blot kunne have dannet selv ud fra højderne på de enkelte børn.

(b) *Var der en anden information, man hellere skulle have registreret?*

Det ville have været mere interessant at have registreret nummeret i søskendeflokken. Så kunne man have undersøgt, om der skulle være en tendens til at højden øgedes (eller formindskedes) med nummeret i søskendeflokken.

Se i det følgende bort fra, at der er flere børn hørende til hver familie (korrelerede data).

Vi har i disse data op til 15 børn fra de enkelte familier (se figuren ovenfor), og vi må formode, at der er en vis (positiv) korrelation mellem disse børns højde. Noget af denne forklares givetvis ud fra moderens og faderens højde, men der kan nemt være uforklaret korrelation f.eks. pga. samme kostvaner el.lign. Dette bør man tage hensyn til for at få korrekte standard errors på sine estimer, og dette skal vi komme tilbage til til allersidst (og I lærer det på den sidste dag af kurset).

3. Her ser vi på en far og hans afkom:

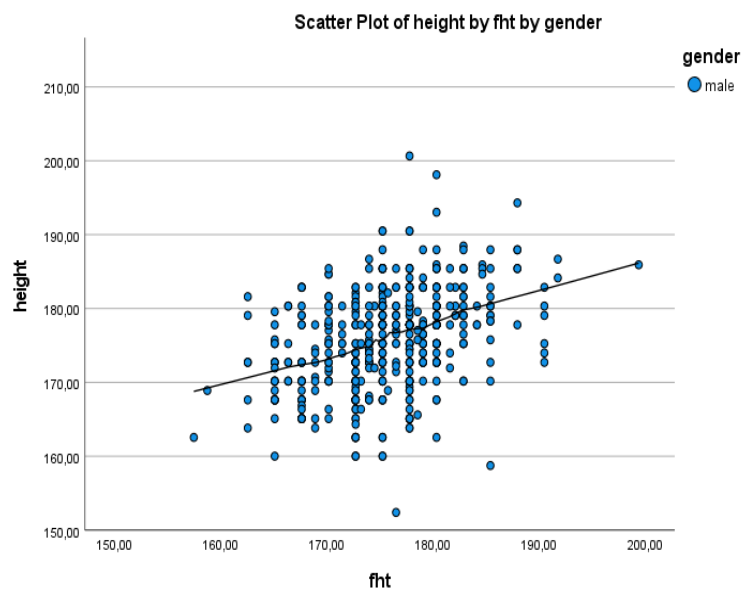
- (a) *Opstil en formel til beregning af den mest sandsynlige højde for en søn, givet kun fars højde.*

Her skal vi kun se på drengebørn, så i dette spørgsmål benytter vi `Data/Select Cases`, sætter flueben i `If condition is satisfied` og under IF skrive `gender="male"`.

Når vi ikke ved noget om sønnens moder, må vi blot se på en model, der relaterer sønners højde til fædres højde. Først laver vi lige en tegning. Vi går ind i `Graph/Chart Builder/Scatter`, sætter `height` over på Y-aksen, og `ght` over på X-aksen.

På den fremkomne figur lægger vi en udglattet `Loess`-kurve ind ved at dobbeltklikke på figuren, klikke `Add Fit Line` at `Total` og vælge `Loess`

Vi får så:



Da der her ses en rimeligt lineær sammenhæng mellem fædres og sønners højde, går vi videre med en simpel lineær regression ved hjælp af menuen **Analyze/Regression/Linear**, hvor man i boksen sætter **height** som **Dependent** og **fht** som **Independent(s)** (et uheldigt navn til forklarende variable...).

Man skal efterfølgende huske at gå ind i **Statistics** og afkrydse **Parameter Estimates** og **Confidence intervals**.

Så får vi outputtet

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,392 ^a	,154	,152	6,13661

a. Predictors: (Constant), fht

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3282,663	1	3282,663	87,170	,000 ^b
	Residual	18038,203	479	37,658		
	Total	21320,866	480			

a. Dependent Variable: height

b. Predictors: (Constant), fht

Coefficients^a

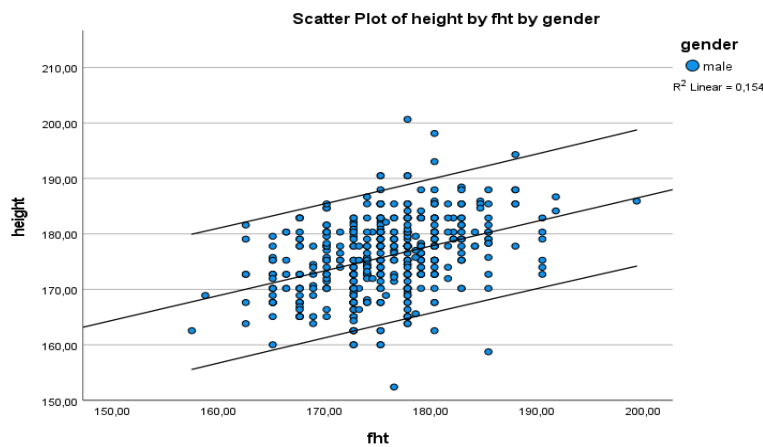
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	97,441	8,403		11,596	,000
	fht	,447	,048	,392	9,337	,000

Coefficients^a

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	80,929	113,953
	fht	,353	,540

a. Dependent Variable: height

Modellen kan illustreres med prediktionsgrænser, som fås ved at dobbeltklikke på scatterplottet fra tidligere og benytte **Add Fit Line at Total** og derefter i **Properties**-boksen afkrydse **Linear** og **Confidence Intervals/Individual**

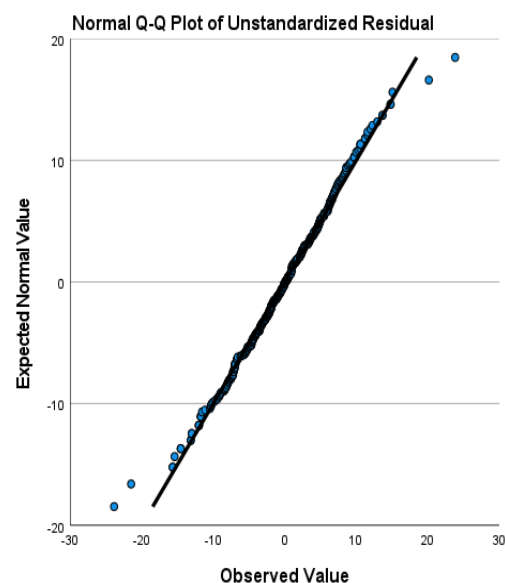
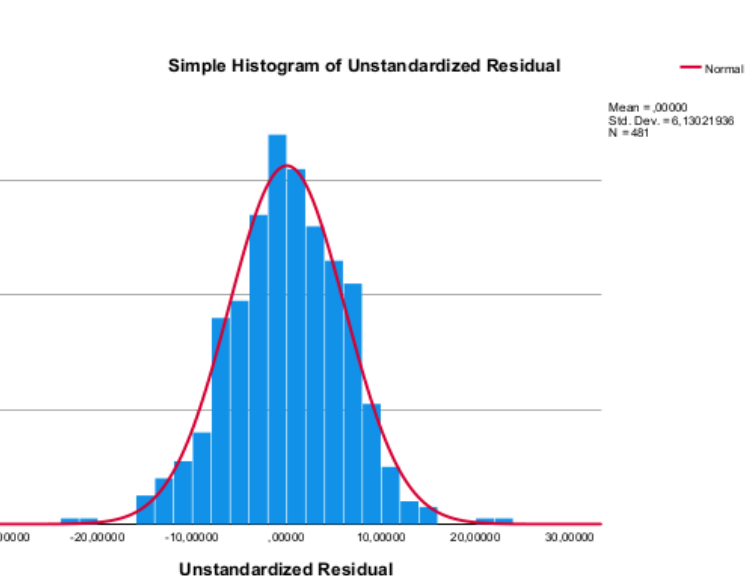


Man bør også lave noget modelkontrol, og til det formål bør man generelt ved udførelsen af selve analysen gemme såvel residualer som forventede værdier. er skal vi dog kun bruge residulaerne, da analysen er så simpel. Man benytter **Save**-knappen, og krydser det af, man gerne vil have, f.eks. under **Predicted values: Unstandardized** og under **Residuals: Unstandardized**. Herved dannes nye variable kaldet **PRE_1** og **RES_1**.

Herefter kan vi f.eks. checke normalfordelingsantagelsen ved at lave histogram og/eller fraktildiagram. Brug **Graphs/Chart Builder** og vælg **Histogram** (det enkleste, dobbeltklik det op i det store felt). Sæt så **RES_1** over på X-aksen, og tryk **OK**, så figuren fremkommer.

For at lægge en normalfordelingskurve oveni, dobbeltklikker man på figuren, klikker på **show distribution curve**-ikonet, afkrydser **Normal** og trykker **Apply/Close**.

Fraktildiagrammet fås direkte fra menuen **Analyze/Descriptive Statistics/Q-Q Plots**, hvor vi sætter **RES_1** over i **Variables** og der trykkes **OK**



Vi kan konstatere, at modelkontrollen ikke giver anledning til bekymring.

Analysen viser en stærkt signifikant (positiv, selvfølgelig) sammenhæng mellem faders og sønners højde, med estimeret regressionskoefficient 0.447 (0.048), og $CI=(0.353, 0.540)$. Det betyder, at for hver ekstra far-cm, forventer vi at sønnen er 0.447 cm højere, altså knap $\frac{1}{2}$ cm.

Formlen til beregning af den mest sandsynlige højde for en søn, givet kun fars højde er således

$$\text{height} = 97.44 + 0.447 * \text{fht}$$

- (b) Hvis far A er 5 cm højere end far B, hvor meget forventer vi så at A's søn er højere end B's søn?

Her skal vi blot bruge resultatet fra vores regressionsanalyse ovenfor, idet den umiddelbart giver den forventede forskel på sønners højde, hvis deres fædres højde afviger med 1 cm. Denne højdeforskel er angivet som hældningen, altså 0.447 cm, med $CI=(0.353, 0.540)$ cm. Vi skal så bare gange dette op med 5, og finder den forventede forskel på A's og B's sønner til 2.24 cm, med $CI=(1.77, 2.70)$ cm.

- (c) *Vil sønner altid forventes at blive højere end deres far?*

Vi kan bare betragte figuren fra spørgsmål 5a og konstatere, at sønner *ikke* altid vil blive højere end deres fædre. Små fædre vil typisk få sønner, der er højere end dem selv, medens høje fædre typisk vil få sønner, der er *lavere* end dem selv (dette kaldes *regression to the mean*), men hvor ligger skæringspunktet? Med andre ord:

- (d) *Hvilken højde skal en far have, for at vi forventer, at sønnen bliver præcis ligeså høj?*

Det kan vi finde ud af, ved at løse ligningen:

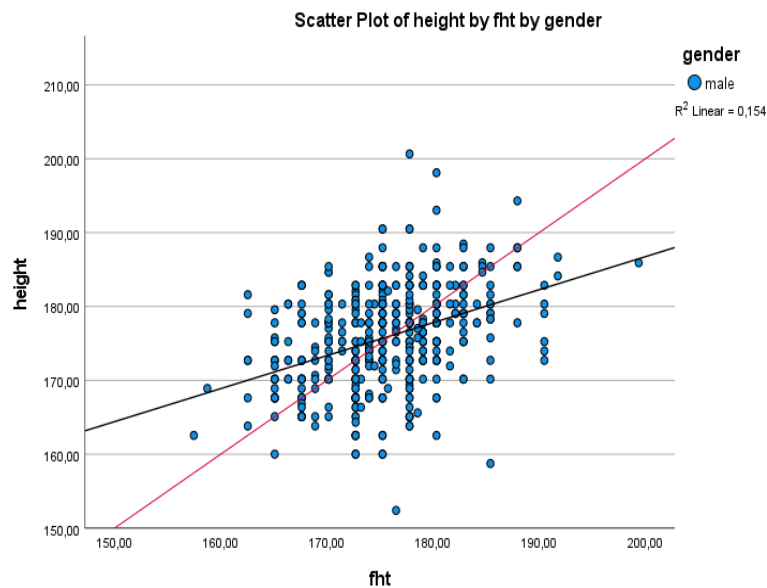
$$x = 97.44 + 0.447 * x$$

og vi finder

$$x = \frac{97.44}{(1 - 0.447)} = 176.2 \text{ (cm)}$$

Vi kan også illustrere det på forskellig vis, dels med et plot, hvor vi udover regressionslinien også har indlagt en identitetslinie ved at dobbeltklikke på figuren, klikke på **Properties**-ikonet og vælger **Add a reference line from Equation**, hvorefter man i **Custom Equation** skriver $y=1*x+0$ og herefter **Apply**.

Evt kan man vælge liniens farve under **Linear**, og her er valgt rød for identitetslinien.



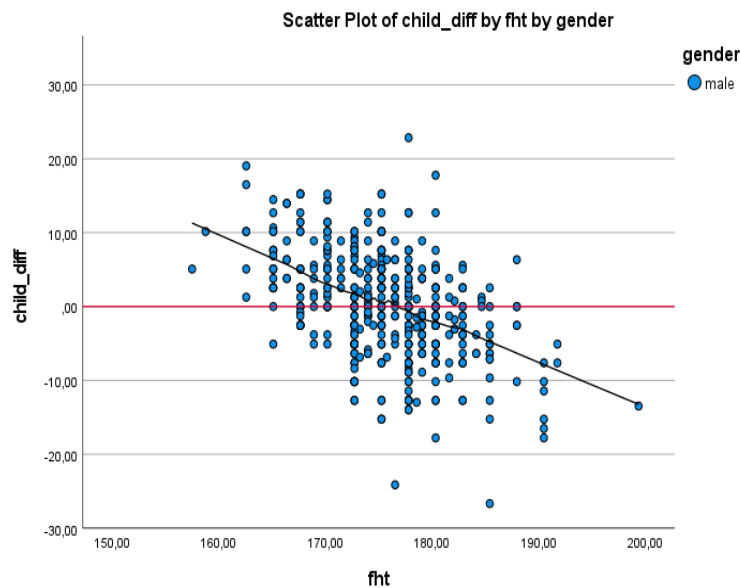
Her ser vi, at denne identitetslinie netop skærer regressionslinien i 176 (hvis man kigger godt efter).

Alternativt kan vi definere forskellen mellem fars og søns højde ved at bruge **Transform/Compute** og definere

`child_diff=height-fht`

og dernæst f.eks tegne denne differens op mod faderens højde, incl. en vandret linie i 0 (brug konstruktionen med **Custom Equation** som beskrevet ovenfor, blot med $y=0$).

Vi får så figuren:



som igen viser, at differensen omkring en faderhøjde på 176 cm tipper over mod at være mere negativ end positiv, svarende til overvejende lavere sønner.

- (e) *Preben er 172 cm høj: Er det sandsynligt, at han får en søn, der er 185 cm høj?*

Her skal vi enten benytte den opskrevne formel og lave udregningen:

$$97.44 + 0.447 * 172 = 174.32 \text{ (cm)}$$

For at finde et 95% prediktionsinterval for højden af Prebens sønner, skal vi bruge spredningen omkring regressionslinien, som ses ovenfor under **Std. Error of the Estimate**: $6.1366 \approx 6.14$, så vi udregner:

$$174.32 \pm 2 \times 6.14 = (162.04, 186.60)$$

Det er derfor ikke meget usædvanligt, at Prebens sønner vil blive så høje som 185 cm, men det ligger dog i den høje ende.

- (f) *Hvor meget lavere vil vi forvente, at en datter af Preben vil være? Husk at benytte en model, der tillader at angive dette estimat med*

CI.

Nu er vi nødt til også at inddrage døtrene i analysen, men inden vi går det, laver vi lige analysen fra spørgsmål 3a for pigerne alene. Derfor ændrer vi i **Data/Select Cases, If condition is satisfied**, så der nu står **gender="female"**, og herefter gentages (dele af) analyserne ovenfor.

Herved får vi outputtet

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,428 ^a	,184	,182	5,41239

a. Predictors: (Constant), fht

b. Dependent Variable: height

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2970,258	1	2970,258	101,395	,000 ^b
	Residual	13211,599	451	29,294		
	Total	16181,857	452			

a. Dependent Variable: height

b. Predictors: (Constant), fht

Coefficients^a

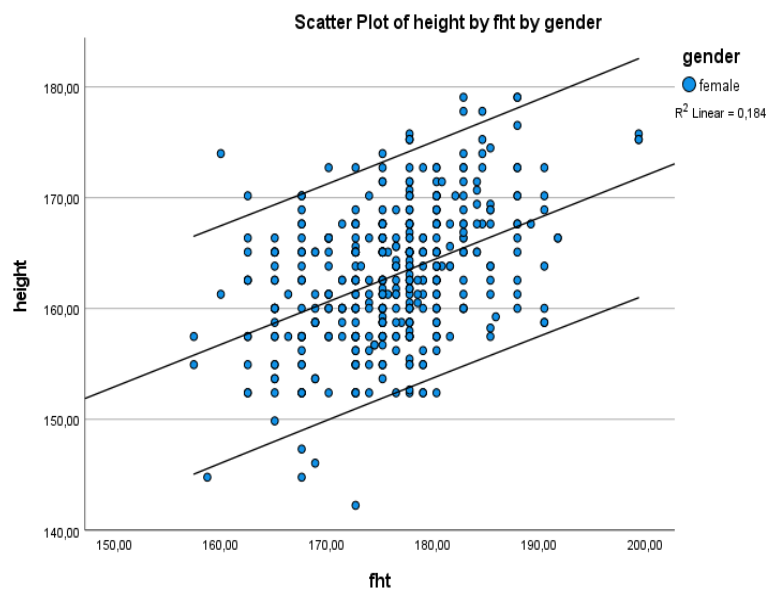
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	95,745	6,666		14,362	,000
	fht	,381	,038	,428	10,069	,000

Coefficients^a

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	82,644	108,846
	fht	,307	,456

a. Dependent Variable: height

med tilhørende plot af prediktionsgrænser (modelkontrol udeladt her):



Vi bemærker igen, at faders højde er en stærkt signifikant prediktor, nu for døtres højde. Desuden ser vi, at 1 cm ekstra for faderen giver forventet 0.38 cm for døtre, medens det gav 0.45 cm for sønner.

Regressionskoefficienten for døtrene er således en smule lavere end for sønnerne, men vi vil i første omgang alligevel antage, at kønsforskellen er uafhængig af fars højde (ingen interaktioner). Vi foretager derfor en multipel regressionsanalyse, med afkommets højde som outcome, og to kovariater: faders højde og barnets køn.

Da kønnet er en kategorisk kovariat, udføres denne som en generel lineær model. Vi benytter derfor **Univariate** under **Analyze/General Linear Model**, sætter **height** i **Dependent Variable**, **gender** i **Fixed Factor(s)** og **fht** i **Covariate(s)**. I **Model** sættes både **gender** og **fht** over i **Model**. Under **Options** husker vi at afkrydse **Parameter Estimates**, og for at lave modelkontrol må vi endvidere gemme residualer og predikterede værdier.

Herved får vi outputtet

Between-Subjects Factors		
		N
gender	female	453
	male	481

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	45825,452 ^a	2	22912,726	681,774	,000
Intercept	11269,361	1	11269,361	335,323	,000
gender	40367,590	1	40367,590	1201,148	,000
fht	6214,139	1	6214,139	184,903	,000
Error	31288,584	931	33,608		
Total	26922144.80	934			
Corrected Total	77114,036	933			

a. R Squared = ,594 (Adjusted R Squared = ,593)

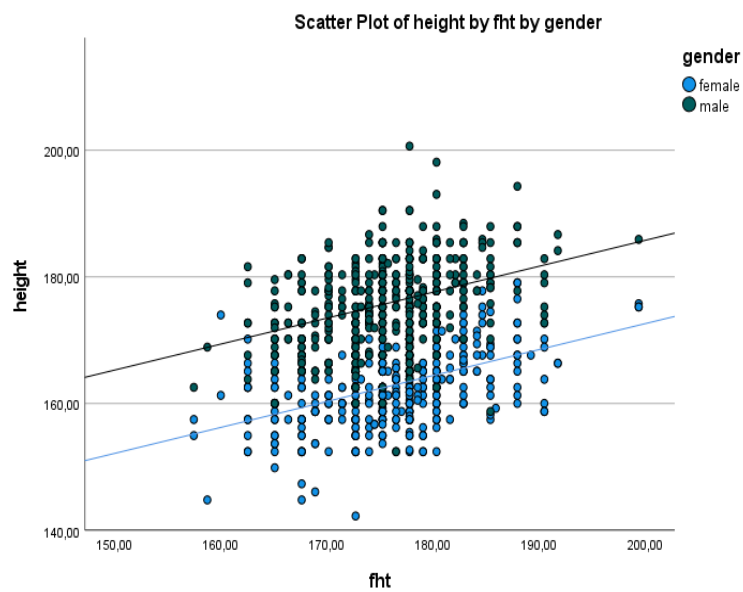
Parameter Estimates

Dependent Variable: height

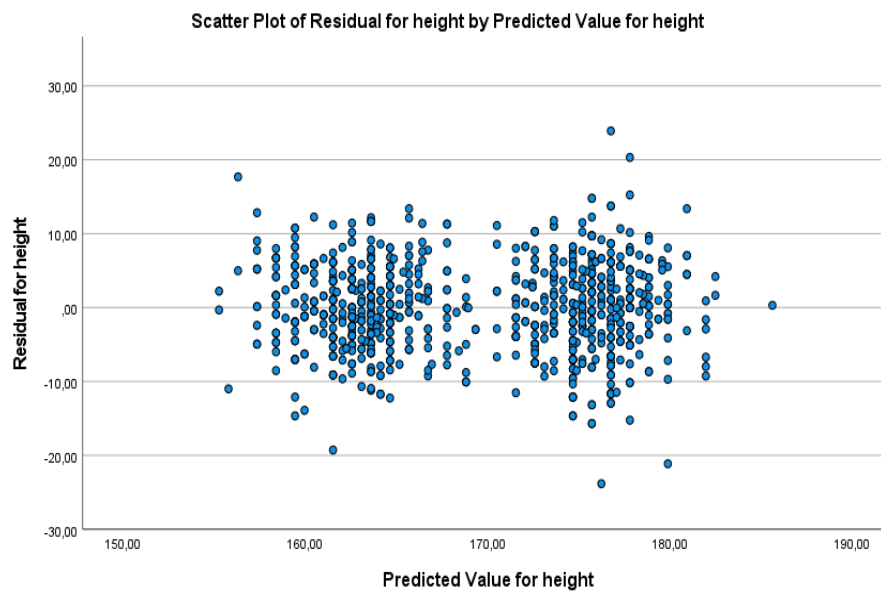
Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	103,783	5,307	19,557	,000	93,369	114,198
[gender=female]	-13,158	,380	-34,658	,000	-13,903	-12,413
[gender=male]	0 ^a	.	.	.	.	.
fht	,410	,030	13,598	,000	,351	,470

a. This parameter is set to zero because it is redundant.

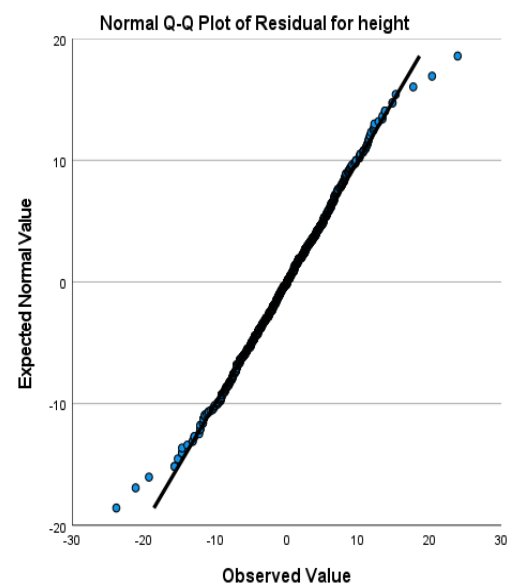
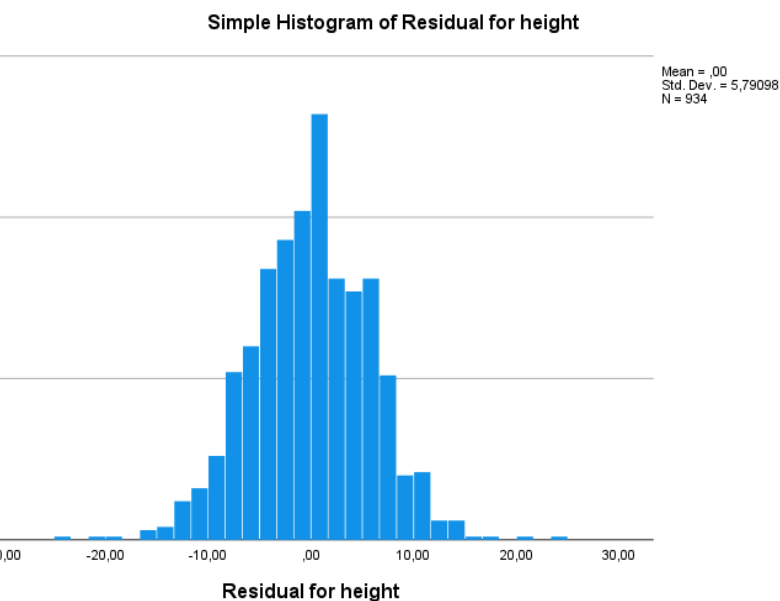
Vi illustrerer modellen de er vist ikke helt parallelle - og dog??, er de lavet ud fra ligningerne!???:



og laver tilhørende modelkontrol plots:



samt illustration af modellen:



Residualplot mod fht mangler!!

Forudsat, at forskellen mellem højden på bror og søster er den samme uafhængigt af faderens højde (svarende til de parallelle regressionslinier i figuren ovenfor), finder vi altså, at en søster forventeligt vil være 13.16 cm lavere end sin bror, $CI=(12.41, 13.90)$ cm.

- (g) *Er den forventede højdeforskel på bror og søster den samme uanset faderens højde?*

I modellen ovenfor antog vi, at den forventede forskel på bror og søsters højde var uafhængig af deres fars højde, men det kan vi jo ikke være sikre på, hvis vi ikke undersøger det.

Vi inkluderer derfor nu interaktionsled mellem **gender** og **fht** under **Model** og laver en ny analyse. Herved får vi outputtet (stærkt beskåret):

Between-Subjects Factors

		N
gender	female	453
	male	481

Parameter Estimates

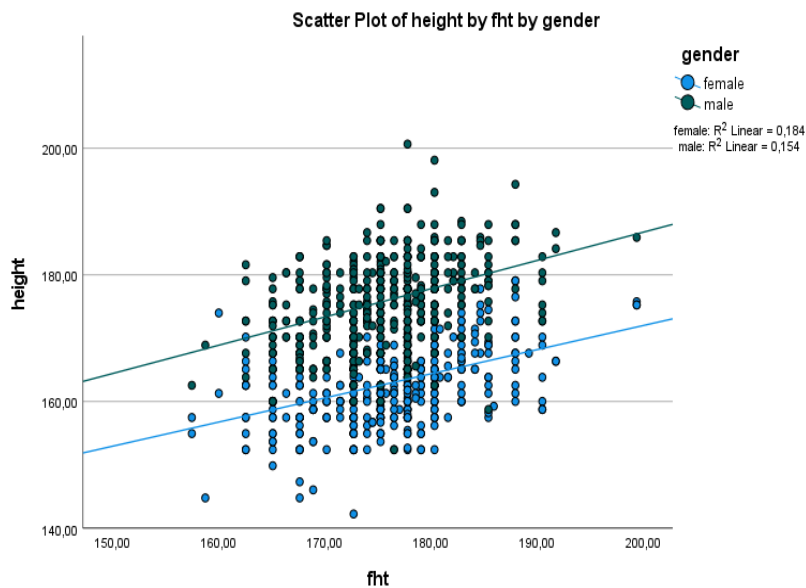
Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	97,441	7,938	12,276	,000	81,863	113,019
[gender=female]	-1,696	10,676	-,159	,874	-22,648	19,257
[gender=male]	0 ^a	.	.	.	.	.
fht	,447	,045	9,884	,000	,358	,535
[gender=female] * fht	-,065	,061	-1,074	,283	-,184	,054
[gender=male] * fht	0 ^a	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

med tilhørende plots:

dårlige farver!!



Vi får her estimeret forskellen på regressionskoefficienterne, altså at døtrenes afhængighed af faders højde er 0.065 lavere end for sønnerne, men at denne forskel ikke er signifikant ($P=0.28$).

Vi kan derfor fastholde, at kønsforskellen på afkommets højde nok ikke afhænger af fars højde, og vi bemærker også, at selv om modellen nu ikke mere antager, at regressionslinierne er parallelle, så er de det næsten alligevel.

Da modellen her ikke afviger nævneværdigt fra den i spørgsmål 3e, undlader vi modelkontrol i denne omgang.

4. Nu ser vi på højden hos et forældrepar.

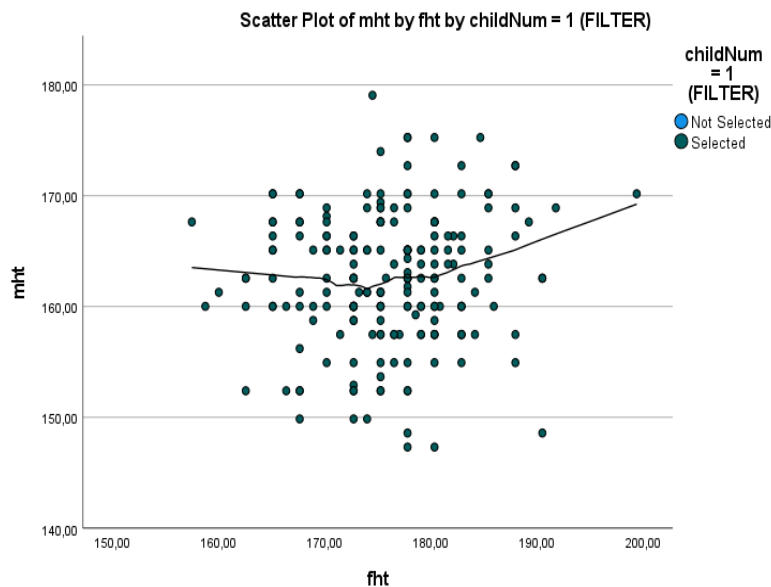
Husk her, at hvert forældrepar typisk optræder adskillige gange i datasættet.

- (a) *Er der tendens til at man vælger en partner, der "passer til" ens egen højde?*

Vi starter med at plote moders højde mod faders højde, med en indlagt loess-kurve.

Her er det vigtigt, at vi begrænser os til at medtage hvert forældrepar en enkelt gang! Dette kan f.eks. gøres ved at begrænse sig til at se på `childNum=1` (der må jo altid være et barn, der er det højeste). Vi bruger altså `Data\Select Cases` og vælger `childNum=1`, hvorefter vi tegner.

Herved får vi figuren:



som ikke tyder på nogen særlig sammenhæng overhovedet. Der synes at være en lille tendens til en stigning til sidst, men dette kan meget vel skyldes den enkelte meget høje mand på ca 2 meter, der finder sig en rimelig høj partner.

For at svare på spørgsmålet, kunne man blot vælge at udregne en korrelationskoefficient og teste, om den kunne være 0, men set i lyset af det næste spørgsmål, er dette ikke nødvendigt. Af pædagogiske årsager gør vi det dog alligevel.

Vi benytter `Analyze/Correlate/Bivariate` og sætter de to variable (`fht` og `mht`) over i `Variables` og sørger for, at `Pearson` er hakket af:

Vi får outputtet

Correlations

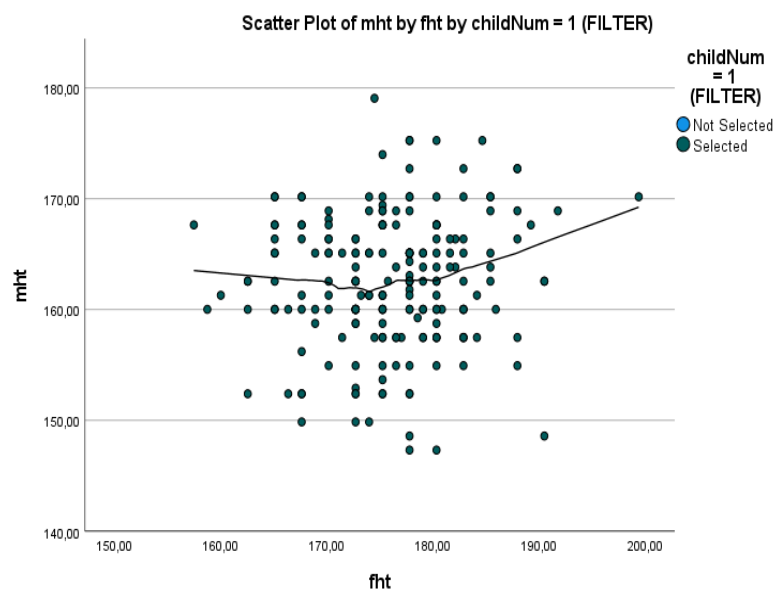
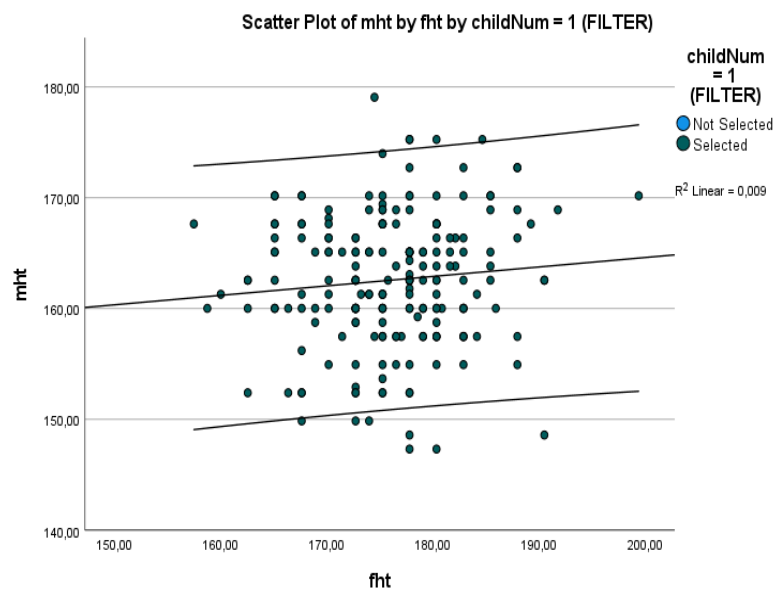
		fht	mht
fht	Pearson Correlation	1	,097
	Sig. (2-tailed)		,167
	N	205	205
mht	Pearson Correlation	,097	1
	Sig. (2-tailed)	,167	
	N	205	205

Vi finder altså en korrelationskoefficient (Pearson) på 0.097 og kan konstatere, at den ikke er statistisk signifikant forskellig fra 0 ($P=0.17$). Ud over det, kan vi ikke rigtigt bruge det til noget.

(b) *Hvor høj vil Prebens partner typisk være?*

Her er det oplagt at foretage en simpel lineær regressionsanalyse, hvor vi vælger moderens højde (**mht**) som outcome og faderens højde (**fht**) som forklarende variabel.

Først ser vi lige på en figur:



og outputtet fra regressionsanalysen bliver:

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,097 ^a	,009	,005	5,91252

a. Predictors: (Constant), fht

b. Dependent Variable: mht

ANOVA^a

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	67,264	1	67,264	1,924	,167 ^b
	Residual	7096,459	203	34,958		
	Total	7163,723	204			

a. Dependent Variable: mht

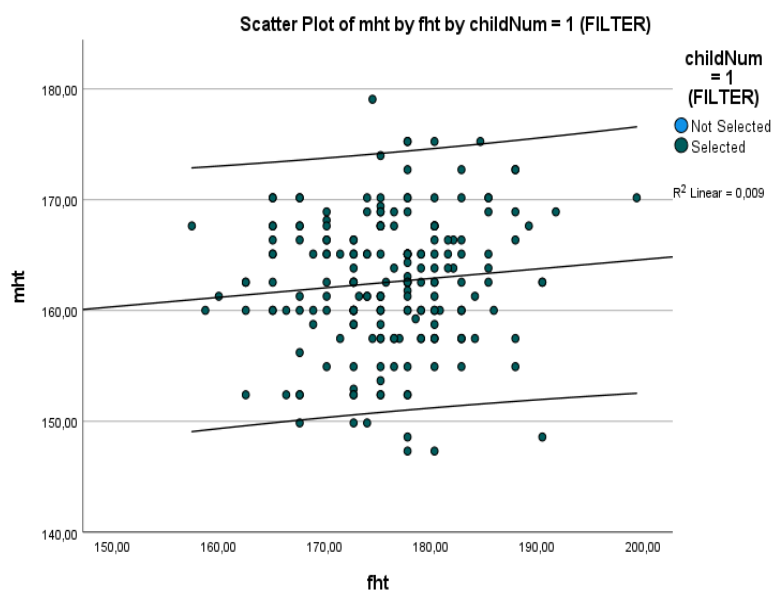
b. Predictors: (Constant), fht

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	147,529	10,847		,000
	fht	,085	,062	,097	,167

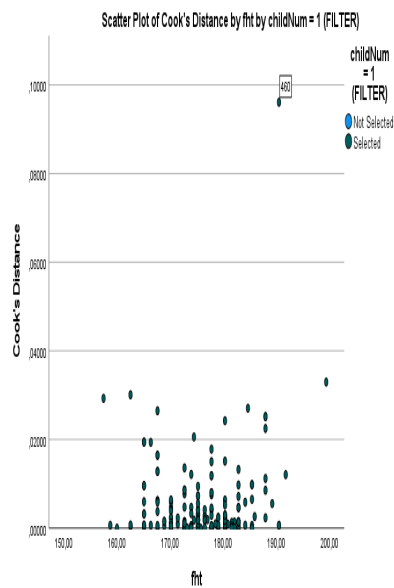
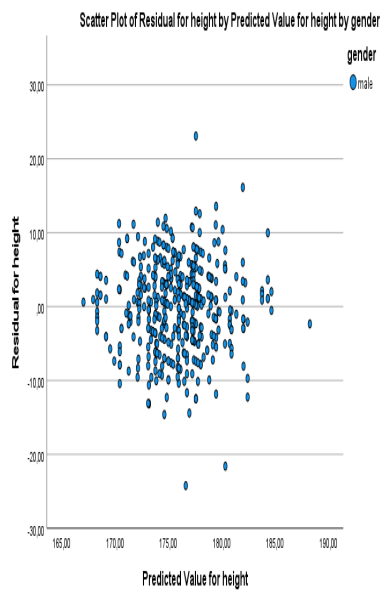
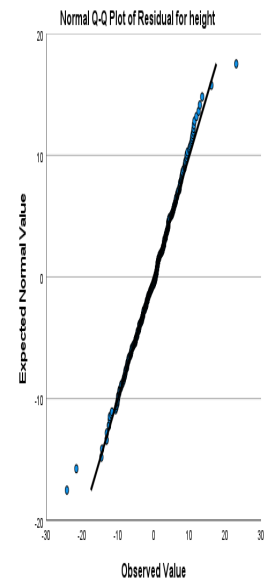
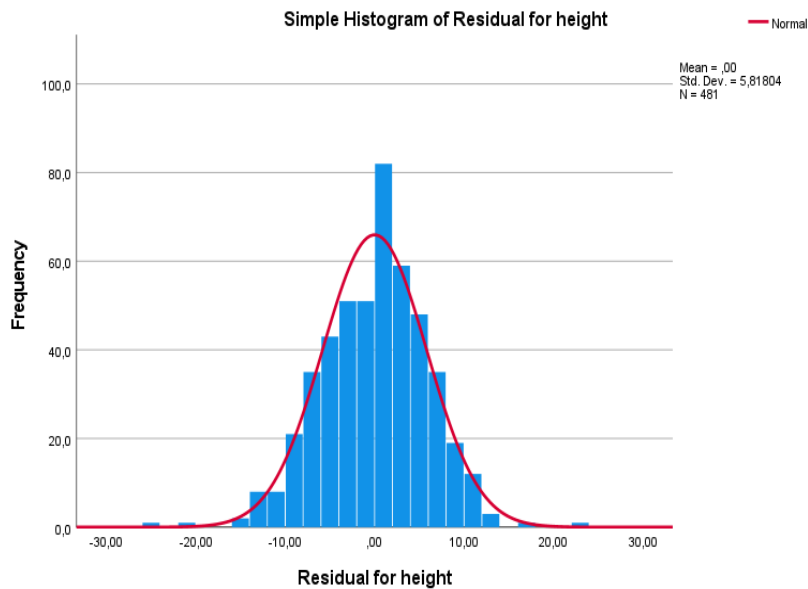
Coefficients ^a			
Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	126,141	168,917
	fht	-,036	,207

a. Dependent Variable: mht

Modellen illustreres med prediktionsgrænser:



og de tilhørende modelkontrol plots giver ikke anledning til bekymring:



Man kan dog se en enkelt observation med ret stor indflydelse på estimerterne. Ved at undersøge sagen nærmere finder man, at det drejer sig om denne observation:

Obs	idnr	family	fht	mht
5	16	5	190.5	148.59

altså ikke den meget høje mand, men en mand på 190.5 cm, der har en meget lav partner på kun 148.6 cm. Uden denne observation ville kurven formentlig gå lidt mere op til sidst.....

Som antydnet i den indledende figur (fra spørgsmål 4a) med loesskurven, finder vi i den simple lineære regression ingen tegn på sammenhæng mellem faders og moders højde, i hvert fald ikke signifikant ($P=0.17$). Bemærk, at denne P -værdi er præcis den samme som den, vi fandt ovenfor for Pearsons korrelationskoefficient.

Den estimerede sammenhæng ses at være

$$\text{mht} = 147.529 + 0.085 * \text{fht}$$

og for at finde den forventede højde for partneren til en mand på 172 cm kan vi nu blot indsætte og finde

$$\text{mht} = 147.529 + 0.085 * 172 = 162.15$$

(c) *Vil det være usædvanligt, hvis hun er højere end ham?*

Her skal vi igen bruge et prediktionsinterval, så vi finder spredningen omkring regressionslinien, altså Root MSE , som ses at være 5.91. Vi finder derfor et 95% prediktionsinterval til at være

$$162.21 \pm 2 \times 5.91 = (150.39, 174.03)$$

Det er derfor ikke specielt usædvanligt, at Prebens partner er mere end 172 cm høj.

5. *Prebens partner, Berit, er 175 cm høj:*

(a) *Opstil en formel til beregning af den mest sandsynlige højde for en søn af Preben og Berit.*

Nu ved vi mere, nemlig at Prebens partner er ret høj, nemlig 175 cm. Det er en del højere end forventet, jf spørgsmål 4b ovenfor.

Vi skal nu inddrage såvel faders som moders højde i en model til at beskrive sønners højde, så vi er nødt til at lave en multipel regressionsanalyse med såvel fars som mors højde som kovariater.

Da vi nu ikke har nogen kategoriske kovariater, kan vi benytte **Analyze/Regression/Linear**, hvor vi sætter **height** over i **Dependent** og både **fht** og **mht** over i **Independent(s)**

Herefter klikker vi **Statistics** og afkrydser **Confidence Intervals**, og i **Save** vælges residualer og predikterede værdier.

Herved får vi outputtet

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	mht, fht ^b	.	Enter

a. Dependent Variable: height

b. All requested variables entered.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,488 ^a	,238	,235	5,83020

a. Predictors: (Constant), mht, fht

b. Dependent Variable: height

Coefficients^a

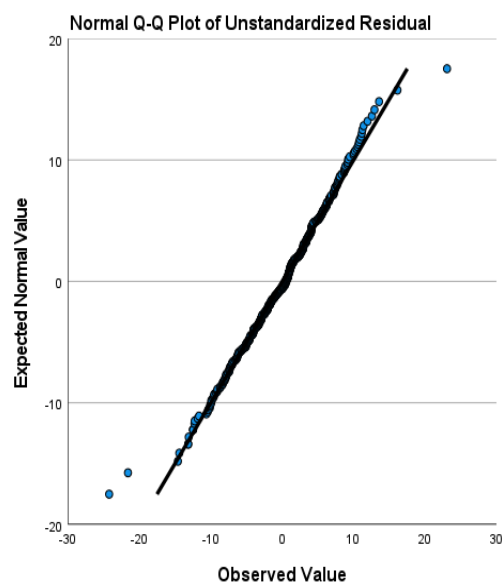
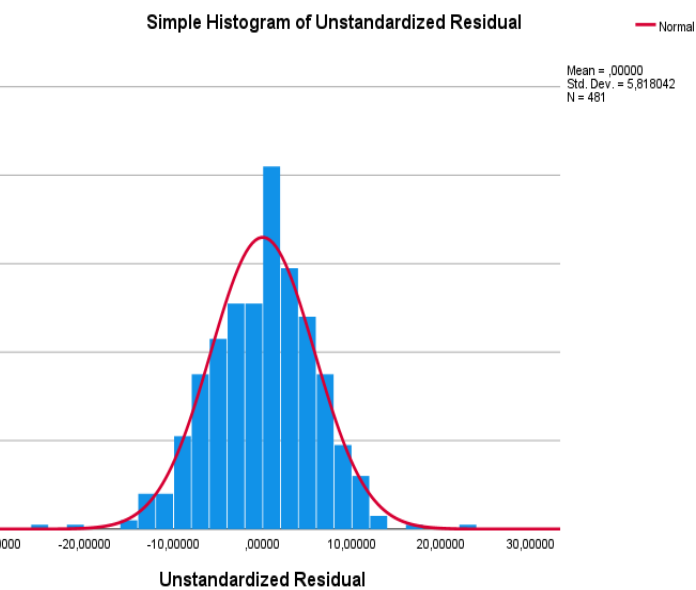
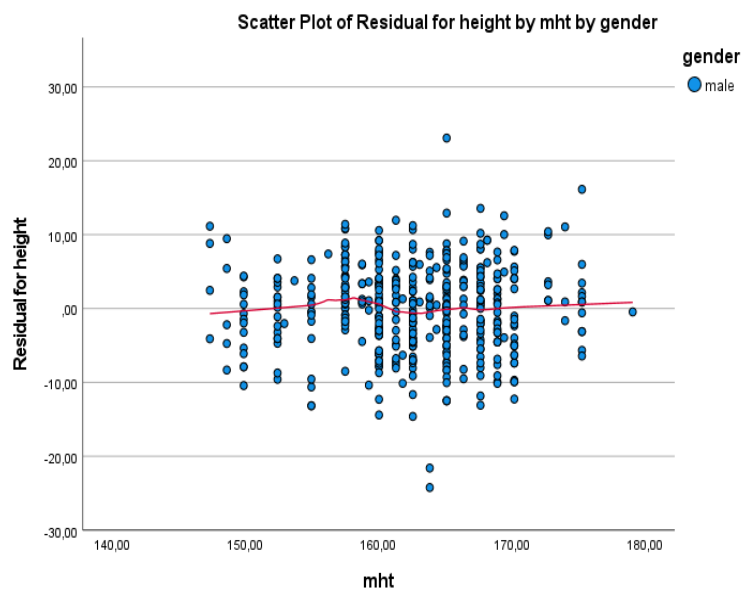
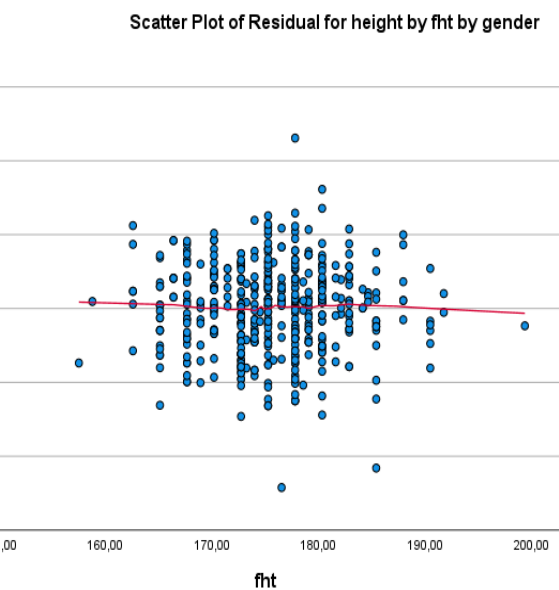
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49,055	10,401		4,716	,000
	fht	,418	,046	,367	9,154	,000
	mht	,329	,045	,291	7,258	,000

Coefficients^a

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	28,616	69,493
	fht	,328	,507
	mht	,240	,418

a. Dependent Variable: height

med tilhørende plots:



Vi bemærker, at såvel faders som moders højde er stærkt signifikante prediktorer for sønnens højde. Desuden ser vi, at 1 cm ekstra for faderen giver forventet 0.42 cm ekstra for sønnen, mens 1 ekstra mor-cm kun giver forventeligt 0.33 cm mere. Denne forskel ser vi nærmere på i et senere spørgsmål.

Formlen til beregning af den forventede højde for sønner af Preben

og Berit er altså

$$\text{height} = 49.055 + 0.418 \times \text{fht} + 0.329 \times \text{mht}$$

og når vi indsætter Prebens og Berits højder finder vi

$$49.055 + 0.418 \times 172 + 0.329 \times 175 = 178.526$$

Et 95% prediktionsinterval udregnes til

$$178.526 \pm 2 \times 5.83 = (166.87, 190.19)$$

Til sammenligning fik vi (i spørgsmål 3e) prediktionsintervallet (161.96, 186.52) cm uden vores viden om Berits højde.

- (b) *Er det sandsynligt, at en søn af dette par vil være 185 cm høj? Sammenlign til svaret fra spørgsmål 3e.*

Som vi kan se af prediktionsintervallet ovenfor, er det nu absolut indenfor det 95% forventelige, at en søn af Preben og Berit kan være 185 cm høj. Forskellen til spørgsmål 3e er, at vi nu ved, at Berit er en usædvanlig høj partner for Preben.

- (c) *Er afhængigheden af fædres og mødres højde den samme? Altså samme forventede forskel på sønners højde for en ekstra cm for hhv far og mor? Kvantificer den øgede højde for 1 far-cm i forhold til 1 mor-cm.*

Vi så i spørgsmål 5a på sønners højde som funktion af fars og mors højde, og her var $\hat{\beta}_F = 0.42 > 0.33 = \hat{\beta}_M$, men hvordan kan vi teste, om denne forskel på regressionskoefficienterne er signifikant?

For at gøre dette, benytter vi et lille trick, idet vi definerer en ny kovariat kaldet **parents=fhtmht+** og idet vi konstaterer, at

$$a \times \text{fht} + b \times \text{mht} = (a - b) \times \text{fht} + b \times \text{parents}$$

kan vi nu estimere forskellen $a - b$, altså forskellen på regressionskoefficienterne hørende til far og mor ved at benytte en model med **mht** skiftet ud med **parents**, altså med kovariaterne **parents** og **fht**.

I denne model får vi outputtet:

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	fht, parents ^b	.	Enter

a. Dependent Variable: height

b. All requested variables entered.

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49,055	10,401		4,716	,000
	parents	,329	,045	,428	7,258	,000
	fht	,089	,067	,078	1,324	,186

Coefficients^a

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	28,616	69,493
	parents	,240	,418
	fht	-,043	,221

a. Dependent Variable: height

Vi noterer os, at højde-øgningen svarende til 1 far-cm er 0.0888 cm større end for 1 mor-cm, med CI=(-0.043, 0.221) og P=0.19, altså ikke statistisk signifikant forskellig fra 0.

I materialet forekommer der jo mange søskende, og den sædvanlige antagelse om uafhængighed er derfor *ikke* opfyldt.

6. *Hvilke konsekvenser tror du, det har for besvarelsen af spørgsmålene 3f og 5c ovenfor? (Du skal ikke udføre analyser her, bare overveje konsekvenser for estimer og/eller standard errors).*

Vi gentager lige spørgsmål 3f:

Hvor meget lavere vil vi forvente, at en datter af Preben vil være?

Husk at benytte en model, der tillader at angive dette estimat med CI.

Nu laver vi en model svarende til den tidligere multiple regression, dog nu med en tilfældig effekt af familie, således at højden for søskende modelleres som værende korrelerede:

Til dette formål benyttes **Analyze/Mixed Models/Linear**, hvor vi sætter **family** over i **Subjects** og klikker **Continue**

Nu sættes **height** i **Dependent Variable**, **gender** i **Factor(s)** og **fht** i **Covariates**. Under **Fixed** sættes både **gender** og **fht** over i **Model** (som **Main Effects**), og under **Random** sættes **family** over i **Model**, og under **Statistics**, afkrydser vi **Parameter estimates**.

Vi får så outputtet:

Model Dimension^a

		Number of Levels	Covariance Structure	Number of Parameters
Fixed Effects	Intercept	1		1
	gender	2		1
	fht	1		1
Random Effects	family ^b	205	Variance Components	1
Residual				1
Total		209		5

a. Dependent Variable: height.

b. As of version 11.5, the syntax rules for the RANDOM subcommand have changed. Your command syntax may yield results that differ from those produced by prior versions. If you are using version 11 syntax, please consult the current syntax reference guide for more information.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% ... Lower Bound
Intercept	106,062363	7,512803	206,834	14,118	,000	91,250873
[gender=female]	-13,218298	,346686	859,156	-38,128	,000	-13,898749
[gender=male]	0 ^b	0	.	.	.	.
fht	,397979	,042682	207,377	9,324	,000	,313832

Estimates of Fixed Effects^a

Parameter	95% Confidence Upper Bound
Intercept	120,873852
[gender=female]	-12,537847
[gender=male]	.
fht	,482127

a. Dependent Variable: height.

b. This parameter is set to zero because it is redundant.

Estimates of Covariance Parameters^a

Parameter		Estimate	Std. Error
Residual		24,538138	1,270926
family	Variance	9,096683	1,540056

a. Dependent Variable: height.

Vi bemærker her, at en søster estimeres til at være 13.22 cm lavere end sin bror, CI=(12.54, 13.90). Dette er kun en anelse anderledes (og snævrere) end det tidligere fundne: 13.16, CI=(12.41, 13.90).

Vi får et snævrere interval, fordi vi nu erkender, at en sammenligning mellem bror og søster foregår *indenfor* en familie, svarende til en *parret sammenligning*. Effekten af parringen er dog lille, idet korrelationen mellem søskendes højde kun estimeres til

$$\frac{9.0967}{9.0967 + 24.5381} = 0.27$$

Vi skal nu se på spørgsmål 5c:

Er afhængigheden af fædres og mødres højde den samme? Alt-så samme forventede forskel på sønners højde for en ekstra cm for hhv far og mor? Kvantificer den øgede højde for 1 far-cm i forhold til 1 mor-cm.

og den tilsvarende analyse, med inddragelse af korrelation fås ganske som beskrevet for analysen ovenfor ved at benytte **Analyze/Mixed Models/Linear**, hvor vi sætter **family** over i **Subjects** og klikker **Continue**

Igen sættes **height** i **Dependent Variable**, men nu sættes intet under **Factor(s)**, men til gengæld såvel **fht** og **mht** i **Covariates**. Under **Fixed** sættes både **fht** og **mht** over i **Model** (som **Main Effects**), og

under **Random** sættes **family** over i **Model**, og under **Statistics**, afkrydser vi **Parameter estimates**.

Så får vi outputtet:

Model Dimension^a

		Number of Levels	Covariance Structure	Number of Parameters
Fixed Effects	Intercept	1		1
	snit	1		1
	halvdiff	1		1
Random Effects	family ^b	179	Variance Components	1
Residual				1
Total		182		5

a. Dependent Variable: height.

b. As of version 11.5, the syntax rules for the RANDOM subcommand have changed. Your command syntax may yield results that differ from those produced by prior versions. If you are using version 11 syntax, please consult the current syntax reference guide for more information.

Estimates of Fixed Effects^a

Parameter	Estimate	Std. Error	df	t	Sig.	95% ... Lower Bound
Intercept	52,642424	12,262341	179,081	4,293	,000	28,445154
snit	,724921	,072497	177,687	9,999	,000	,581854
halvdiff	,103810	,078808	173,621	1,317	,189	-,051734

Estimates of Fixed Effects^a

Parameter	95% Confidence . Upper Bound
Intercept	76,839693
snit	,867988
halvdiff	,259354

a. Dependent Variable: height.

Estimates of Covariance Parameters^a

Parameter		Estimate	Std. Error
Residual		27,243723	2,161992
family	Variance	6,894241	2,013936

a. Dependent Variable: height.

Her finder vi en forskel på 0.1038 på de to regressionsestimater, mod 0.0888 uden hensyntagen til korrelationen. Usikkerheden på estimatet er øget lidt her (fra 0.0670 før til 0.0788 nu), svarende til, at dette er en analyse, der *ikke* foregår indenfor familier. På grund af korrelationen indenfor familier har vi ikke helt så megen information, som vi troede, og det resulterer i en større usikkerhed på estimatet.

Igen er korrelationen dog ganske beskeden:

$$\frac{6.8942}{6.8942 + 27.2437} = 0.20$$