

Hjemmeopgavebesvarelse i SAS

Basal statistik for sundhedsvidenskabelige forskere, efterår 2021

Et gammelt studie af Galton, 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 i samme omgang redefinerer vi variabelen **gender** ved at indlæse den i **sex** og derefter lave **gender** til en karaktervariabel. Desuden indlæses **family** som karaktervariabel, da der i min version var en familie med nummer 136A.

```
data ind;
infile "data_til_opgave.txt" firstobs=2;
input idnr family $ father mother sex childNum childHeight;

if sex=2 then gender="female";
if sex=1 then gender="male";
run;
```

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:

```
proc means mean min max data=ind;
  class gender;
  var father mother childNum childHeight;
run;
```

hvorved vi får:

The MEANS Procedure

gender	N Obs	Variable	Mean	Minimum	Maximum
female	453	father	69.2602649	62.0000000	78.5000000
		mother	64.1487859	58.0000000	70.5000000
		childNum	4.9955850	1.0000000	15.0000000
		childHeight	64.1039735	56.0000000	70.5000000
male	481	father	69.1376299	62.0000000	78.5000000
		mother	64.0332640	58.0000000	70.5000000
		childNum	2.2577963	1.0000000	10.0000000
		childHeight	69.2340956	60.0000000	79.0000000

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 indfører linierne med de nye 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 variabelen `childNum` mon betyder.

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

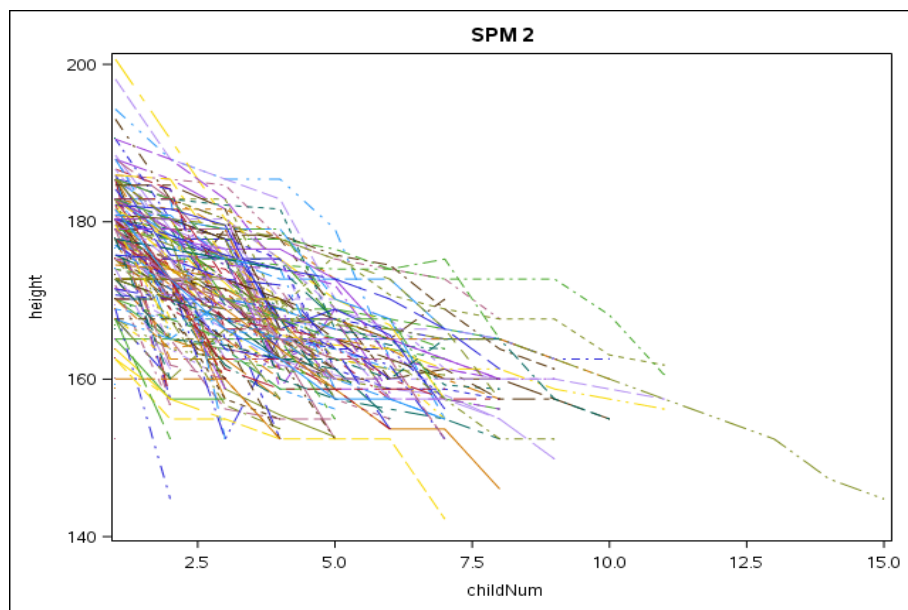
FAMILY HEIGHTS. (from REF.)
(add the inches to every entry on this Table)

	Father	Mother	Son's order of height	Daughter's order of height
1	17.5	7.0	12.2	9.2, 9.0, 9.0
2	18.5	6.5	13.5, 13.5	8.5, 8.5
3	18.0	about 4.0	11.0	8.0
4	18.0	4.0	10.5, 12.5	7.0, 4.5, 3.0
5	18.0	4.5	12.0, 9.0, 8.0	6.5, 2.5, 2.5
6	14.0	2.0		4.5
7	14.0	2.0	10.5, 14.0, 13.0, 13.0	10.5, 4.0
8	14.0	6.5		10.5, 2.0, 6.0
9	14.5	6.0		6.0
10	14.0	5.5		5.5
11	14.0	2.0	14.0, 10.0	2.0, 2.0, 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	8.5	11.0, 11.5	6.7
16	13.0	about 7.0	12.0, 10.5, 14.5, 10.5, 14.5	2.7, 6.5, 4.5, 3.5
17	13.0	4.5	14.0, 12.0, 11.5, 2.5	6.5, 2.5
18	12.0	4.0		6.0, 4.5, 4.0
19	12.2	3.0		2.7
20	12.7	4.0	13.5, 13.0, 12.7	10.0, 9.0, 2.5, 2.0, 6.0
21	12.0	2.0	13.0	2.5, 2.0
22	12.0	about 7.0	12.0, 11.0	7.0
23	12.0	5.0	14.2, 10.5, 11.5	6.0, 3.5, 2.0, 2.0
24	12.0	5.5		5.5

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:

```
ods graphics / imagename="spm2";
series Y=height X=childNum / group=family;
run;
```



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).

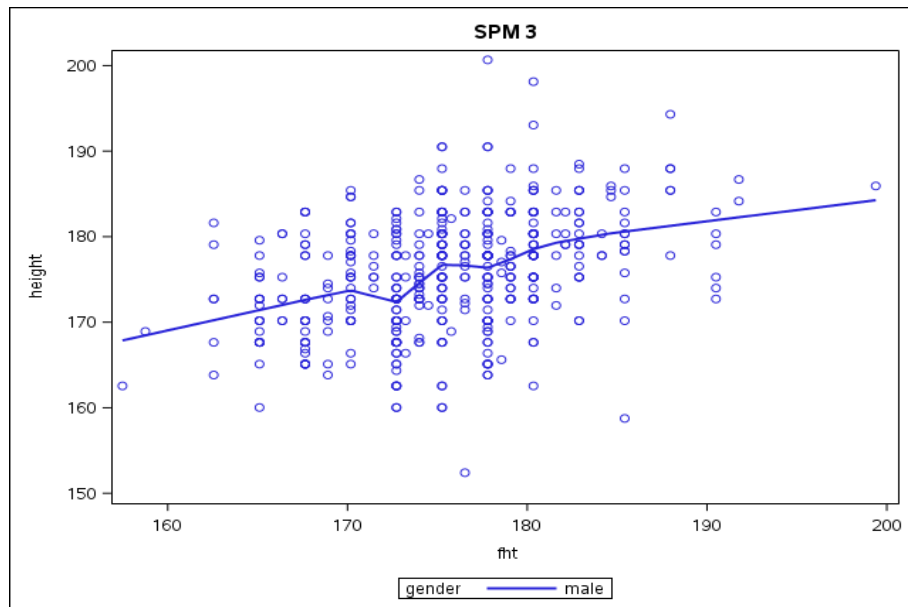
Vi har i disse data op til 15 børn fra de enkelte familier (se figuren ovenfor), og vi må formode, at der er en vis (positiv) korrelation mellem disse børns højde. Noget af denne forklares givetvis ud fra moderens og faderens højde, men der kan nemt være uforklaret korrelation f.eks. pga. samme kostvaner 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.*

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:

```
ods graphics / imagename="spm3a-loess";  
proc sgplot data=a1; where gender="male";  
loess Y=height X=fht / group=gender;  
run;  
ods graphics off;
```



Da der her ses en rimeligt lineær sammenhæng mellem fædres og sønners højde, går vi videre med en simpel lineær regression, idet vi også tilføjer to **estimate**-sætninger, som vi skal bruge i senere spørgsmål:

```
ods graphics / imagename="spm3a-glm";
proc glm plots=all data=a1; where gender="male";
class gender;
model height = fht / solution clparm;
estimate "son with father=172" intercept 1 fht 172;
estimate "5 cm extra for father" fht 5;
run;
ods graphics off;
```

Dette giver outputtet

The GLM Procedure

Class Level Information

Class	Levels	Values
gender	1	male

Number of Observations Read 481
 Number of Observations Used 481

Dependent Variable: height

R-Square	Coeff Var	Root MSE	height Mean
0.153965	3.489596	6.136615	175.8546

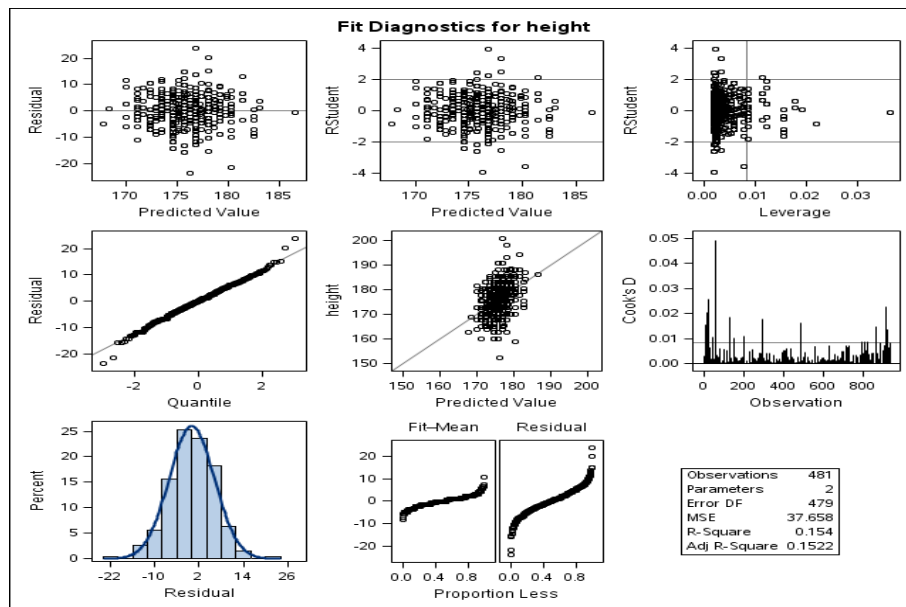
Parameter	Estimate	Standard Error	t Value	Pr > t
son with father=172	174.242844	0.32877380	529.98	<.0001
5 cm extra for father	2.232613	0.23912731	9.34	<.0001

Parameter	95% Confidence Limits	
son with father=172	173.596827	174.888861
5 cm extra for father	1.762745	2.702481

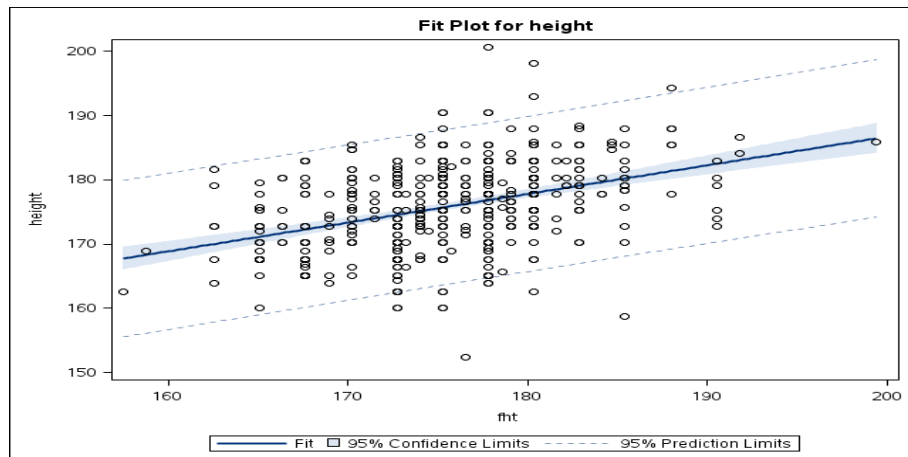
Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	97.44095582	8.40326896	11.60	<.0001
fht	0.44652260	0.04782546	9.34	<.0001

Parameter	95% Confidence Limits	
Intercept	80.92913013	113.95278151
fht	0.35254897	0.54049624

med tilhørende modelkontrol plots:



samt regressions-grafik:



Analysen viser en stærkt signifikant (positiv, selvfølgelig) sammenhæng mellem faders og sønners højde, med estimeret regressionskoefficient 0.4465 (0.0478), og $CI=(0.3525, 0.5405)$. Det betyder, at for hver ekstra far-cm, forventer vi at sønnen er 0.4465 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

$$\text{height} = 97.44 + 0.4465 * \text{fht}$$

- (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 0.4465 cm, med $CI=(0.3525, 0.5405)$ cm. Vi skal så bare gange dette op med 5, og finder den forventede forskel på A's og B's sønner til 2.23 cm, med $CI=(1.76, 2.70)$ cm.

Dette ses også af den ene af de to **estimate**-sætninger:

```
estimate "5 cm extra for father" fht 5;
```

Parameter	Estimate	Standard Error	t Value	Pr > t
5 cm extra for father	2.232613	0.23912731	9.34	<.0001

Parameter	95% Confidence Limits	
5 cm extra for father	1.762745	2.702481

(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:

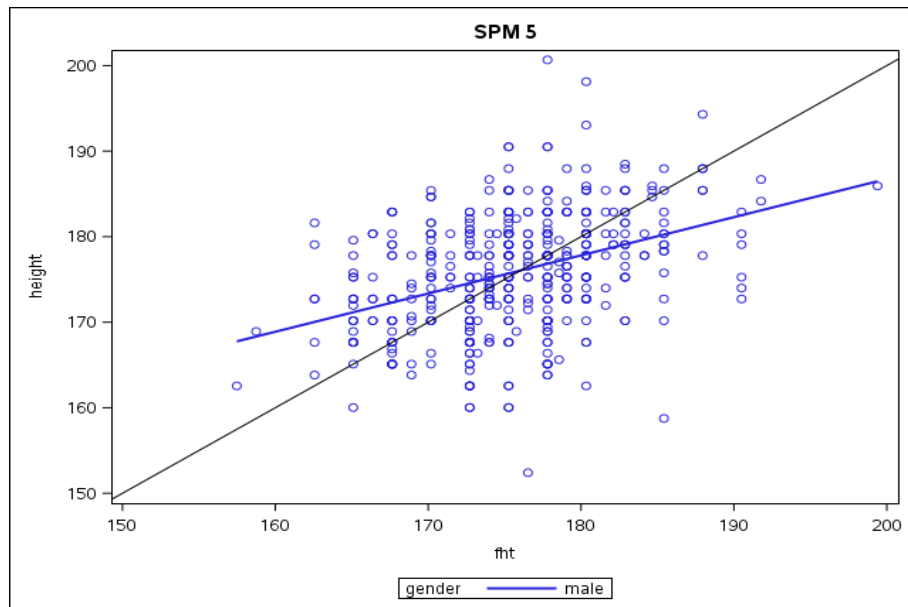
$$x = 97.44 + 0.4465 * x$$

og vi finder

$$x = \frac{97.44}{(1 - 0.4465)} = 176.04 \text{ (cm)}$$

Vi kan også illustrere det på forskellig vis, dels med et plot, hvor vi har indlagt en identitetslinie:

```
ods graphics / imagename="spm5-identity";
proc sgplot data=a1; where gender="male";
  loess Y=height X=fht / group=gender;
  lineparm x=0 y=0 slope=1;
run;
ods graphics off;
```



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

Alternativt kan vi definere forskellen mellem fars og søns højde ved (på passende sted) at tilføje sætningen

```
child_diff=height-fht;
```

og dernæst f.eks tegne denne differens op mod faderens højde, incl. en vandret linie i 0:

```
ods graphics / imagename="spm5-diff-loess";
proc sgplot data=a1; where gender="male";
loess Y=child_diff X=fht / group=gender;
refline 0 / axis=y;
run;
ods graphics off;
```

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:

$$97.44 + 0.4465 * 172 = 174.24 \text{ (cm)}$$

eller vi kan blot afløse fra `estimate`-sætningen i koden ovenfor:

```
estimate "son with father=172" intercept 1 fht 172;
```

Parameter	Estimate	Standard		t Value	Pr > t
		Error			
son with father=172	174.242844	0.32877380	529.98	<.0001	
95% Confidence Limits					
son with father=172	173.596827	174.888861			

For at svare på spørgsmålet, er det imidlertid vigtigt at gøre sig klart, at det her *ikke* er et konfidensinterval, vi skal bruge, men et prediktionsinterval.

For at finde et 95% prediktionsinterval for højden af Prebens sønner, skal vi bruge spredningen omkring regressionslinien, som ses ovenfor under RootMSE: $6.1366 \approx 6.14$, så vi udregner:

$$174.24 \pm 2 \times 6.14 = (161.96, 186.52)$$

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:

```
ods graphics / imagename="spm3e-glm";
proc glm plots=all data=a1; where gender="female";
class gender;
model height = fht / solution clparm;
estimate "daughter with father=172" intercept 1 fht 172;
run;
ods graphics off;
```

Herved får vi outputtet

```
The GLM Procedure

      Class Level Information

Class          Levels    Values
gender          1      female

Number of Observations Read      453
Number of Observations Used      453

Dependent Variable: height
```

R-Square	Coeff Var	Root MSE	height Mean
0.183555	3.324075	5.412394	162.8241

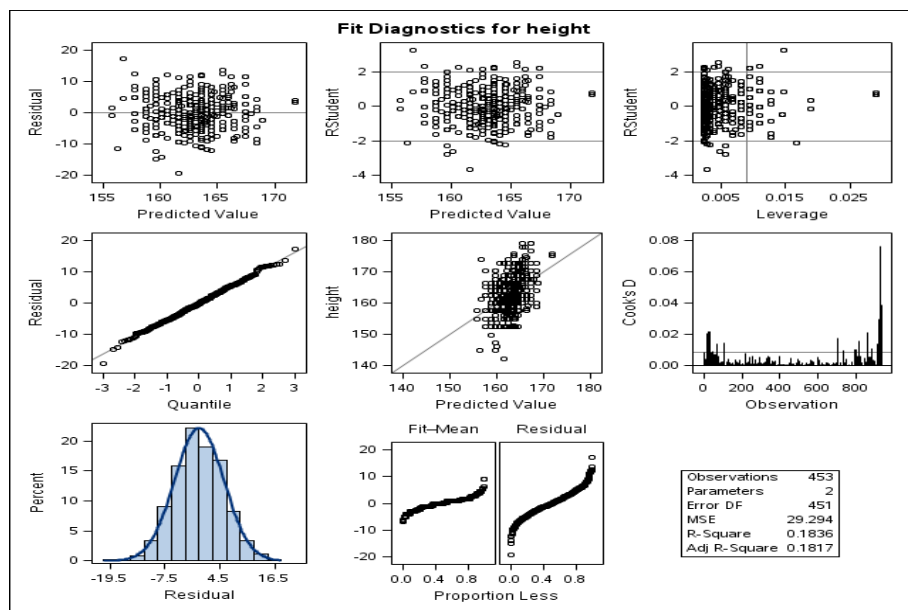
Parameter	Estimate	Standard Error	t Value	Pr > t
daughter with father=172	161.328984	0.29447022	547.86	<.0001

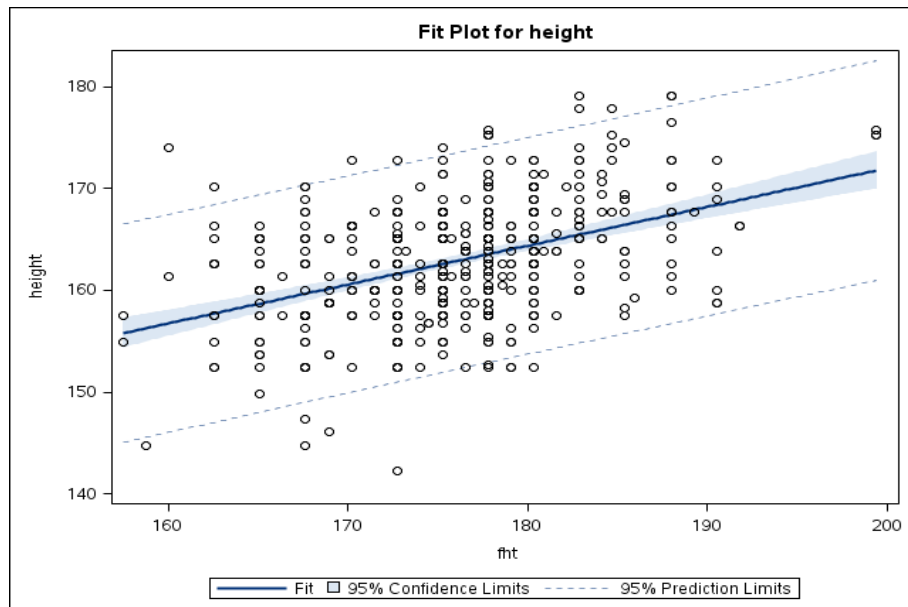
Parameter	95% Confidence Limits	
daughter with father=172	160.750280	161.907688

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	95.74523480	6.66644367	14.36	<.0001
fht	0.38130087	0.03786693	10.07	<.0001

Parameter	95% Confidence Limits	
Intercept	82.64408695	108.84638265
fht	0.30688333	0.45571840

med tilhørende plots:





Vi bemærker igen, at faders højde er en stærkt signifikant prediktorer, 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:

```
ods graphics / imagename="spm3e-multipel-gender";
proc glm plots=all data=a1;
class gender;
model height = gender fht / solution clparm;
run;
ods graphics off;
```

Herved får vi outputtet

The GLM Procedure

Class Level Information

Class	Levels	Values
gender	2	female male

Number of Observations Read	934
Number of Observations Used	934

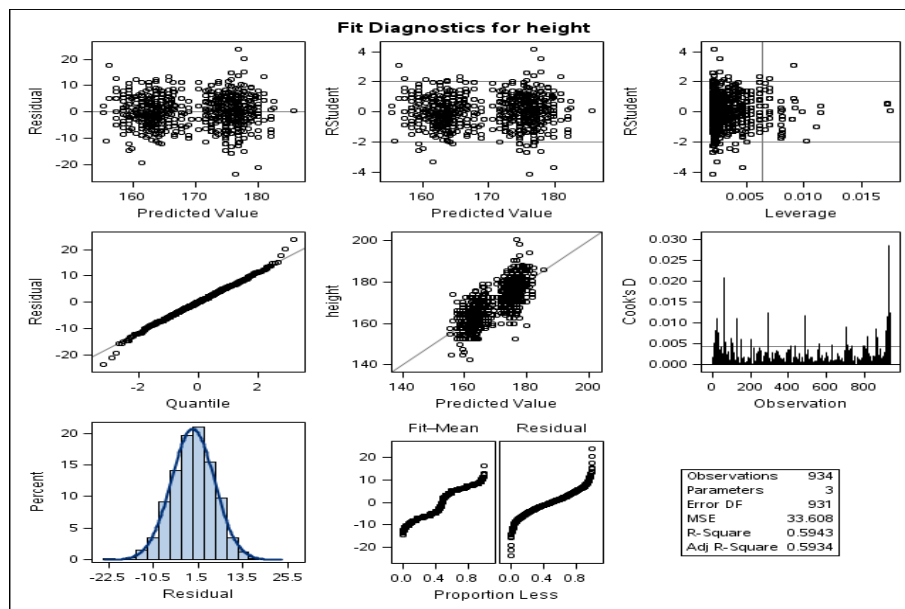
Dependent Variable: height

R-Square	Coeff Var	Root MSE	height Mean
0.594256	3.419476	5.797198	169.5347

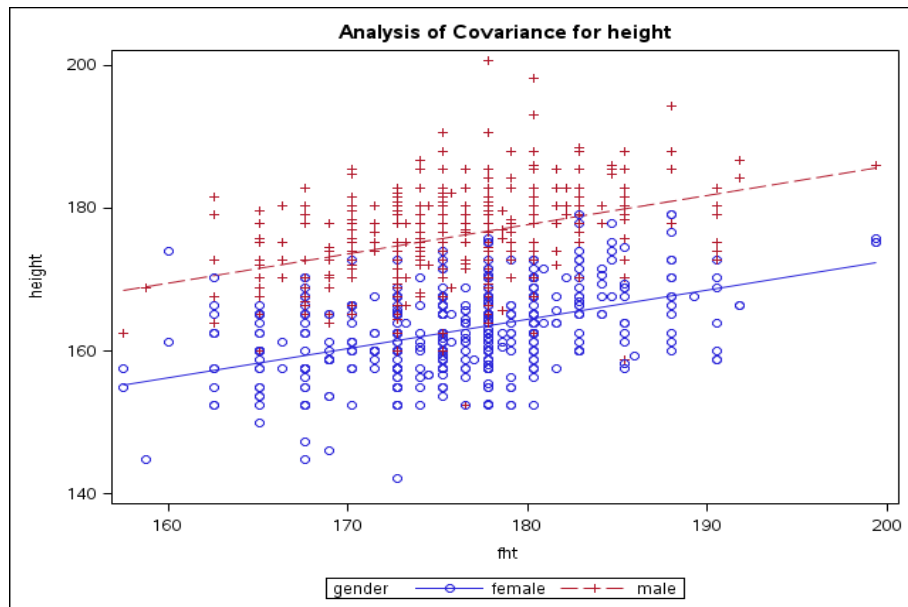
Parameter		Estimate	Standard Error	t Value	Pr > t
Intercept		103.7832589	5.30676284	19.56	<.0001
gender	female	-13.1583489	0.37966722	-34.66	<.0001
gender	male	0.0000000	.	.	.
fht		0.4104067	0.03018159	13.60	<.0001

Parameter		95% Confidence Limits	
Intercept		93.3686555	114.1978624
gender	female	-13.9034517	-12.4132462
gender	male	.	.
fht		0.3511749	0.4696385

med tilhørende modelkontrol plots:



samt illustration af modellen:



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**, og indsætter samtidig en **estimate**-sætning til at se på den specifikke forskel på Prebens datter og søn:

```
ods graphics / imagename="spm3f-glm-interaktioner";
proc glm plots=all data=a1;
class gender;
model height = fht gender gender*fht / solution clparm;
estimate "Prebens datter vs søn" gender 1 -1 gender*fht 172 -172;
```

```
run;
ods graphics off;
```

Herved får vi outputtet

The GLM Procedure

Class Level Information

Class	Levels	Values
gender	2	female male

Number of Observations Read	934
Number of Observations Used	934

Dependent Variable: height

R-Square	Coeff Var	Root MSE	height Mean
0.594759	3.419193	5.796718	169.5347

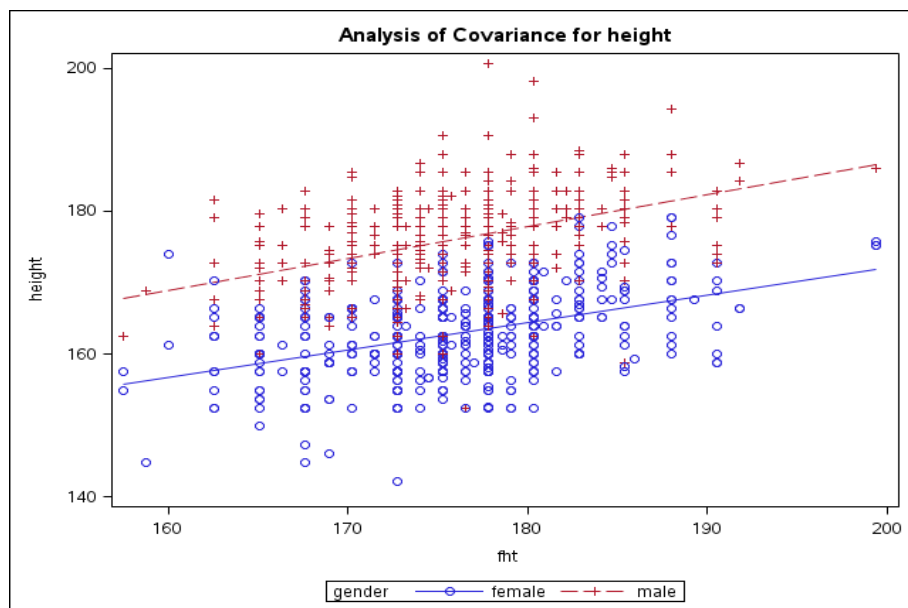
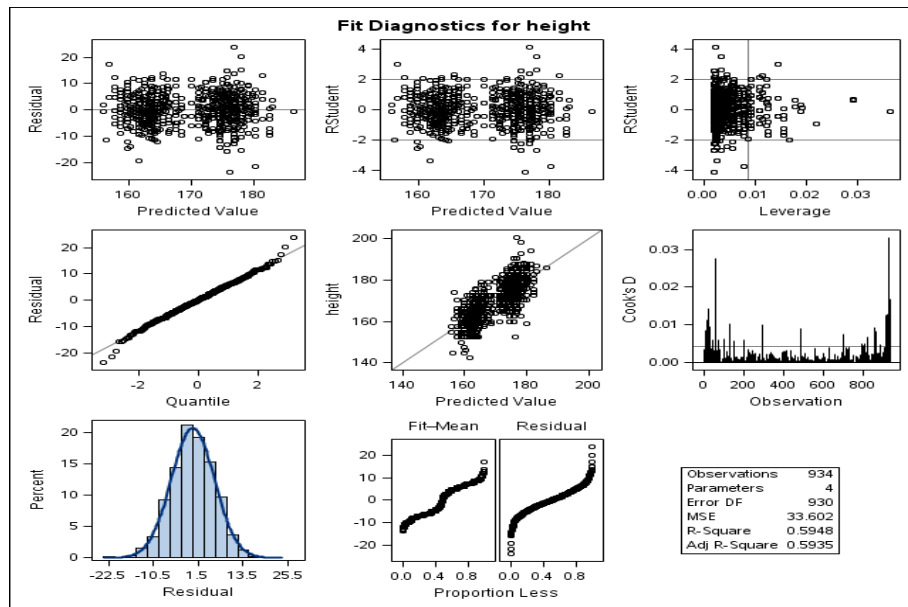
Parameter	Estimate	Standard Error	t Value	Pr > t
Prebens datter vs søn	-12.9138596	0.44262201	-29.18	<.0001

Parameter	95% Confidence Limits	
Prebens datter vs søn	-13.7825133	-12.0452059

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	97.44095582 B	7.93782552	12.28	<.0001
fht	0.44652260 B	0.04517649	9.88	<.0001
gender female	-1.69572102 B	10.67642404	-0.16	0.8738
gender male	0.00000000 B	.	.	.
fht*gender female	-0.06522174 B	0.06070986	-1.07	0.2830
fht*gender male	0.00000000 B	.	.	.

Parameter	95% Confidence Limits	
Intercept	81.86282973	113.01908191
fht	0.35786293	0.53518228
gender female	-22.64839624	19.25695420
gender male	.	.
fht*gender female	-0.18436594	0.05392246
fht*gender male	.	.

med tilhørende plots:



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.

Vi bemærker dog, at der specifikt for Prebens børn (eller snarere for børn af fædre på 172 cm) gælder, at en datter forventeligt kun er 12.91 cm lavere end sin bror, $CI=(12.05, 13.78)$.

```
estimate "Prebens datter vs søn" gender 1 -1 gender*fht 172 -172;
```

Parameter	Estimate	Standard Error	t Value	Pr > t
Prebens datter vs søn	-12.9138596	0.44262201	-29.18	<.0001

Parameter	95% Confidence Limits	
Prebens datter vs søn	-13.7825133	-12.0452059

4. Nu ser vi på højden hos et forældrepar.

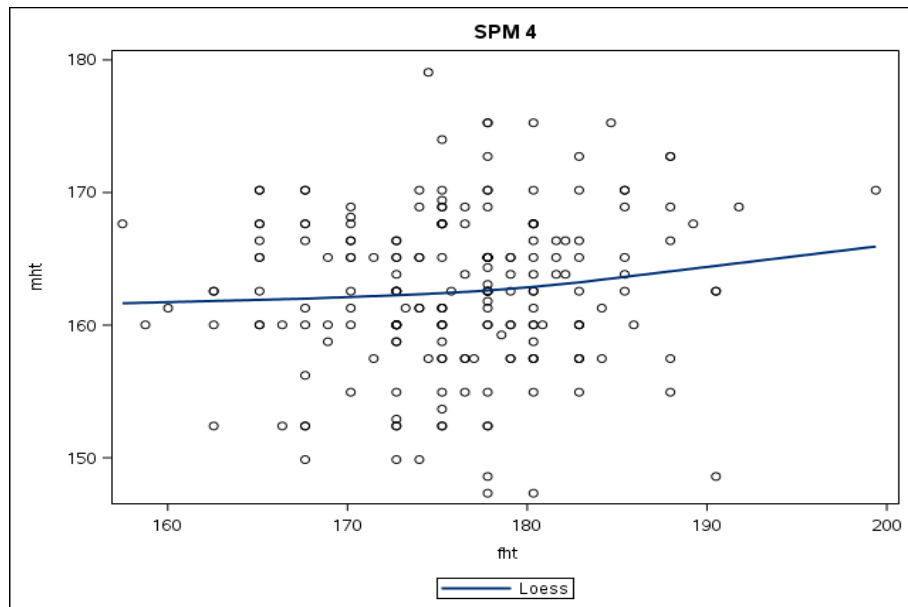
Husk her, at hvert forældrepar typisk optræder adskillige gange i datasættet.

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

Vi starter med at plotte moders højde mod faders højde, med en indlagt loess-kurve:

```
ods graphics / imagename="spm4-loess";
proc sgplot data=a1; where childNum=1;
loess Y=mht X=fht;
run;
ods graphics off;
```

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.

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 skriver altså:

```
ods graphics / imagename="spm4-corr";
proc corr data=a1; where childNum=1;
var mht fht;
run;
ods graphics off;
```

og får outputtet

```
The CORR Procedure
  2 Variables:   mht       fht

                        Simple Statistics

Variable           N           Mean        Std Dev        Sum
mht                205       162.56496       5.92590       33326
fht                205       176.06289       6.72387       36093

                        Simple Statistics
Variable      Minimum      Maximum
mht          147.32000      179.07000
fht          157.48000      199.39000

Pearson Correlation Coefficients, N = 205
Prob > |r| under H0: Rho=0

           mht           fht
mht      1.00000       0.09690
           0.1669
fht      0.09690       1.00000
           0.1669
```

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:

```
ods graphics / imagename="spm4-glm";
proc glm plots=all data=a1; where childNum=1;
model mht = fht / clparm;
estimate "typisk partner M=172" intercept 1 fht 172;
output out=ny p=yhat r=resid cookd=cook;
run;
ods graphics off;
```

Vi finder så:

The GLM Procedure

Number of Observations Read 205
Number of Observations Used 205

Dependent Variable: mht

R-Square Coeff Var Root MSE mht Mean
0.009389 3.637022 5.912523 162.5650

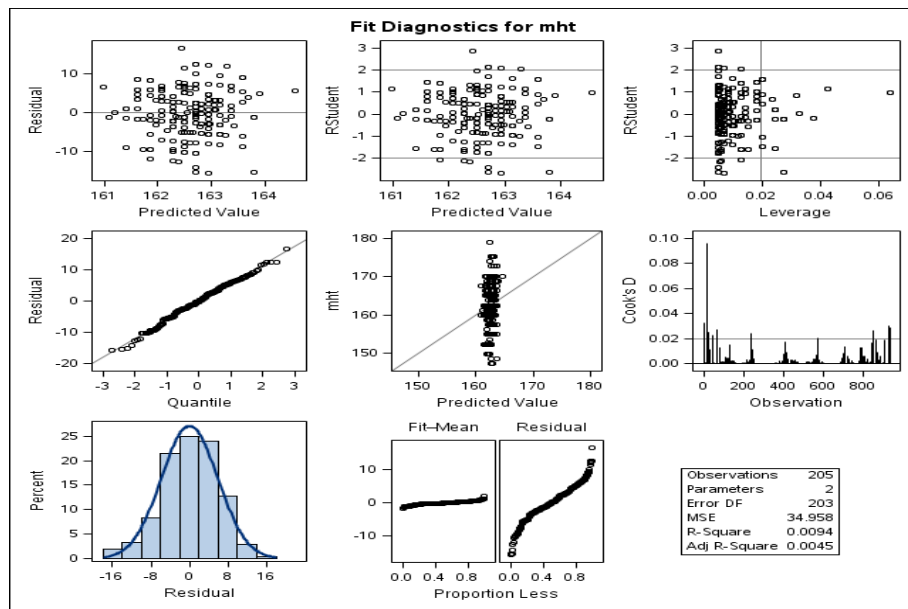
Parameter	Estimate	Standard Error	t Value	Pr > t
typisk partner M=172	162.217987	0.48279771	336.00	<.0001

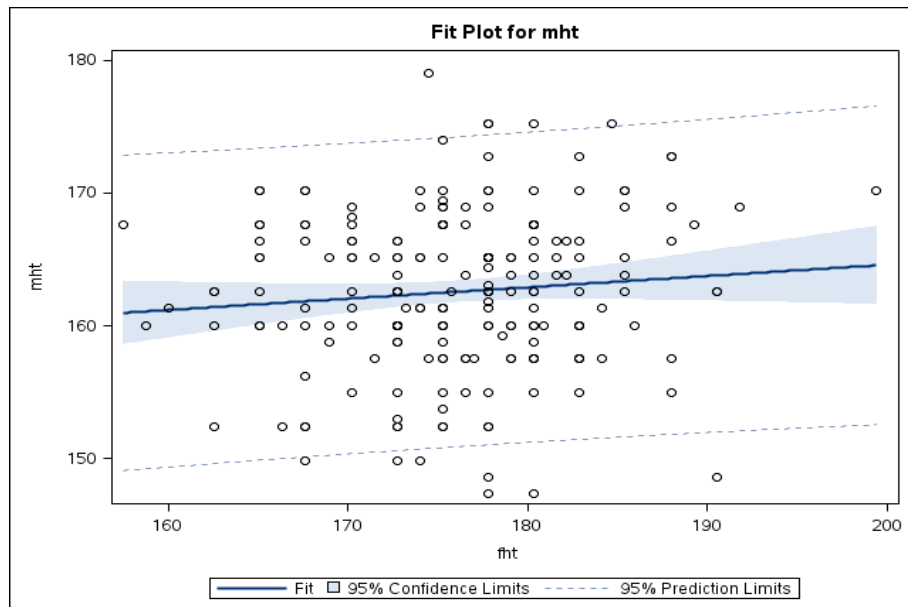
Parameter	95% Confidence Limits	
typisk partner M=172	161.266046	163.169929

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	147.5292589	10.84728737	13.60	<.0001
fht	0.0853996	0.06156564	1.39	0.1669

Parameter	95% Confidence Limits	
Intercept	126.1414580	168.9170599
fht	-0.0359905	0.2067897

med tilhørende modelkontrol plots, som ikke giver anledning til bekymring:





Man kan dog se en enkelt observation med ret stor indflydelse på estimerterne. Ved hjælp af `output`-sætningen i koden ovenfor kan vi finde ud af, hvilken observation, det drejer sig om:

```
proc print data=ny; where cook>0.05;
var idnr family fht mht;
run;
```

hvorved vi finder

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 loess-kurven, 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

$$\text{mht} = 147.53 + 0.085 * \text{fht}$$

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

$$\text{mht} = 147.53 + 0.085 * 172 = 162.15$$

På grund af afrunding er dette ikke eksakt det samme som det (mere nøjagtige), som vi får fra `estimate`-sætningen

```
estimate "typisk partner M=172" intercept 1 fht 172;
```

som ses at give 162.22 cm, med konfidensinterval CI=(161.27, 163.17) cm.

(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:

```
ods graphics / imagename="spm5-multipel";
proc glm plots=all data=a1; where gender="male";
class gender;
model height = fht mht / solution clparm;
estimate "son with father=172, mother 175" intercept 1 fht 172 mht 175;
run;
ods graphics off;
```

Herved får vi outputtet

The GLM Procedure

Class Level Information

Class	Levels	Values
gender	1	male

Number of Observations Read	481
Number of Observations Used	481

Dependent Variable: height

R-Square	Coeff Var	Root MSE	height Mean
0.237938	3.315353	5.830201	175.8546

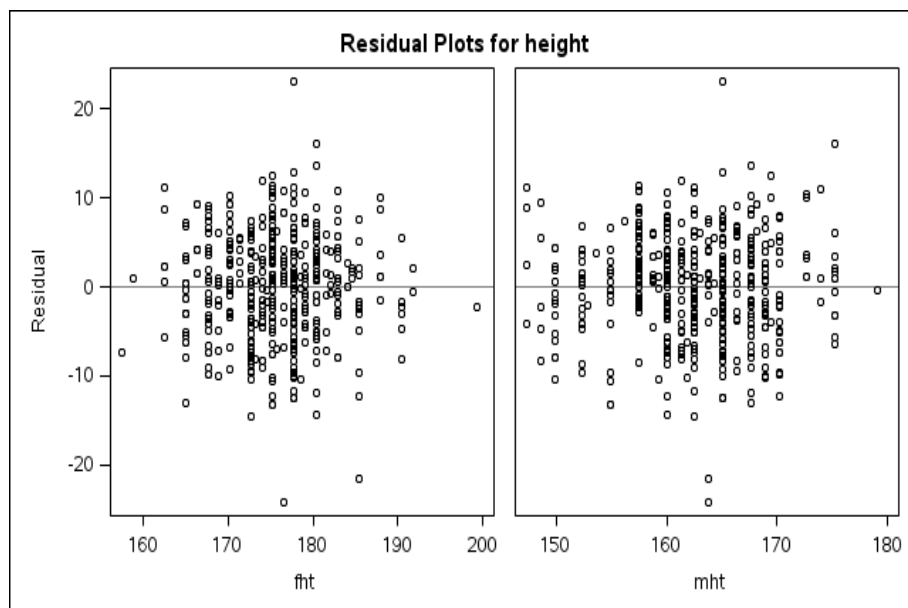
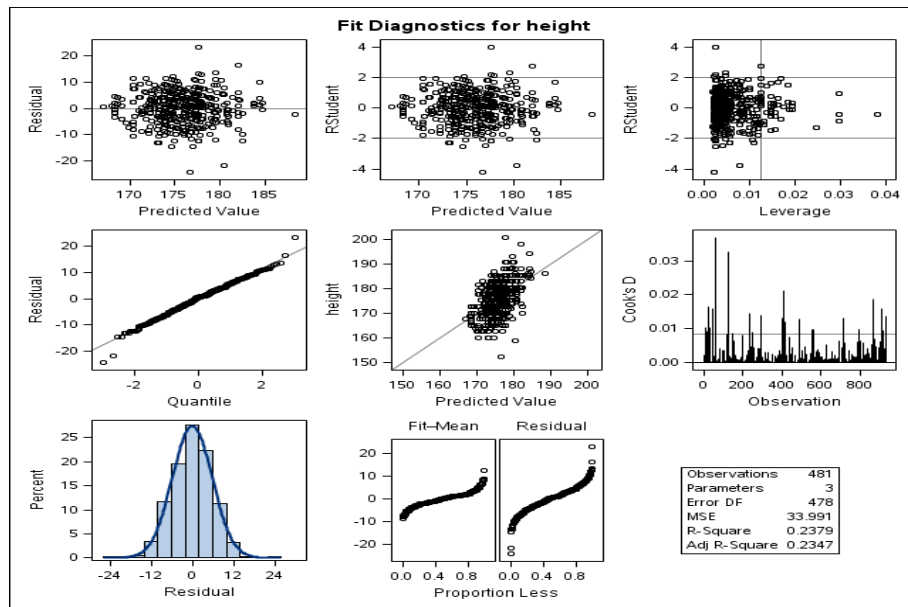
Parameter	Estimate	Standard Error	t Value
son with father=172, mother 175	178.409563	0.65359247	272.97

Parameter	Pr > t	95% Confidence Limits
son with father=172, mother 175	<.0001	177.125294 179.693833

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	49.05454497	10.40137228	4.72	<.0001
fht	0.41755622	0.04561239	9.15	<.0001
mht	0.32877342	0.04530087	7.26	<.0001

Parameter	95% Confidence Limits
Intercept	28.61648010 69.49260984
fht	0.32793064 0.50718180
mht	0.23975996 0.41778688

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, medens 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

$$\text{height} = 49.054 + 0.418 \times \text{fht} + 0.329 \times \text{mht}$$

og når vi indsætter Prebens og Berits højder finder vi

$$49.054 + 0.418 \times 172 + 0.329 \times 175 = 178.525$$

En mere nøjagtig udregning fås fra `estimate`-sætningen:

```
estimate "son with father=172, mother 175" intercept 1 fht 172 mht 175;
```

Parameter	Estimate	Standard Error	t Value
son with father=172, mother 175	178.409563	0.65359247	272.97

Parameter	Pr > t	95% Confidence Limits
son with father=172, mother 175	<.0001	177.125294 179.693833

Desuden ser vi, at sønner af Preben og Berit nu estimeres til at være gennemsnitlig 178.41 cm høje, med CI=(177.13, 179.69), og med 95% prediktionsinterval på

$$178.41 \pm 2 \times 5.83 = (166.75, 190.07)$$

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?

Det er faktisk rigtigt nemt, for det svarer bare til at teste, om differensen mellem de to regressionskoefficienter er 0, altså

$$H_0 : \beta_F - \beta_M = 0$$

Vi benytter derfor en `estimate`-sætning til estimation af denne differens, med konfidensinterval:

```
proc glm plots=all data=a1; where gender="male";  
class gender;  
model height = fht mht / solution clparm;  
estimate "differens for fht og mht" fht 1 mht -1;  
run;
```

hvilket giver os det yderligere output (resten er vist ovenfor i forbindelse med spørgsmål 5a):

Parameter	Estimate	Standard Error	t Value	Pr > t
differens for fht og mht	0.08878280	0.06703931	1.32	0.1860

Parameter	95% Confidence Limits	
differens for fht og mht	-0.04294538	0.22051098

Vi noterer os, at højde-øgningen svarende til 1 far-cm er 0.0888cm 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:

```
ods graphics / imagename="spm6-3e-mixed";
proc mixed plots=all data=a1;
class family gender;
model height = gender fht / s cl;
random intercept / subject=family;
run;
ods graphics off;
```

hvorved vi får outputtet:

The Mixed Procedure

Model Information

Data Set	WORK.A1
Dependent Variable	height
Covariance Structure	Variance Components
Subject Effect	family
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
family	205	1 10 100 101 102 103 104 105 90 91 92 93 94 95 96 97 98 99
gender	2	female male

Dimensions

Covariance Parameters	2
Columns in X	4
Columns in Z per Subject	1

Subjects	205
Max Obs per Subject	15

Number of Observations

Number of Observations Read	934
Number of Observations Used	934
Number of Observations Not Used	0

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	family	9.0967
Residual		24.5381

Solution for Fixed Effects

Effect	gender	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept		106.06	7.5128	203	14.12	<.0001	0.05
gender	female	-13.2183	0.3467	728	-38.13	<.0001	0.05
gender	male	0	.	.	.	.	.
fht		0.3980	0.04268	728	9.32	<.0001	0.05

Effect	gender	Lower	Upper
Intercept		91.2492	120.88
gender	female	-13.8989	-12.5377
gender	male	.	.
fht		0.3142	0.4818

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, er:

```
ods graphics / imagename="spm6-5c-mixed";
proc mixed plots=all data=a1; where gender="male";
class family gender;
model height = fht mht / s cl;
random intercept / subject=family;
estimate "differens for fht og mht" fht 1 mht -1 / cl;
run;
ods graphics off;
```

som giver outputtet:

The Mixed Procedure

Model Information

Data Set	WORK.A1
Dependent Variable	height
Covariance Structure	Variance Components
Subject Effect	family
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Class Level Information

Class	Levels	Values
family	179	1 100 101 102 103 104 105 106 84 85 86 87 88 89 90 91 92 93 95 96 97 99
gender	1	male

Dimensions

Covariance Parameters	2
Columns in X	3
Columns in Z per Subject	1
Subjects	179
Max Obs per Subject	10

Number of Observations

Number of Observations Read	481
Number of Observations Used	481
Number of Observations Not Used	0

Covariance Parameter Estimates

Cov Parm	Subject	Estimate
Intercept	family	6.8942
Residual		27.2437

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t	Alpha
Intercept	52.6424	12.2623	176	4.29	<.0001	0.05
fht	0.4144	0.05278	302	7.85	<.0001	0.05
mht	0.3106	0.05429	302	5.72	<.0001	0.05

Effect	Lower	Upper
Intercept	28.4423	76.8426
fht	0.3105	0.5182
mht	0.2037	0.4174

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t
differens for fht og mht	0.1038	0.07881	302	1.32	0.1887

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:

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