

Vejledende besvarelse hjemmeopgave E21

Bemærk at denne besvarelse er baseret på alle 934 observationer.

Data indlæses fra URL med

```
g <- read.csv("http://publicifsv.sund.ku.dk/~sr/BasicStatistics/datasets/hjemmeopgave/dataE21.csv")
```

Vi laver `gender` til en faktor med labels, så vi ikke behøver at huske koderne:

```
g$gender <- factor(g$gender, labels=c("male", "female"))
```

Inden vi går i gang med spørgsmålene, checker vi lige at der ikke er nogle skøre værdier for vores variable, f.eks. ved at udregne summary statistics for hvert køn:

```
> table( g$gender )

male female
 481    453
> tapply( g$father, g$gender, summary )
$male
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 62.0   68.0   69.0   69.1   70.5   78.5

$female
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 62.0   68.0   69.5   69.3   71.0   78.5
> tapply( g$mother, g$gender, summary )
$male
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 58.0   63.0   64.0   64.0   66.0   70.5

$female
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 58.0   63.0   64.0   64.2   65.5   70.5
> tapply( g$childNum, g$gender, summary )
$male
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1.00   1.00   2.00   2.26   3.00   10.00

$female
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 1      3      5      5      6     15
> tapply( g$childHeight, g$gender, summary )
$male
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 60.0   67.5   69.2   69.2   71.0   79.0
```

```
$female
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  56.0   62.5   64.0   64.1   65.5   70.5
```

Vi ser her, at materialet består af 453 piger og 481 drenge, og at summary statistics for de 3 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):

```
g$fht <- g$father*2.54;
g$mht <- g$mother*2.54;
g$height <- g$childHeight*2.54
```

2. Se på (dele af) originalmaterialet og vurder hvad variablen child-Num mon betyder.

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

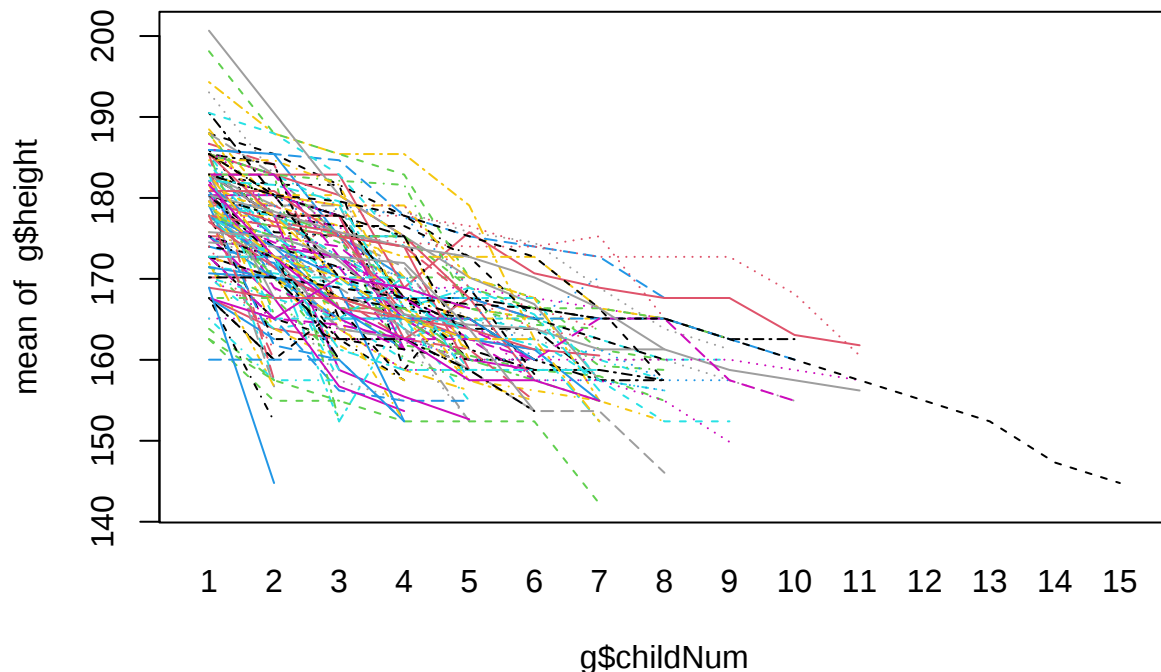
FAMILY HEIGHTS. from R.F.F.
(could be inches in every entry in the Table.)

	Father	Mother	Sons in order of height	Daughters in order of height
1	15.5	7.0	13.2	9.2, 9.0, 9.0
2	15.5	6.5	13.5, 13.5	5.5, 5.5
3	15.0	about 4.0	11.0	8.0
4	15.0	4.0	10.5, 7.5	7.0, 4.5, 3.0
5	15.0	1.5	12.0, 9.0, 7.0	6.5, 2.5, 2.5
6	14.0	5.0		4.5
7	14.0	2.0	16.5, 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, 7.0, 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	2.0, 7.0	
15	13.0	6.5	11.0, 10.5	6.7
16	13.0	about 5.0	12.0, 10.5, 10.2, 10.2, 9.2	2.7, 6.5, 4.5, 3.5
17	13.0	4.5	14.0, 13.0, 11.5, 2.5	6.5, 2.3
18	13.0	4.0		6.0, 4.5, 4.0
19	13.2	3.0		2.7
20	12.7	4.0	13.2, 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	13.0, 11.0	7.0
23	12.0	5.0	14.2, 10.5, 9.5	6.0, 5.5, 5.0, 2.0
24	12.0	5.5		5.5

Galton Papers/UCL

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 hvor vi plotter højden mod `childnum` med en separat linje i forskellige farver/linjetyper for hver af de 205 familier:

```
interaction.plot( g$childNum, g$family, g$height, col=1:205,
                  legend=F)
```



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 ved at sortere højderne på de enkelte afkom inden for hver familie.

(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).

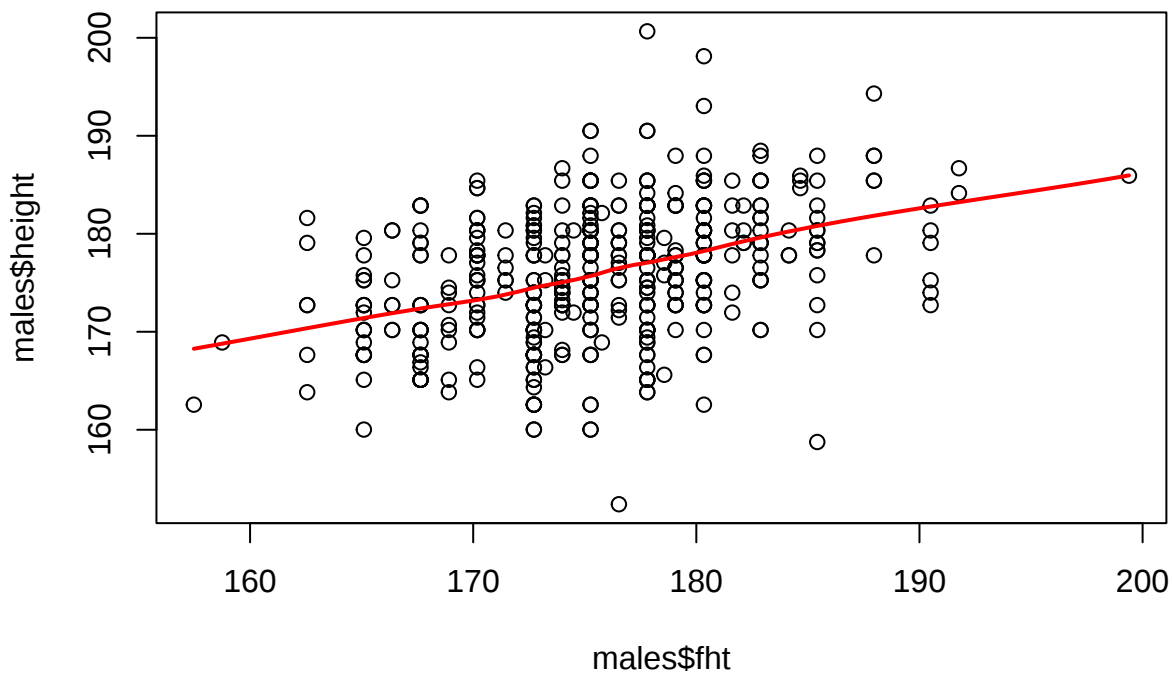
Vi har i disse data op til 15 børn fra de enkelte familier (se figuren ovenfor), og vi må formode, at der er en vis (positiv) korrelation mellem disse børns højde. Noget af denne forklares givetvis ud fra moderens og faderens højde, men der kan nemt være uforklaret korrelation f.eks. pga. samme kostvaner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

```
males <- subset(g,gender=="male")
scatter.smooth( males$height ~ males$fht, lpars=list(col="red",lwd=2))
```

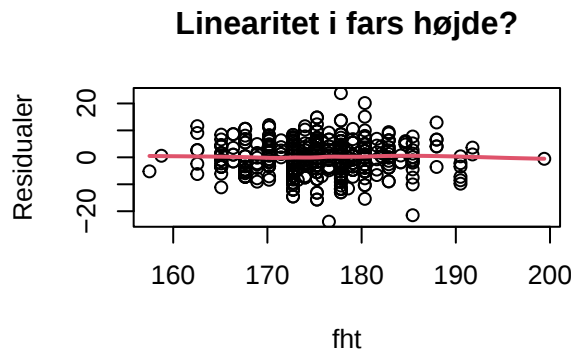
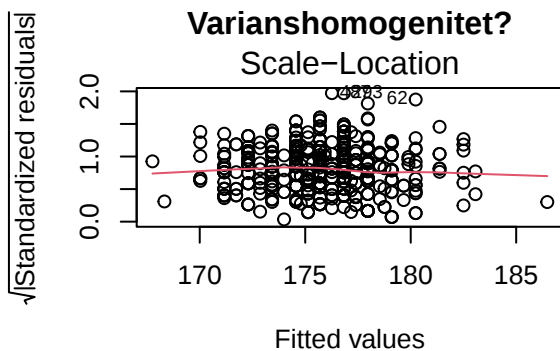
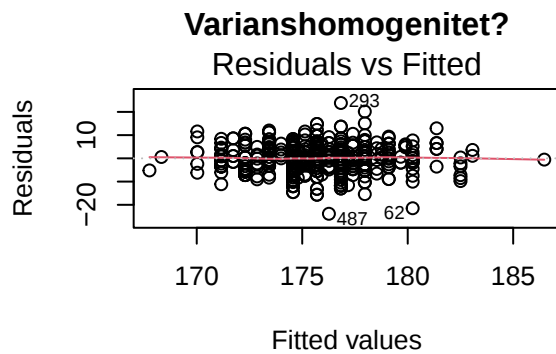
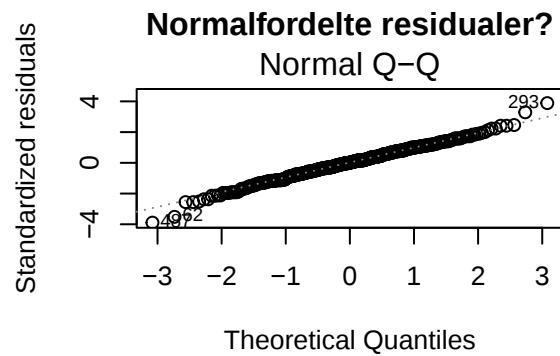


Da der her ses en rimeligt lineær sammenhæng mellem fædres og sønners højde, går vi videre med en simpel lineær regression.

```
lm1 <- lm( height ~ fht, data=males )
```

Inden vi rapporterer tallene fra denne analyse, kaster vi lige et blik på modelkontrollen:

```
par(mfrow=c(2,2))
plot( lm1, which=2, main="Normalfordelte residualer?" )
plot( lm1, which=1, main="Varianshomogenitet?" )
plot( lm1, which=3, main="Varianshomogenitet?" )
scatter.smooth( resid(lm1) ~ males$fht,lpars=list(col=2,lwd=2),
               main="Linearitet i fars højde?", xlab="fht", ylab="Residualer")
```



```
par(mfrow=c(1,1))
```

Antagelserne om normalfordeling, varianshomogenitet og linearitet ser ud til at være opfyldt.

Nu til resulaterne:

```
summary(lm1)
```

Call:

```
lm(formula = height ~ fht, data = males)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.866	-3.841	0.105	4.119	23.827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	97.4410	8.4033	11.60	<2e-16 ***
fht	0.4465	0.0478	9.34	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.14 on 479 degrees of freedom

Multiple R-squared: 0.154, Adjusted R-squared: 0.152

F-statistic: 87.2 on 1 and 479 DF, p-value: <2e-16

```
confint(lm1)
```

	2.5 %	97.5 %
(Intercept)	80.9291	113.9528
fht	0.3525	0.5405

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 0.5 cm.

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

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

(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.

```
pred <- predict(lm1,
newdata = data.frame(fht= c(0,5),
gender = c('male','male')),
interval = 'confidence')
```

(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:

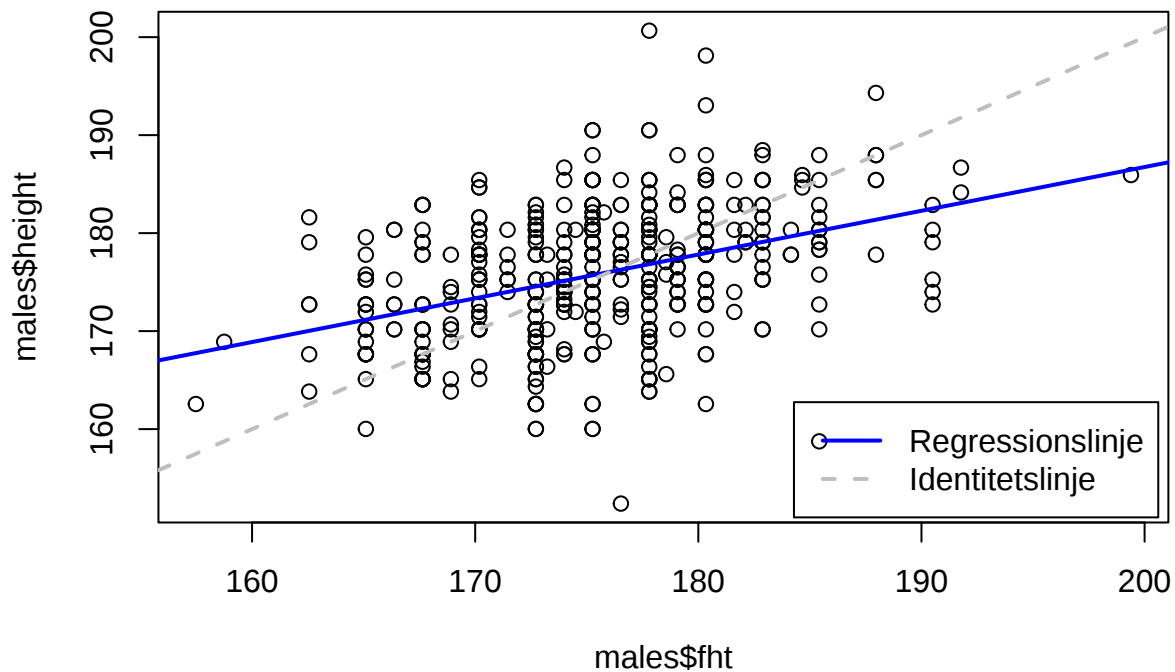
$$x = 97.44 + 0.4465 \times x$$

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 identitetslinje:

```
plot( males$height ~ males$fht )
abline(lm1, col="blue",lwd=2)
abline(a=0,b=1, col="gray",lwd=2, lty=2)
legend("bottomright",c("Regressionslinje","Identitetslinje"),inset=.01,
      lty=1:2,col=c("blue","gray"),lwd=2)
```



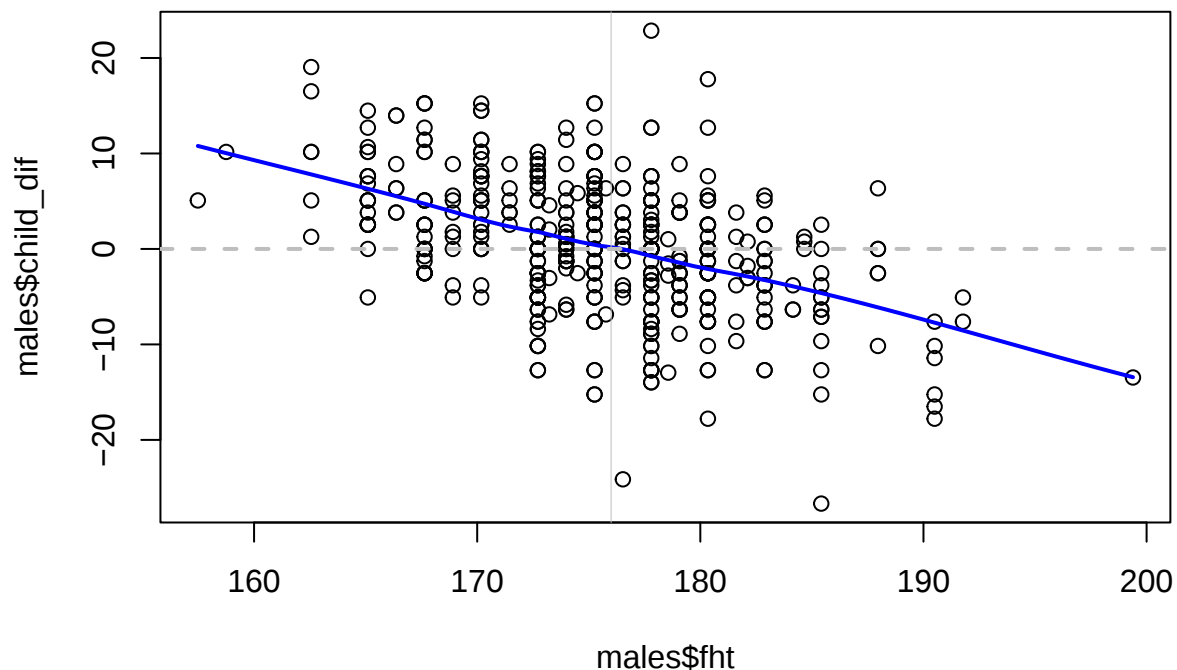
hvor vi ser, at denne identitetslinie netop skærer regressionslinjen i 176 (hvis man kigger godt efter).

Alternativt kan vi definere forskellen mellem fars og søns højde ved (på passende sted) at tilføje en variabel som angiver forskellen mellem søn og fars højde:

```
males$child_dif <- males$height-males$fht
```

og dernæst f.eks tegne denne differens incl en udglatning op mod faderens højde, samt en vandret linie i 0 og en lodret linje i 176:

```
scatter.smooth( males$child_dif ~ males$fht, lpars=list(col="blue",lwd=2) )
abline(a=0,b=0, col="gray",lwd=2, lty=2)
abline(v=176,col="gray",lwd=.5)
```



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:

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

For at svare på spørgsmålet, er det 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 i summary ovenfor under Residual standard error at være 6.14, så vi udregner:

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

Vi kan også bare benytte `predict()` og samtidigt beregne et prædiktionsinterval:

```
new <- data.frame(fht=172)
predict( lm1, newdata = new, interval="prediction")
```

```
      fit    lwr    upr
1 174.2 162.2 186.3
```

R benytter det “helt rigtige 2-tal” og har heller ikke afrunding af tallene, så derfor får vi den mest præcise beregning på den måde.

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?

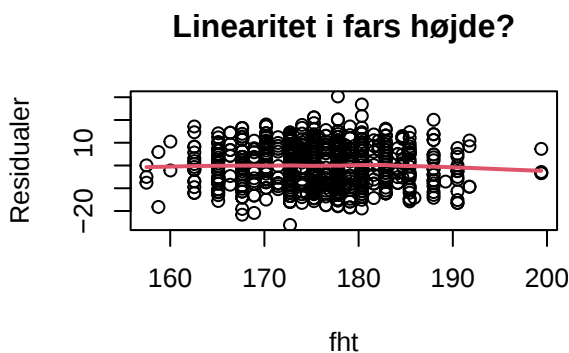
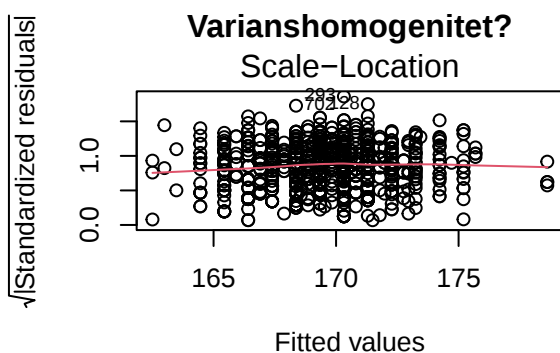
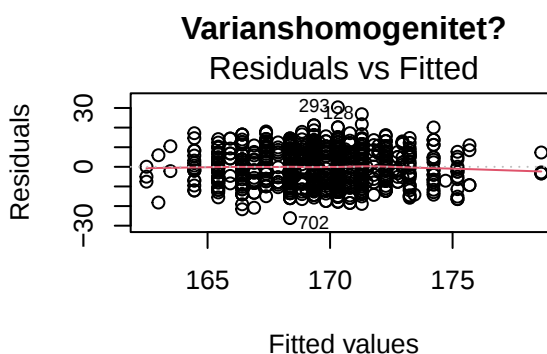
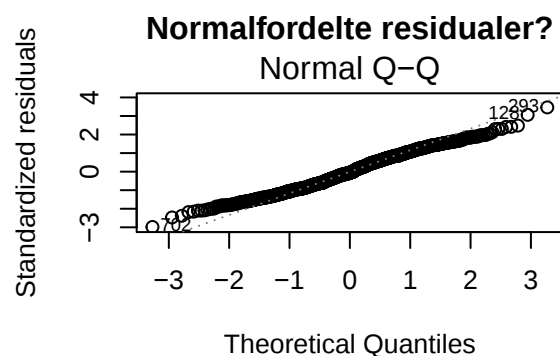
Husk at benytte en model, der tillader at angive dette estimat med CI.

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:

```
female <- subset(g, gender="female")
lm2 <- lm( height ~ fht, data=female )
```

Modelkontrol - igen nydeligt:

```
par(mfrow=c(2,2))
plot( lm2, which=2, main="Normalfordelte residualer?" )
plot( lm2, which=1, main="Varianshomogenitet?" )
plot( lm2, which=3, main="Varianshomogenitet?" )
scatter.smooth( resid(lm2) ~ female$fht, lpars=list(col=2,lwd=2),
               main="Linearitet i fars højde?", xlab="fht",
               ylab="Residualer")
```



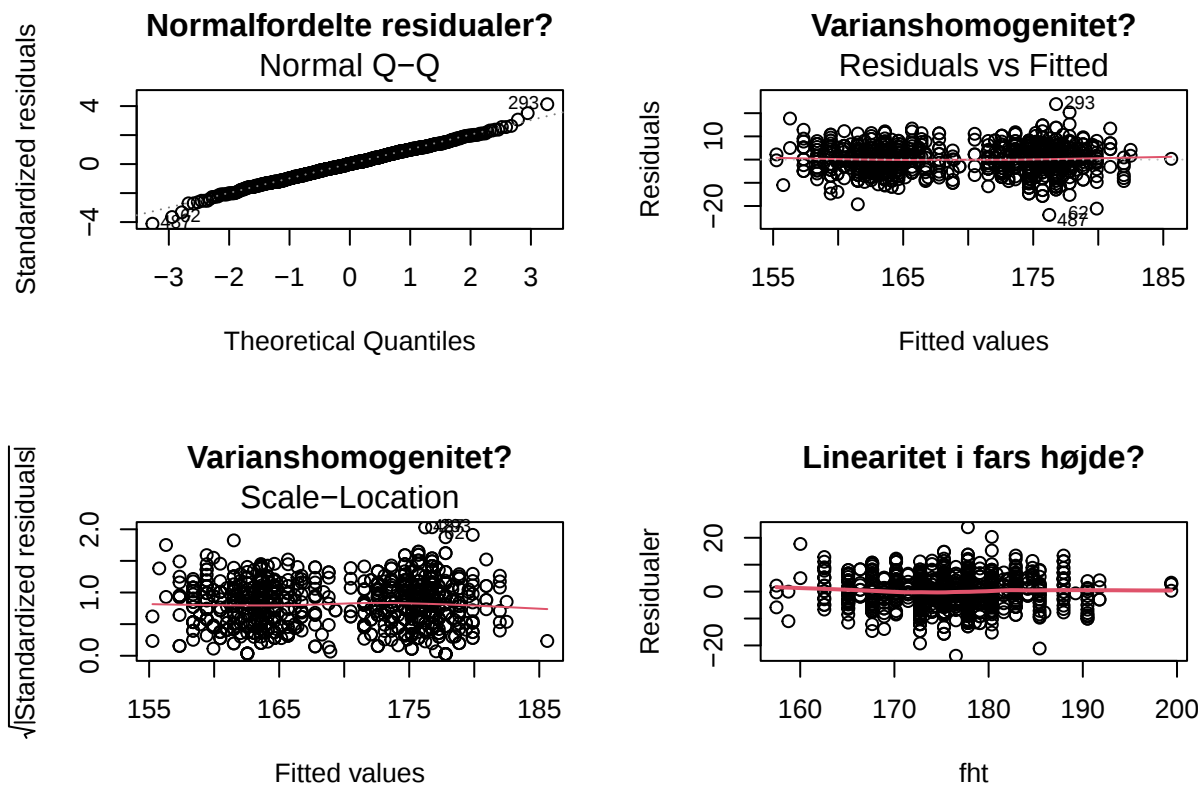
```
par(mfrow=c(1,1))
```

Vi bemærker igen, at faders højde er en stærkt signifikant prædiktør, 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. Regressionsskoefficienten 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:

```
lm3 <- lm( height ~ gender + fht, data=g)
```

En hurtig modelkontrol

```
par(mfrow=c(2,2))
plot( lm3, which=2, main="Normalfordelte residualer?" )
plot( lm3, which=1, main="Varianshomogenitet?" )
plot( lm3, which=3, main="Varianshomogenitet?" )
scatter.smooth( resid(lm3) ~ g$fht,lpars=list(col=2,lwd=2),
               main="Linearitet i fars højde?", xlab="fht",
               ylab="Residualer")
```



```
par(mfrow=c(1,1))
```

inden vi tager et kig på estimerterne:

```
summary(lm3)
```

Call:

```
lm(formula = height ~ gender + fht, data = g)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.832	-3.805	0.007	4.045	23.906

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  103.7833     5.3068   19.6   <2e-16 ***
genderfemale -13.1583     0.3797  -34.7   <2e-16 ***
fht           0.4104     0.0302   13.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 5.8 on 931 degrees of freedom
Multiple R-squared:  0.594, Adjusted R-squared:  0.593
F-statistic: 682 on 2 and 931 DF,  p-value: <2e-16

```

```
confint(lm3)
```

```

              2.5 %    97.5 %
(Intercept)  93.3687 114.1979
genderfemale -13.9035 -12.4132
fht           0.3512   0.4696

```

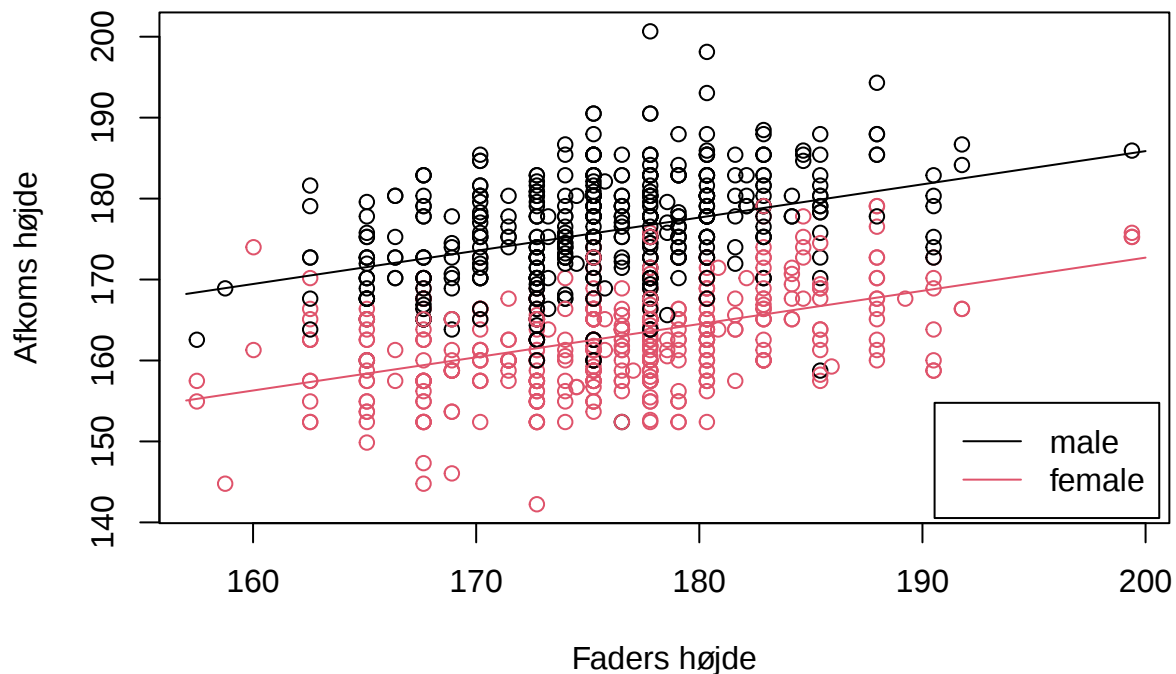
Forudsat, at forskellen mellem højden på bror og søster er den samme uafhængigt af faderens højde (svarende til de parallelle regressionslinier i figuren ovenfor), finder vi altså, at en søster forventeligt vil være 13.16 cm lavere end sin bror, CI=(12.41, 13.90) cm.

En illustration af modellen:

```

plot( g$height ~ g$fht, col=g$gender, xlab="Faders højde", ylab="Afkoms højde")
# tapply(g$fht,g$gender,range)
new1 <- data.frame( fht=157:200, gender="male")
p1 <- predict( lm3, newdata=new1 )
new2 <- data.frame( fht=157:200, gender="female")
p2 <- predict( lm3, newdata=new2 )
lines( new1$fht, p1, col=1)
lines( new2$fht, p2, col=2)
legend("bottomright",c("male","female"),col=1:2,lty=1,inset=.01)

```



(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:

```
lm3 <- lm( height ~ gender + fht + gender*fht, data=g)
summary(lm3)
```

Call:

```
lm(formula = height ~ gender + fht + gender * fht, data = g)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.866	-3.822	-0.012	4.042	23.827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	97.4410	7.9378	12.28	<2e-16 ***
genderfemale	-1.6957	10.6764	-0.16	0.87
fht	0.4465	0.0452	9.88	<2e-16 ***
genderfemale:fht	-0.0652	0.0607	-1.07	0.28

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.8 on 930 degrees of freedom

Multiple R-squared: 0.595, Adjusted R-squared: 0.593

F-statistic: 455 on 3 and 930 DF, p-value: <2e-16

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, $CI=(12.05, 13.78)$:

```
g$fht172 <- g$fht-172
lm4a <- lm( height ~ gender + fht172 + gender*fht172, data=g)
summary(lm4a)
```

Call:

```
lm(formula = height ~ gender + fht172 + gender * fht172, data = g)
```

Residuals:

Min	1Q	Median	3Q	Max
-23.866	-3.822	-0.012	4.042	23.827

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	174.2428	0.3106	561.05	<2e-16 ***

```

genderfemale      -12.9139      0.4426  -29.18   <2e-16 ***
fht172            0.4465       0.0452   9.88   <2e-16 ***
genderfemale:fht172 -0.0652      0.0607  -1.07    0.28
---

```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.8 on 930 degrees of freedom
Multiple R-squared: 0.595, Adjusted R-squared: 0.593
F-statistic: 455 on 3 and 930 DF, p-value: <2e-16

```
confint(lm4a)
```

```

                2.5 %    97.5 %
(Intercept)    173.6334 174.85233
genderfemale   -13.7825 -12.04521
fht172          0.3579  0.53518
genderfemale:fht172 -0.1844 0.05392

```

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?

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) - med mindre en af de 4 observationer vi har slettet, er det første barn i en familie. Vi ønsker derfor at snuppe den første linje for hver familie. Det kan vi klare vha kommandoen `duplicated()` som returnerer værdien `FALSE` første gang den støder på en værdi i en vektor, `TRUE` ellers. Benytter vi den på vores familie-id-variabel `family` vil den således være `FALSE` for første linje i hver familie. Vi kan bytte rundt på `TRUE` og `FALSE` med `!` og dermed plukke den første linje for hver familie med:

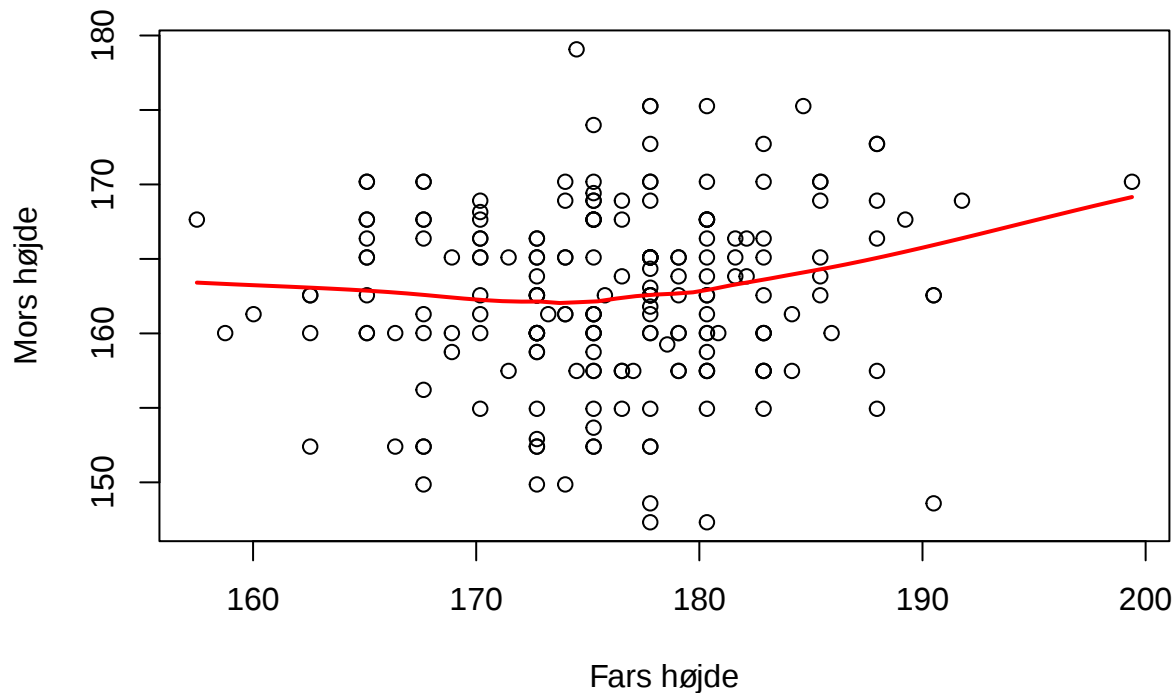
```
g1 <- subset(g,!duplicated(family))
```

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

```

scatter.smooth( g1$mht ~ g1$fht, lpars=list(col="red",lwd=2),
               xlab="Fars højde", ylab="Mors højde")

```



Det ser ikke ud til at der er 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.

```
cor.test( g1$mht, g1$fht )
```

Pearson's product-moment correlation

```
data: g1$mht and g1$fht
t = 1.4, df = 203, p-value = 0.2
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.04068  0.23087
sample estimates:
cor
0.0969
```

Vi finder altså en korrelationskoefficient (Pearson) på 0.097 og kan konstatere, at den ikke er statistisk signifikant forskellig fra 0 ($P=0.20$). Ud over det, kan vi ikke rigtigt bruge det til noget.

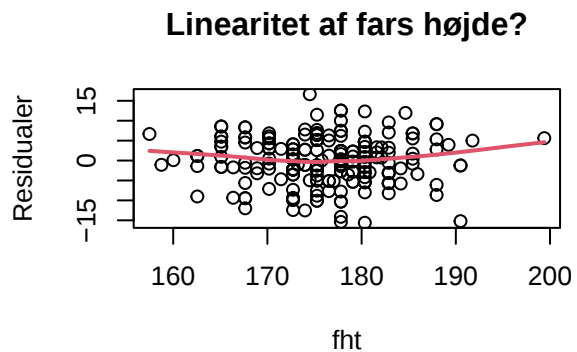
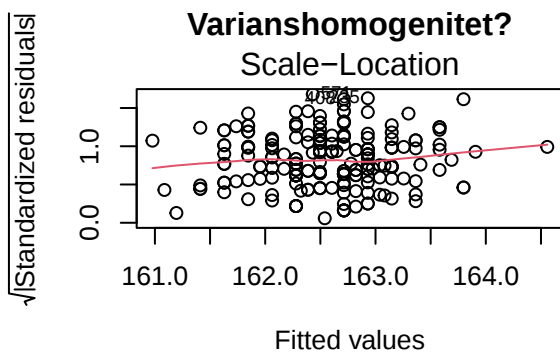
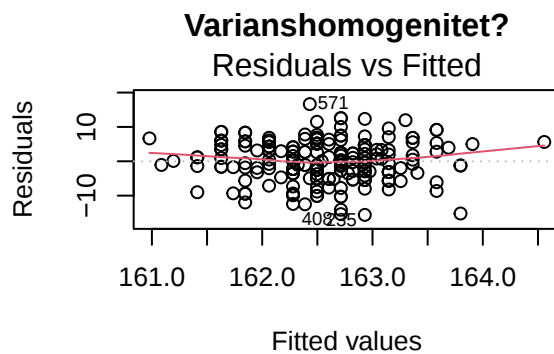
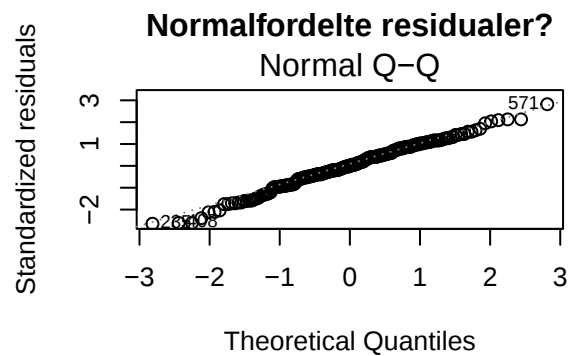
(b) Hvor høj vil Prebens partner typisk være?

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:

```
lm5 <- lm( mht ~ fht, data=g1 )
```

Modelkontrol inden vi kigger på estimerne:

```
par(mfrow=c(2,2))
plot( lm5, which=2, main="Normalfordelte residualer?" )
plot( lm5, which=1, main="Varianshomogenitet?" )
plot( lm5, which=3, main="Varianshomogenitet?" )
scatter.smooth( resid(lm5) ~ g1$fht,lpars=list(col=2,lwd=2),
               main="Linearitet af fars højde?", xlab="fht",
               ylab="Residualer")
```



```
par(mfrow=c(1,1))
```

Ingen grund til bekymring her og vi kan derfor tage et kig på estimerne:

```
summary(lm5)
```

Call:

```
lm(formula = mht ~ fht, data = g1)
```

Residuals:

Min	1Q	Median	3Q	Max
-15.61	-3.20	0.02	4.31	16.64

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	147.5293	10.8473	13.60	<2e-16 ***
fht	0.0854	0.0616	1.39	0.17

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.91 on 203 degrees of freedom
Multiple R-squared: 0.00939, Adjusted R-squared: 0.00451
F-statistic: 1.92 on 1 and 203 DF, p-value: 0.167

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

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

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

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

Vi vil jo også gerne have et CI på dette estimat for den forventede højde og bestemmer derfor prædiktionen vha `predict()`:

```
nyedata <- data.frame(fht=172)
predict( lm5, nyedata, interval="confidence")
```

	fit	lwr	upr
1	162.2	161.3	163.2

som ses at give 162.2cm, med 95% konfidensinterval $CI=(161.27, 163.17\text{cm})$.

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

Her skal vi igen bruge et prædiktionsinterval. Dette gøres vha

```
nyedata <- data.frame( fht=172 )
predict( lm5, nyedata, interval="prediction" )
```

	fit	lwr	upr
1	162.2	150.5	173.9

Vi finder dermed et 95% prædiktionsinterval til at være (150.3; 173.7) og 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.

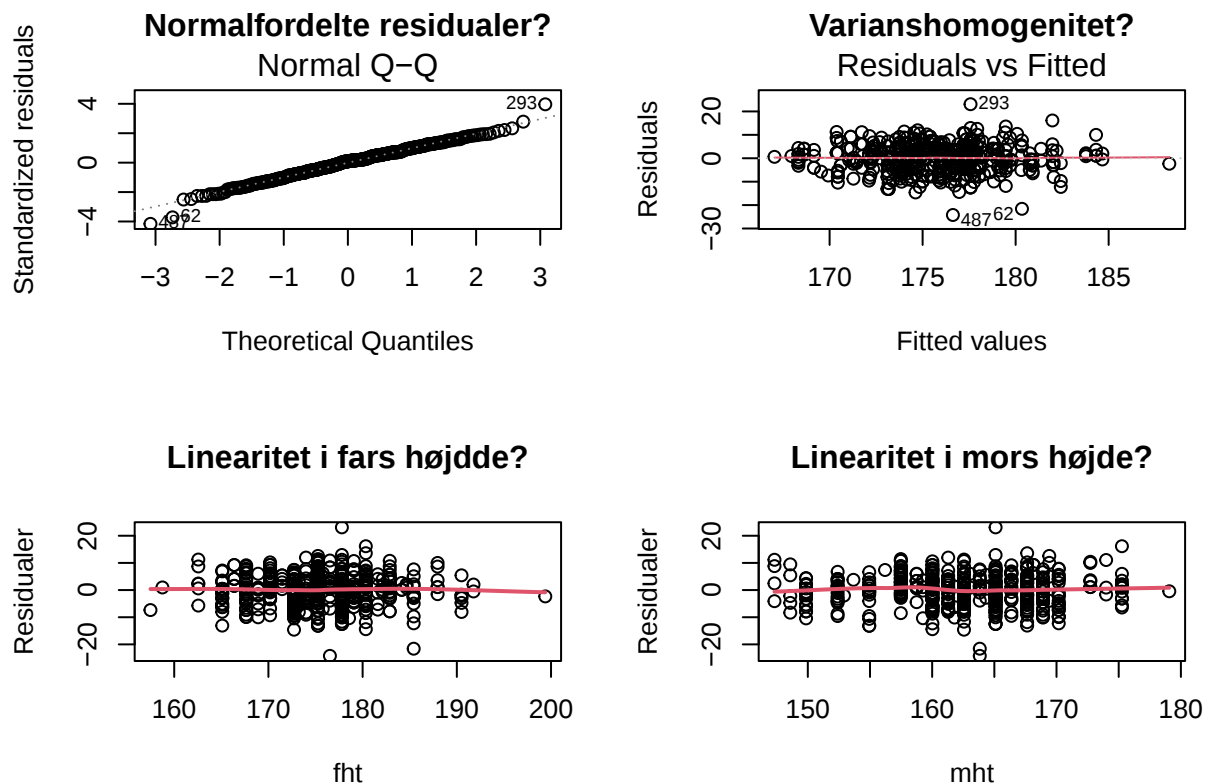
Nu ved vi mere, nemlig at Prebens partner er ret høj, nemlig 175 cm. Det er en del højere end forventet, jf spørgsmål 4b ovenfor.

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:

```
lm6 <- lm( height ~ fht + mht, data=males )
```

Modelkontrol inden vi kigger på estimerne:

```
par(mfrow=c(2,2))
plot( lm6, which=2, main="Normalfordelte residualer?" )
plot( lm6, which=1, main="Varianshomogenitet?" )
scatter.smooth( resid(lm6) ~ males$fht,lpars=list(col=2,lwd=2),
               main="Linearitet i fars højde?", xlab="fht",
               , ylab="Residualer")
scatter.smooth( resid(lm6) ~ males$mht,lpars=list(col=2,lwd=2),
               main="Linearitet i mors højde?", xlab="mht",
               , ylab="Residualer")
```



```
par(mfrow=c(1,1))
```

Selve estimerne med konfidensintervaller er:

```
summary( lm6 )
```

Call:

```
lm(formula = height ~ fht + mht, data = males)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-24.229	-3.946	0.562	3.825	23.083

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.0545	10.4014	4.72	3.2e-06 ***
fht	0.4176	0.0456	9.15	< 2e-16 ***
mht	0.3288	0.0453	7.26	1.6e-12 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.83 on 478 degrees of freedom

Multiple R-squared: 0.238, Adjusted R-squared: 0.235

F-statistic: 74.6 on 2 and 478 DF, p-value: <2e-16

```
confint( lm6 )
```

	2.5 %	97.5 %
(Intercept)	28.6165	69.4926
fht	0.3279	0.5072
mht	0.2398	0.4178

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 0.42cm (95%CI 0.33-0.51cm) ekstra for sønnen, medens 1 ekstra mor-cm kun giver forventeligt 0.33cm mere (95%CI 0.24-0.42). 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 - og tilmed med et CI - kan vi få vha `predict()`:

```
nyedata <- data.frame( fht=172, mht=175)
predict( lm6, nyedata, interval="confidence")
```

	fit	lwr	upr
1	178.4	177.1	179.7

Vi estimerer altså gennemsnitshøjden til sønner af fædre på 172cm, mødre på 175cm, til at være 187.4 (95%CI 177.1-179.7).

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

Prædiktionsintervallet for en søn af dette par finder vi nu nemt med

```
predict( lm6, nyedata, interval="prediction")
```

```
      fit    lwr    upr  
1 178.4 166.9 189.9
```

dvs fra 166.9 til 189.9cm.

Til sammenligning fik vi (i spørgsmål 3e) prediktionsintervallet (161.96, 186.52) cm uden vores viden om Berits højde. 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? Kvantificér 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 $b_{far} = 0.418 > 0.329 = b_{mor}$, men hvordan kan vi teste, om denne forskel på regressionskoefficienterne er signifikant?

Det svarer bare til at teste, om differensen mellem de to regressionskoefficienter er 0, altså

$$H_0 : b_{far} - b_{mor} = 0$$

Den estimerede forskel mellem de to hældninger er altså $0.418 - 0.329 = 0.089$ og det er denne, som vi ønsker at teste lig 0.

Løsning 1:

En nem løsning er at definere en ny forklarende variabel, som er summen af mors og fars højde. Vi kan omskrive modellen indholdende **fht** og **mht** som vi benyttede ovenfor:

$$a + b_{far} \times \text{fht} + b_{mor} \times \text{mht} = a + (b_{far} - b_{mor}) \times \text{fht} + b_{mor}(\text{fht} + \text{mht})$$

og ser at hvis vi laver en regression med **fht** og **fht+mht** bliver koefficienten til **fht** netop forskellen mellem fars og mors hældning. Dette er blot en omskrivning af modellen (prædiktioner vil være præcis de samme), dvs det er i bund og grund en samme model (på statistiksprog siger vi at det er en *omparameterisering* af modellen).

Vi laver nu analysen:

```
males$mfht <- males$mht + males$fht  
lm7 <- lm( height ~ fht + mfht, data=males )  
summary(lm7)
```

Call:

```
lm(formula = height ~ fht + mfht, data = males)
```

Residuals:

Min	1Q	Median	3Q	Max
-24.229	-3.946	0.562	3.825	23.083

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	49.0545	10.4014	4.72	3.2e-06 ***

```
fht      0.0888    0.0670    1.32    0.19
mfht     0.3288    0.0453    7.26  1.6e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 5.83 on 478 degrees of freedom
Multiple R-squared:  0.238, Adjusted R-squared:  0.235
F-statistic: 74.6 on 2 and 478 DF,  p-value: <2e-16
```

```
confint(lm7)
```

```
                2.5 %  97.5 %
(Intercept) 28.61648 69.4926
fht         -0.04295  0.2205
mfht         0.23976  0.4178
```

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. Vi kan altså ikke vise, at der skulle være forskel på effekten af mors højde og fars højde på afkommets højde.

Løsning 2

Vi kan også beregne differensen af de to hældninger ved at benytte `glht()` (se e-bogens kapitel om kvantificering, hvor der også er en video). Vi har 3 koefficienter i vores model

```
coef(lm6)
```

```
(Intercept)      fht      mht
      49.0545    0.4176    0.3288
```

og ønsker at bestemme koefficient nr 2 minus koefficient nr 3. Det svarer til at tage 1 gange koefficient nr 2 og lægge -1 gange koefficient nr 3 til:

```
library(multcomp)
glht6 <- glht( lm6, linfct=rbind(c(0,1,-1)))
summary(glht6)
```

Simultaneous Tests for General Linear Hypotheses

```
Fit: lm(formula = height ~ fht + mht, data = males)
```

Linear Hypotheses:

```
      Estimate Std. Error t value Pr(>|t|)
1 == 0    0.0888    0.0670    1.32    0.19
(Adjusted p values reported -- single-step method)
```

```
confint(glht6)
```

Simultaneous Confidence Intervals

```
Fit: lm(formula = height ~ fht + mht, data = males)
```

```
Quantile = 1.965
95% family-wise confidence level
```

Linear Hypotheses:

	Estimate	lwr	upr
1 == 0	0.0888	-0.0429	0.2205

Vi får præcis samme tal som i løsning 1.

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 estimater 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:

```
library(nlme)
lme1 <- lme(height ~ gender+fht, data = g, random =~ 1 | family,
            na.action=na.exclude)
summary(lme1)
```

Linear mixed-effects model fit by REML

Data:	g	
AIC	BIC	logLik
5795	5819	-2892

Random effects:

Formula:	~1 family	
	(Intercept)	Residual
StdDev:	3.027	4.964

Fixed effects: height ~ gender + fht

	Value	Std.Error	DF	t-value	p-value
(Intercept)	106.0	7.539	721	14.07	0
genderfemale	-13.2	0.349	721	-37.81	0
fht	0.4	0.043	202	9.29	0

Correlation:

	(Intr)	gndrfm
genderfemale	0.011	
fht	-0.999	-0.033

Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-----	----	-----	----	-----

```
-4.076485 -0.579351 0.004282 0.620982 3.631061
```

```
Number of Observations: 926
```

```
Number of Groups: 204
```

```
intervals(lme1)
```

```
Approximate 95% confidence intervals
```

```
Fixed effects:
              lower    est.    upper
(Intercept)  91.2415 106.043 120.8437
genderfemale -13.8850 -13.200 -12.5143
fht           0.3136  0.398  0.4825
attr(,"label")
[1] "Fixed effects:"
```

```
Random Effects:
Level: family
              lower    est.    upper
sd((Intercept)) 2.564 3.027 3.575
```

```
Within-group standard error:
lower    est.    upper
4.718 4.964 5.224
```

Vi bemærker her, at en søster estimeres til at være 13.20 cm lavere end sin bror, CI=(12.51, 13.89). 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

$$\frac{3.027^2}{3.027^2 + 4.964^2} = 0.27$$

Vi skal nu se på spørgsmål 5c:

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.

Den tilsvarende analyse, med inddragelse af korrelation, er:

```
lme2 <- lme(height ~ fht+mfht, data = males, random =~ 1 | family,
             na.action=na.exclude)
summary(lme2)
```

```
Linear mixed-effects model fit by REML
```

```
Data: males
AIC BIC logLik
3029 3050 -1510
```

```
Random effects:
```

```
Formula: ~1 | family
```

```

      (Intercept) Residual
StdDev:      2.647      5.226

Fixed effects: height ~ fht + mfht
              Value Std.Error DF t-value p-value
(Intercept)  52.68      12.311 298   4.279  0.0000
fht           0.10       0.079 175   1.318  0.1891
mfht          0.31       0.055 175   5.684  0.0000
Correlation:
  (Intr) fht
fht  -0.013
mfht -0.657 -0.745

Standardized Within-Group Residuals:
      Min       Q1       Med       Q3      Max
-3.948e+00 -5.849e-01 -6.784e-05  5.961e-01  3.340e+00

Number of Observations: 476
Number of Groups: 178

```

```
intervals(lme2)
```

Approximate 95% confidence intervals

```

Fixed effects:
      lower    est.    upper
(Intercept) 28.4484 52.6767 76.9050
fht          -0.0519  0.1044  0.2608
mfht          0.2024  0.3101  0.4178
attr(,"label")
[1] "Fixed effects:"

```

```

Random Effects:
Level: family
      lower    est.    upper
sd((Intercept)) 1.991 2.647 3.519

```

```

Within-group standard error:
lower    est.    upper
4.833 5.226 5.651

```

Her finder vi en forskel på 0.1044 på de to regressionsestimater, mod 0.0888 uden hensyntagen til korrelationen. Usikkerheden på estimatet er øget lidt her (fra 0.067 før til 0.079 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 / et bredere CI (her -0.052 til 0.261).

Igen er korrelationen dog ganske beskedn:

$$\frac{2.647^2}{2.647^2 + 5.226^2} = 0.20$$