

Opgavebesvarelse, brain weight

(Matthews & Farewell: Using and Understanding Medical Statistics, 2nd. ed.)

For 20 nyfødte mus er der i tabellen nedenfor anført oplysning om

- kuldstørrelsen (fra 3 til 12 mus i et kuld)
- kropsvægt i g
- hjernevægt i g

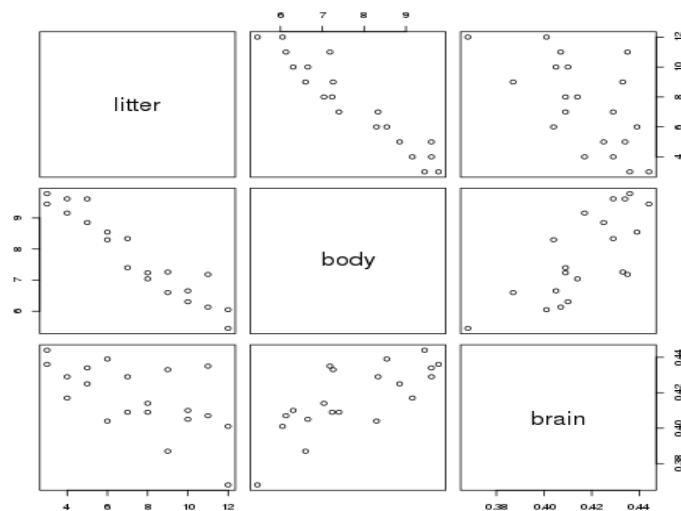
Data er indlagt på filen `brain.txt` med tre kolonner, svarende til de tre variable: `litter`, `body` og `brain`. Disse variabelnavne er anført i øverste linie af datafilen.

Vi ønsker at udtale os om hjernevægtens afhængighed af kuldstørrelsen.

Først indlæser vi ved at skrive

```
mus <-  
read.table("http://publicifsv.sund.ku.dk/~lts/basal/data/brain.txt",  
header=T)
```

Vi har nu 20 observationer, hver med 3 variable, og vi kan få et plot af variablene parvis ved simpelthen at skrive `plot(mus)`:

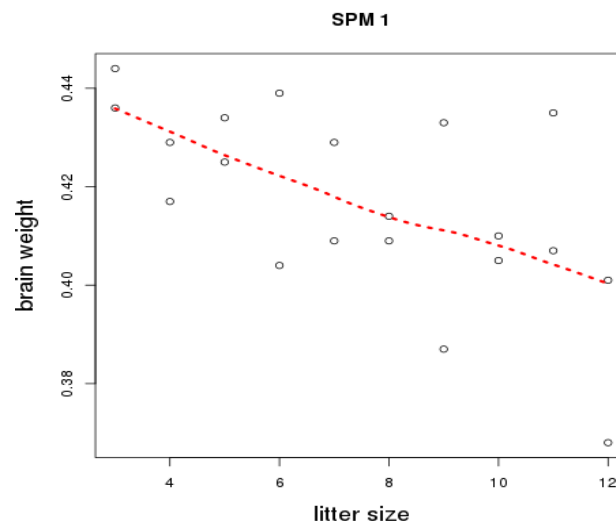


1. Vurder løseligt ud fra en tegning, om forudsætningerne for at foretage lineær regressionsanalyse (med hjernevægt som respons og kuld størrelsen som forklarende variabel) ser ud til at være opfyldt. I hvilken retning går en evt. sammenhæng?

Ovenfor ses (bl.a.) et scatterplot af hjernevægten mod kuld størrelsen. Hvis vi skal se det separat, og samtidig lægge en udglattet kurve oveni, skriver vi

```
with(mus, scatter.smooth(litter, brain, main="SPM 1", ylab="brain weight",  
xlab="litter size", cex.lab=1.5))
```

hvorved vi får figuren



På tegningen spores en negativ afhængighed af kuld størrelse, således at store kuld fører til mindre hjerner.

2. Bestem estimer for afskæring og hældning i en lineær regressionsanalyse af hjernevægt, med kuld størrelse som forklarende variabel. Forklar resultatet med ord.

Ved hjælp af en simpel lineær regressionsanalyse af hjernevægt, med kuld størrelse som forklarende variabel, estimerer vi denne relation:

```
model1 = lm(brain ~ litter, data=mus)
```

og finder resultatet

```
> summary(model1)
```

Call:

```
lm(formula = brain ~ litter, data = mus)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.030600	-0.006742	0.000183	0.007650	0.032367

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.447000	0.009625	46.443	< 2e-16 ***
litter	-0.004033	0.001198	-3.366	0.00344 **

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

Residual standard error: 0.01539 on 18 degrees of freedom

Multiple R-squared: 0.3862, Adjusted R-squared: 0.3521

F-statistic: 11.33 on 1 and 18 DF, p-value: 0.003445

```
> confint(model1)
```

	2.5 %	97.5 %
(Intercept)	0.426779125	0.46722087
litter	-0.006551127	-0.00151554

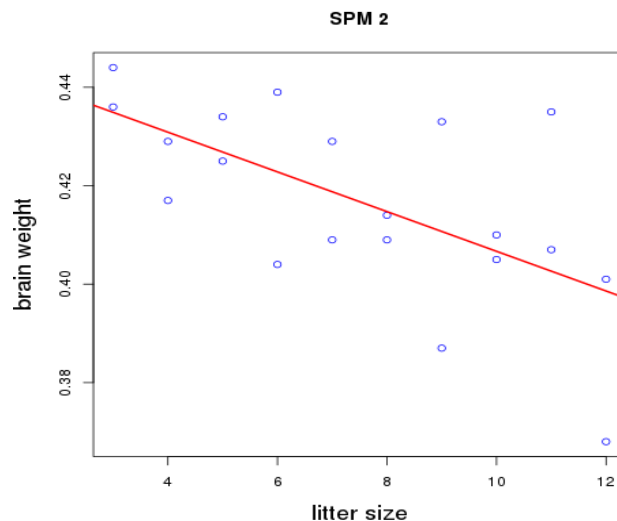
Vi ser (svarende til tegningen), at hældningskoefficienten er negativ

$$\hat{\beta}_1 = -0.0040(0.0012)$$

Dette betyder, at vi for hver ekstra mus i kullet forventer en hjernevægt på gennemsnitligt 0.004g mindre for den enkelte mus.

Vi indlægger regressionslinien på tegningen ovenfor ved at benytte `abline`:

```
plot(mus$litter, mus$brain, main="SPM 2", ylab="brain weight",  
xlab="litter size", cex.lab=1.5, col="blue")  
abline(model1, col="red", lwd=2)
```



(a) *Hvad er den forventede hjernevægt for en mus fra et kuld på 5?*

For at svare på dette spørgsmål, laver vi en ny data frame, kun med en enkelt værdi (5) af `litter`, og predikterer så ud fra denne:

```
litter5 <- data.frame(litter=5)
predict(model1, newdata=litter5, interval="confidence")
```

Vi får så resultatet:

```
> predict(model1, newdata=litter5, interval="confidence")
      fit      lwr      upr
1 0.4268333 0.4172459 0.4364208
```

altså et estimat på 0.427 gram, med konfidensgrænser (0.417, 0.436) gram.

(b) *Er det usædvanligt at se en hjernevægt på 0.4 g for en sådan mus?*

Da dette spørgsmål handler om hjernevægten hos en enkelt mus, skal vi se på prediktionsgrænser i stedet for konfidensgrænser, og

disse dannes ud fra estimatet ved at benytte ± 2 gange residualspredningen, som i outputtet ovenfor ses at være 0.01539.

Vi finder derfor grænserne

$$0.427 \pm 2 \times 0.01539 = (0.396, 0.458)$$

Det er altså lige på kanten af at være usædvanligt med en hjernevægt på 0.4 g for sådanne mus.

Vi kan finde det lidt mere nøjagtigt i R ved at skrive:

```
> predict(model1, newdata=litter5, interval="prediction")
           fit           lwr           upr
1 0.4268333 0.3931005 0.4605661
```

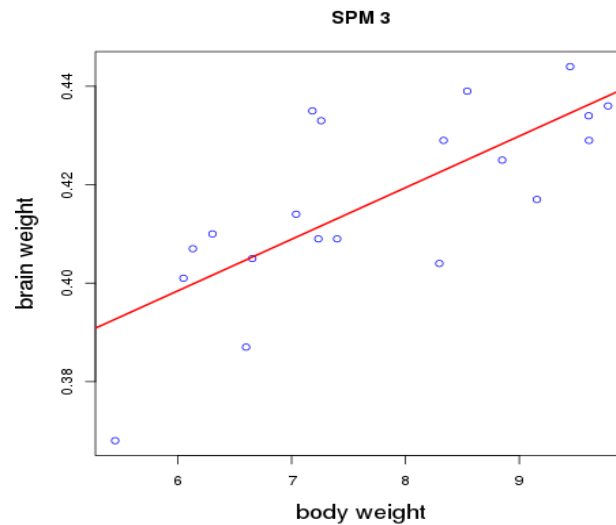
altså intervallet (0.393, 0.461) g.

3. *Undersøg tilsvarende hjernevægtens afhængighed af kropsvægten.*

Nu ser vi tilsvarende på hjernevægtens afhængighed af kropsvægten, hvor regressionsanalysen giver nedenstående resultater, med tilhørende figur:

```
model2 = lm(brain ~ body, data=mus)

plot(mus$body, mus$brain, main="SPM 3", ylab="brain weight",
      xlab="body weight", cex.lab=1.5, col="blue")
abline(model2, col="red", lwd=2)
```



```
> summary(model2)
```

Call:

```
lm(formula = brain ~ body, data = mus)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.024679	-0.004910	-0.001179	0.007464	0.024161

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.335568	0.017318	19.377	1.66e-13 ***
body	0.010479	0.002203	4.756	0.000158 ***

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

Residual standard error: 0.01308 on 18 degrees of freedom

Multiple R-squared: 0.5569, Adjusted R-squared: 0.5323

F-statistic: 22.62 on 1 and 18 DF, p-value: 0.0001578

```
> confint(model2)
```

	2.5 %	97.5 %
(Intercept)	0.29918471	0.37195158
body	0.00585007	0.01510793

Vi ser heraf, at store mus har store hjerner, idet en musehjerne i gennemsnit er 0.010g tungere, når musen vejer 1g mere.

4. *Er der signifikant korrelation mellem kuldstorelse og kropsvægt?*

Vi undersøger nu, ved hjælp af korrelationstest (såvel Pearson som Spearman), om der er signifikant korrelation mellem kuldstorelse og kropsvægt.

Vi udregner de to forskellige versioner af en korrelationskoefficient ved at skrive

```
> with(mus,cor.test(litter,body))

Pearson's product-moment correlation

data:  litter and body
t = -13.615, df = 18, p-value = 6.443e-11
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.9822552 -0.8869040
sample estimates:
      cor
-0.9547218

> with(mus,cor.test(litter,body,method="spearman"))

Spearman's rank correlation rho

data:  litter and body
S = 2594.8, p-value = 1.306e-10
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
-0.9509502

Warning message:
In cor.test.default(litter, body, method = "spearman") :
  Cannot compute exact p-value with ties
```

Hvad enten vi benytter parametrisk (Pearson) eller nonparametrisk (Spearman) korrelation, er der helt klart en sammenhæng mellem kuldstorelse og kropsvægt.

Noget helt andet er så *fortolkningen* af korrelationskoefficienten. Vi kan i hvert fald ikke gå ud fra, at vores observationer passer med en

tidimensional normalfordeling, bl.a. fordi der er 2 af hver kuldstørrelse.

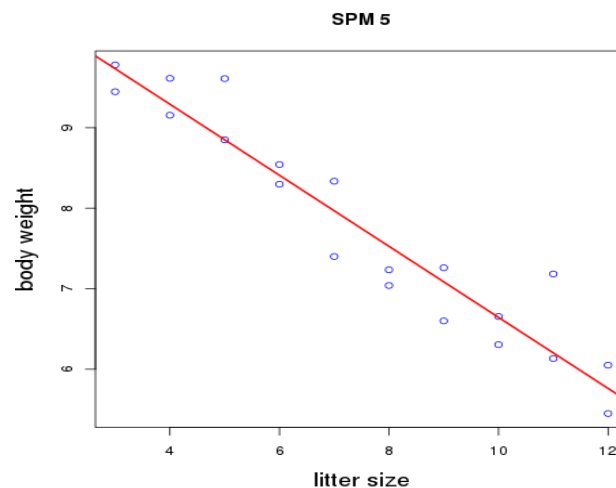
Det ser ud som om denne litter er **valgt** på **systematisk** måde, og den faktiske størrelse af korrelationskoefficienten (hvad enten den er parametrisk eller ej) kan derfor ikke tillægges nogen fornuftig mening (idet korrelationskoefficientens størrelse som bekendt afhænger af samplingmetoden).

5. *Undersøg kropsvægten som funktion af kuldstørrelse og giv en biologisk fortolkning af resultatet.
Hvad er den forventede kropsvægt for en mus fra et kuld på 5?*

Den relevante tegning er

```
model3 = lm(body ~ litter, data=mus)
```

```
plot(mus$litter, mus$body, main="SPM 5", ylab="body weight",  
xlab="litter size", cex.lab=1.5, col="blue")  
abline(model3, col="red", lwd=2)
```



og modellen giver nedenstående resultater:

```
> summary(model3)
```

```

Call:
lm(formula = body ~ litter, data = mus)

Residuals:
    Min       1Q   Median       3Q      Max
-0.56772 -0.29649 -0.03498  0.20320  0.98025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.05642    0.26027   42.48 < 2e-16 ***
litter      -0.44124    0.03241  -13.62 6.44e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4163 on 18 degrees of freedom
Multiple R-squared:  0.9115, Adjusted R-squared:  0.9066
F-statistic: 185.4 on 1 and 18 DF,  p-value: 6.443e-11

> confint(model3)
                2.5 %      97.5 %
(Intercept) 10.5096026 11.6032338
litter      -0.5093289 -0.3731559

```

Også for kropsvægten ses en negativ sammenhæng til kuld størrelsen, idet hældningskoefficienten bliver signifikant negativ (i parentes bemærket er P-værdien præcis den samme, som vi fandt ovenfor i forbindelse med test af den parametriske korrelationskoefficient):

$$\hat{\beta} = -0.441(0.032)$$

Dette betyder, at vi for hver ekstra mus i kullet forventer en kropsvægt på gennemsnitligt 0.441g (knap et halvt gram) mindre for hver enkelt mus.

Estimatet for kropsvægten for mus fra kuld med 5 individer fås ligesom tidligere ved at skrive:

```

> predict(model3, newdata=litter5, interval="confidence")
      fit      lwr      upr
1 8.850206 8.59094 9.109472
> predict(model3, newdata=litter5, interval="prediction")
      fit      lwr      upr
1 8.850206 7.937999 9.762413

```

altså et estimat på 8.85 g, med konfidensinterval $CI=(8.59, 9.11)$ gram og prediktionsinterval $(7.94, 9.76)$ gram.

6. *Foretag en multipel regressionsanalyse med hjernevægt som responsvariabel og såvel kuld størrelse som kropsvægt som forklarende variable.*

Vi har nu set, at der er signifikante sammenhænge mellem alle de tre målte størrelser:

- kropsvægten er negativt relateret til kuld størrelsen
- hjernevægten er positivt relateret til kropsvægten
- hjernevægten er negativt relateret til kuld størrelsen

Vi kan nu med rette spørge os selv, om den lavere hjernevægt blandt mus fra store kuld simpelthen er betinget af, at musene som sådan er mindre i store kuld og derfor også har mindre hjerner. For at undersøge denne påstand skulle man ideelt råde over data fra forskellige størrelser musekuld, hvor musene alle var lige tunge. Hvis hjernevægten her også kunne vises at falde med kuld størrelsen, kunne vi konstatere at kuld størrelse havde en *direkte* effekt på hjernevægten og ikke kun en effekt via kropsvægten.

Sådanne data har vi naturligvis ikke i praksis, men en multipel regressionsanalyse er præcis designet til at svare på dette spørgsmål, idet de enkelte effekter her netop fortolkes som effekten af den relevante kovariat for *fastholdt* værdi af alle de øvrige.

Vi foretager altså en multipel regressionsanalyse med hjernevægt som responsvariabel og såvel kuld størrelse som kropsvægt som forklarende variable:

```
> model4 = lm(brain ~ litter + body, data=mus)
> summary(model4)
```

Call:

```
lm(formula = brain ~ litter + body, data = mus)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.0230428	-0.0098897	0.0006377	0.0092065	0.0180602

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	0.178771	0.075213	2.377	0.02947 *
litter	0.006671	0.003128	2.132	0.04785 *
body	0.024260	0.006769	3.584	0.00229 **

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

Residual standard error: 0.01196 on 17 degrees of freedom

Multiple R-squared: 0.6504, Adjusted R-squared: 0.6093

F-statistic: 15.81 on 2 and 17 DF, p-value: 0.000132

```
> confint(model4)
```

	2.5 %	97.5 %
(Intercept)	2.008551e-02	0.33745581
litter	7.088313e-05	0.01327158
body	9.978796e-03	0.03854132

Vi ser her, at begge kovariater er signifikante (kuldstørrelse er lige på kanten med $P=4.8\%$, medens kropsvægten er tydelig med $P=0.2\%$).

- (a) *Hvad er den forventede hjernevægt for en mus fra et kuld på 5 og en kropsvægt på 10 g?*

Nu skal vi prediktere ud fra 2 kovariater, så vi danner en ny data frame, med værdier af begge disse:

```
l5b10 <- data.frame(litter=5, body=10)
```

Herefter predikterer vi ligesom i de tidligere modeller:

```
> predict(model4, newdata=l5b10, interval="confidence")
      fit      lwr      upr
1 0.4547274 0.4366846 0.4727702
> predict(model4, newdata=l5b10, interval="prediction")
      fit      lwr      upr
1 0.4547274 0.4237155 0.4857393
```

Estimatet bliver altså 0.455 gram, med konfidensinterval (0.437, 0.473) gram, og prediktionsgrænser (0.424, 0.486) gram.

For sådanne mus er det altså **stærkt usædvanligt** at have en hjernevægt på kun 0.4 g.

(b) *Forklar forskellen til resultatet fra spørgsmål 2a.*

I dette spørgsmål ser vi på hjernevægten hos mus, der fra et kuld med 5 individer har opnået en kropsvægt på 10 gram. Dette er ret usædvanlige mus, idet den typiske kropsvægt for sådanne mus var 8.85 g, som vi så det i spørgsmål 5. Prediktionsintervallet (fra spørgsmål 5) for sådanne kropsvægte er (7.94, 9.76) gram, som det ses nedenfor

```
> predict(model3, newdata=litter5, interval="prediction")
      fit      lwr      upr
1 8.850206 7.937999 9.762413
```

og mus fra sådanne kuld når altså sjældent op på 10 gram.

7. *Hvilken biologisk fortolkning har forskellen på koefficienten til litter i model 1 og model 3 i jeres udfyldte skema nedenfor?*

Vi sammenfatter resultaterne i tabellen:

	Respons: brain		koefficient til			
Model nr.	Kovariater		litter	body	residual spredning	% forklaret variation
1	kuldstørrelse		-0.004 (0.001)	–	0.015	38.6
2	kropsvægt		–	0.010 (0.002)	0.013	55.7
3	kuldstørrelse og kropsvægt		0.007 (0.003)	0.024 (0.007)	0.012	65.0

Bemærk her, at afhængigheden af kuldstørrelse skifter fortegn, idet den er negativ, når **body** ikke er med i modellen (model nr. 1), men positiv,

når body medtages (model nr. 3). Fortolkningen af dette følger nedenfor.

Vi kan af dette konkludere

- at hjernevægten er positivt relateret til kropsvægten
 - Hvis mus A vejer $1g$ mere end mus B, forventer vi også at mus A's hjerne vejer i gennemsnit $0.010g$ mere end mus B's (Bemærk, at vi i denne situation typisk har, at mus A kommer fra en mindre kuldstørrelse end mus B)
 - Hvis mus A og B kommer fra to kuld af **samme størrelse**, forventer vi dog en forskel i hjernevægt i gennemsnit på hele $0.024g$, fordi A må betegnes som en usædvanlig stor mus fra en sådan kuldstørrelse.
- at hjernevægten er relateret til kuldstørrelsen, på følgende måde
 - Hvis mus C kommer fra en kuldstørrelse på en flere end mus D, da forventer vi, at mus C's hjerne vejer lidt mindre, i gennemsnit $0.004g$ mindre (Bemærk, at vi i denne situation typisk har, at mus C tillige har en mindre kropsvægt end mus D)
 - Hvis mus C og D alligevel **vejer det samme**, da vil vi forvente, at mus C's hjerne vejer *mest*, (nemlig i gennemsnit $0.007g$ mere end mus D's), fordi mus C er en usædvanlig stor mus fra en sådan kuldstørrelse.

Den interessante konklusion er, at hjernevægten hos mus fra store kuld er **relativt større** set i forhold til kropsvægten end for mus fra små kuld. Den biologiske fortolkning af dette er, at den øgede konkurrence om næringen for store musekuld ikke går ud over hjernen i samme grad som den går ud over kropsvægten.

Dette genfindes i øvrigt for mennesker, idet tvillinger (trillinger, firlinger osv.) generelt er mindre og tyndere, men med vitale organer, der er prioriteret, så de ikke er (meget) mindre end hos andre nyfødte.

Opgavebesvarelse, biomasse

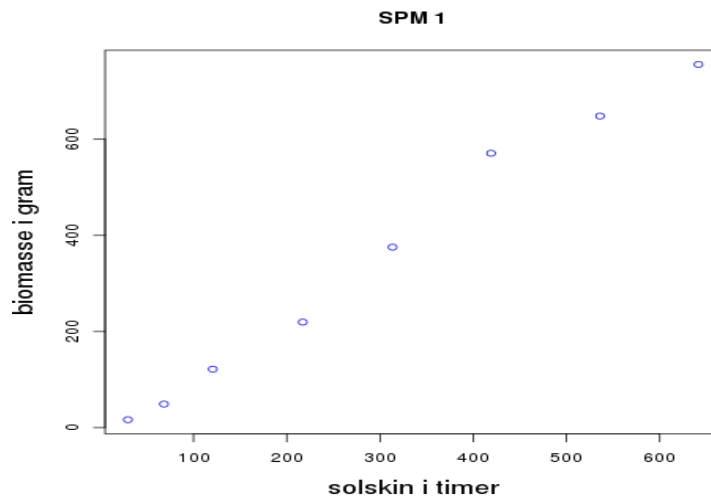
Nedenstående tabel angiver sammenhørende observationer af det kumulerede antal solskinstimer og biomassen af sojabønner over en 8-ugers periode efter spiring. Biomassen er målt som den gennemsnitlige tørvægt i gram af fire uafhængige planter.

<i>nr.</i>	<i>soltimer</i>	<i>biomasse</i>
1	29.7	16.6
2	68.4	49.1
3	120.7	121.7
4	217.2	219.6
5	313.5	375.5
6	419.1	570.8
7	535.9	648.2
8	641.5	755.6

1. *Ser det ud til, at der en lineær sammenhæng mellem de to variable?*

Data indlæses

```
bio <-  
read.table("http://publicifsv.sund.ku.dk/~lts/basal/data/biomasse.txt",  
header=T)  
  
plot(bio$sol, bio$biomass, main="SPM 1", ylab="biomasse i gram",  
xlab="solskin i timer", cex.lab=1.5, col="blue")
```



Plottet ser jo rimeligt lineært ud. En nøjere granskning kan evt. foretages som modelkontrol efter fit af modellen, men da der er så få observationer, giver det ikke rigtigt mening. Bemærk, at en høj korrelation (specielt hvis det er en Spearman) **ikke** sikrer, at sammenhængen er lineær!

Under antagelse om en lineær regressionsmodel med normalfordelte afvigelser, ønskes følgende spørgsmål besvaret:

2. *Giv et estimat for hældningen, med tilhørende 95% sikkerhedsinterval. Undersøg, om hældningen kan antages at være 1.*

Vi skal foretage en sædvanlig lineær regression med **biomass** som respons og **soltimer** som forklarende variabel:

```
> model1 = lm(biomass ~ sol, data=bio)
> summary(model1)
```

```
Call:
lm(formula = biomass ~ sol, data = bio)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-31.052 -14.738  -4.175   5.488  66.428
```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -27.56895   19.84220  -1.389    0.214
sol           1.26925    0.05503   23.063 4.36e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.65 on 6 degrees of freedom
Multiple R-squared:  0.9888, Adjusted R-squared:  0.987
F-statistic: 531.9 on 1 and 6 DF,  p-value: 4.355e-07

> confint(model1)
              2.5 %      97.5 %
(Intercept) -76.121061 20.983157
sol           1.134583  1.403909

```

Vi ser af ovenstående output, at effekten af solskin er stærkt signifikant, idet et test af hældning = 0 giver $T=23.06$, svarende til en P-værdi, der er mindre end 0.0001. Bemærk, at der i outputtet **ikke** findes noget test for linearitet!

Vi er imidlertid ikke blot interesserede i at påvise en effekt af solskin, vi vil *kvantificere* denne i form af et konfidensinterval for hældningen, som ovenfor ses at blive (1.13, 1.40).

Vi kan altså sige, at intervallet (1.13, 1.40) g/time med 95% sandsynlighed indeholder den sande tilvækst i biomasse efter en enkelt solskinstime, og dermed kan vi også svare på, om hældningen kan antages at være 1: Det kan den *ikke*, da 1 ikke er indeholdt i dette konfidensinterval.

Da værdien 1 ikke er indeholdt i dette interval, kan vi således konkludere, at på et 5% signifikansniveau kan vi ikke acceptere hypotesen om at hældningen er 1 (vi forkaster hypotesen $H_0 : \beta = 1$ på et 5% niveau).

Vi kan også direkte udregne et test for $H_0 : \beta = 1$ ved at omskrive hypotesen til $H_0 : \beta - 1 = 0$ og udregne teststørrelsen:

$$\frac{1.2692 - 1}{0.055} = 4.89$$

eller ved i R at skive det *lidt kryptiske*:

```
> t.value=(coef(model1)[2]-1)/coef(summary(model1))[2,2]
> p.value=2*(1-pt(t,6))
> cbind(t.value,p.value)
      t.value      p.value
sol 4.892374 0.002731841
```

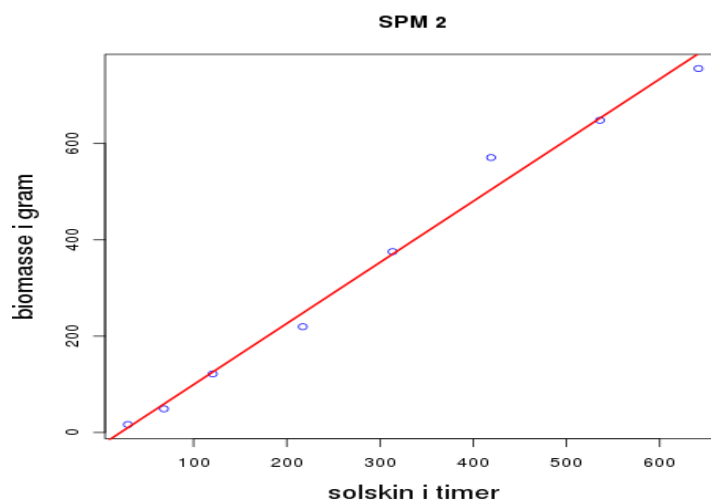
Denne hypotese må altså forkastes, med en P-værdi på 0.0027.

Nu er der jo egentlig heller ikke rigtigt nogen grund til, at β skulle være 1, så vi burde slet ikke have opstillet sådan en hypotese, bare fordi $\hat{\beta}$ så ud til at være tæt på 1!

Der er jo nok ikke nogen, der har indrettet enhederne så viseligt, at timer netop skulle svare til gram....

Den tilhørende figur, med indtegnet linie får vi ved at skrive:

```
plot(bio$sol, bio$biomass, main="SPM 2", ylab="biomasse i gram",
      xlab="solskin i timer", cex.lab=1.5, col="blue")
abline(model1, col="red", lwd=2)
```



3. *Undersøg om interceptet kan antages at være 0.
Hvad bliver hældningsestimatet under denne hypotese?
Hvad sker der med spredningsestimatet for hældningen ved overgang fra modellen med intercept til modellen uden intercept (dvs. intercept=0)?*

Fra outputtet svarende til den ovenfor udførte lineære regressionsanalyse finder vi estimatet for afskæringen (interceptet) med tilhørende spredning (standard error) til

$$\hat{\alpha} = -27.5690(19.8422)$$

T-test størrelsen for test af hypotesen $H_0 : \alpha = 0$ står også i output som -1.39, med en tilhørende P-værdi på 0.21, svarende til, at vi *ikke* kan forkaste denne hypotese (bemærk at vi ikke hermed har *bevist*, at $\alpha = 0$, vi har blot ikke her evidens for det modsatte).

Hvis vi antager, at interceptet er 0, skal vi reestimere hældningen ved at foretage en lineær regressionsanalyse gennem (0,0). Modellen hedder altså

$$Y_i = \beta X_i + \varepsilon_i$$

og den kan fittes i R ved at tilføje -1 i modellen:

```
model2 = lm(biomass ~ sol -1, data=bio)
```

Output bliver:

```
> summary(model2)
```

Call:

```
lm(formula = biomass ~ sol - 1, data = bio)
```

Residuals:

Min	1Q	Median	3Q	Max
-42.572	-26.359	-18.987	-1.848	64.924

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
sol	1.20705	0.03408	35.42	3.71e-09 ***

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

Residual standard error: 34.75 on 7 degrees of freedom
Multiple R-squared: 0.9945, Adjusted R-squared: 0.9937
F-statistic: 1255 on 1 and 7 DF, p-value: 3.712e-09

```
> confint(model2)
          2.5 %    97.5 %
sol 1.12647 1.287637
```

Vores nye hældningsestimat, med tilhørende spredning (standard error) bliver således:

$$\hat{\beta} = 1.2071(0.0341)$$

medens vi i modellen **med intercept** fik

$$\hat{\beta} = 1.2692(0.0550)$$

Vi bemærker, at vi ved at tvinge interceptet til at være 0 (større end det oprindelige estimat, som jo var negativt) har fået et mindre hældningsestimat. Dette sker p.g.a. den negative korrelation mellem disse estimater. Vi bemærker endvidere, at spredningen på estimatet er faldet betragteligt (vi vinder generelt præcision ved at smide insignifikante effekter væk, specielt hvis disse er korrelerede med de interessante effekter). Populært kan man sige at vi nu arbejder i en model med *større viden*, hvilket naturligvis øger vores sikkerhed.

4. *Bestem et 95% sikkerhedsinterval for den estimerede biomasse produktion når det kumulerede antal solskinstimer når op på 200, for modellen med hhv. uden intercept. Forklar forskellen.*

Modellen uden intercept er den letteste. Her er den estimerede biomasse ved 200 solskinstimer blot givet ved

$$200\hat{\beta} = 200 \times 1.2071 = 241.42$$

og usikkerheden er tilsvarende givet ved

$$s.e.(200 \times \hat{\beta}) = 200 \times s.e.(\hat{\beta}) = 200 \times 0.0341 = 6.82$$

således at konfidensgrænserne bliver

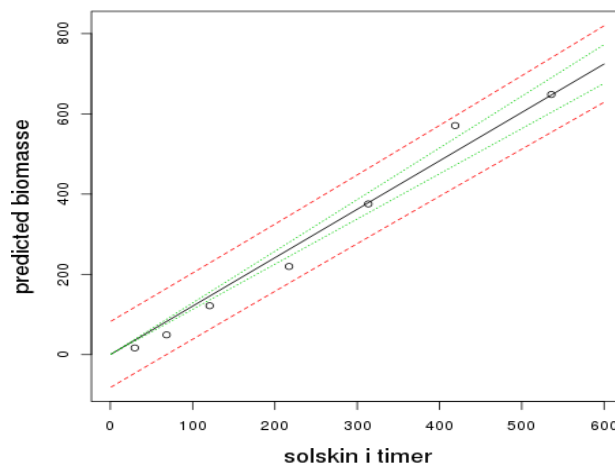
$$200 \times (1.2071 \pm 2.365 \times 0.0341) = (225.29, 257.55)$$

Bemærk, at vi her anvender t-fraktilen svarende til 7 frihedsgrader (i stedet for som tidligere 6). Det er naturligvis fordi vi nu kun har en enkelt parameter i modellen.

Men vi behøver jo ikke at håndregne, for vi kan få R til at gøre det:

```
> sol200 <- data.frame(sol=200)
> predict(model2, newdata=sol200, interval="confidence")
      fit      lwr      upr
1 241.4107 225.294 257.5274
> predict(model2, newdata=sol200, interval="prediction")
      fit      lwr      upr
1 241.4107 157.6682 325.1533
```

Konfidensgrænserne (det er de grønne) kommer til at se således ud:

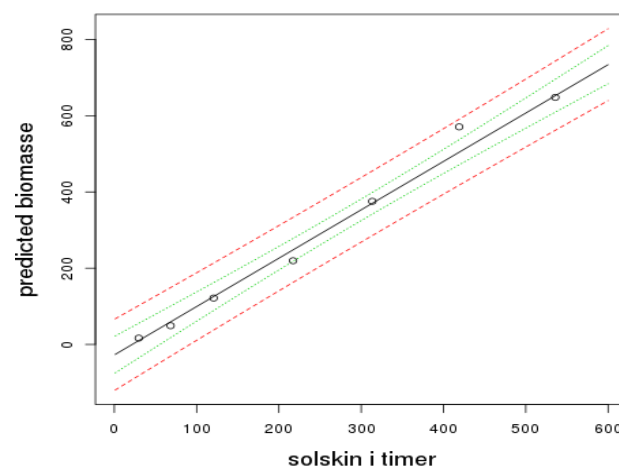


Bemærk her, at konfidensgrænserne er smallest nede ved 0, idet vi har indføjet ekstra *viden* i vores model, nemlig, at interceptet er 0.

For modellen **med** intercept er det vanskeligt at udregne grænserne med håndkraft, men nemt ved brug af R:

```
> predict(modell1, newdata=sol200, interval="confidence")
      fit      lwr      upr
1 226.2803 195.3685 257.1921
> predict(modell1, newdata=sol200, interval="prediction")
      fit      lwr      upr
1 226.2803 140.616 311.9445
```

De tilhørende konfidensgrænser fremgår af nedenstående figur:



Vi ser, at modellen *uden intercept* giver et noget højere predikteret udbytte ved 200 solskinstimer (241.42 mod 226.28 i modellen med intercept), svarende til, at vi stadig er 'i nærheden af 0', hvor linien jo er blevet løftet ved at smide (det negative) intercept ud. Konfidensgrænserne er tillige væsentligt smallere, igen på grund af den øgede præcision i en model med kun en parameter.