

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

<i>kuldstørrelse</i> litter	<i>kropsvægt</i> body	<i>hjernevægt</i> brain
3	9.447	0.444
3	9.780	0.436
4	9.155	0.417
4	9.613	0.429
5	8.850	0.425
5	9.610	0.434
6	8.298	0.404
6	8.543	0.439
7	7.400	0.409
7	8.335	0.429
8	7.040	0.414
8	7.235	0.409
9	6.600	0.387
9	7.260	0.433
10	6.305	0.410
10	6.655	0.405
11	7.183	0.435
11	6.133	0.407
12	5.450	0.368
12	6.050	0.401

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.

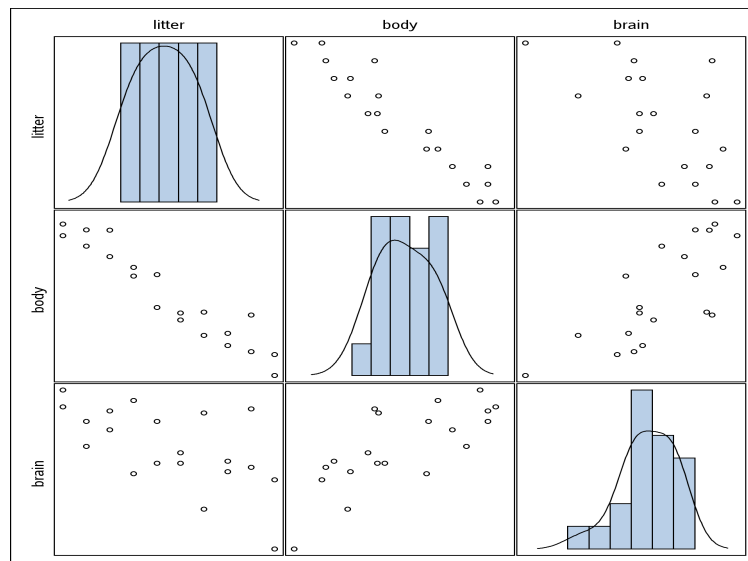
Først indlæser vi ved at skrive

```
data mus;  
infile "http://staff.pubhealth.ku.dk/~lts/basal/data/brain.txt" URL firstobs=2;  
input litter body brain;  
run;
```

Vi har nu 20 observationer, hver med 3 variable, og nedenfor ses de tre plottet parvist mod hinanden. Dette plot kan kun laves med en passende opgraderet version af SAS, ved at skrive:

```
proc sgscatter data=mus;
    matrix _numeric_ / diagonal=(kernel histogram);
run;
```

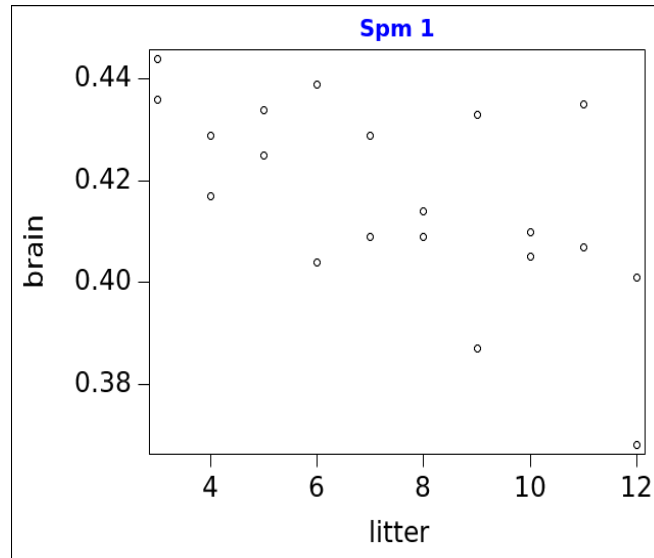
hvorved vi får:



1. Vurder løseligt ud fra en tegning, om forudsætningerne for at foretage lineær regressionsanalyse (med hjernevægt som respons og kuldstørrelsen 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 kuldstørrelsen. Hvis vi skal se det separat, skriver vi

```
proc sgplot data=mus;
    scatter Y=brain X=litter;
run;
```



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

2. *Bestem estimater for afskæring og hældning i en lineær regressionsanalyse af hjernevægt, med kuldstørrelse som forklarende variabel. Forklar resultatet med ord.*

Ved hjælp af en simpel lineær regressionsanalyse af hjernevægt, med kuldstørrelse som forklarende variabel, estimerer vi denne relation, og da vi efterfølgende får brug for en `estimate`-sætning, benytter vi `glm` til at gøre det:

```
proc glm plots=all data=mus;
  model brain=litter / solution clparm;
  estimate "brain for litter=5" intercept 1 litter 5;
run;
```

og finder

The GLM Procedure

```
Number of Observations Read      20
Number of Observations Used      20
```

Dependent Variable: brain

```
R-Square      Coeff Var      Root MSE      brain Mean
0.386227      3.693827      0.015394      0.416750
```

```
Parameter              Estimate      Standard
                        Error          t Value      Pr > |t|
brain for litter=5      0.42683333      0.00456346      93.53      <.0001
```

```
Parameter              95% Confidence Limits
brain for litter=5      0.41724586      0.43642080
```

```
Parameter              Estimate      Standard
                        Error          t Value      Pr > |t|
Intercept              0.4470000000      0.00962476      46.44      <.0001
litter                 -.0040333333      0.00119842      -3.37      0.0034
```

```
Parameter              95% Confidence Limits
Intercept              0.4267791254      0.4672208746
litter                 -.0065511267      -.0015155400
```

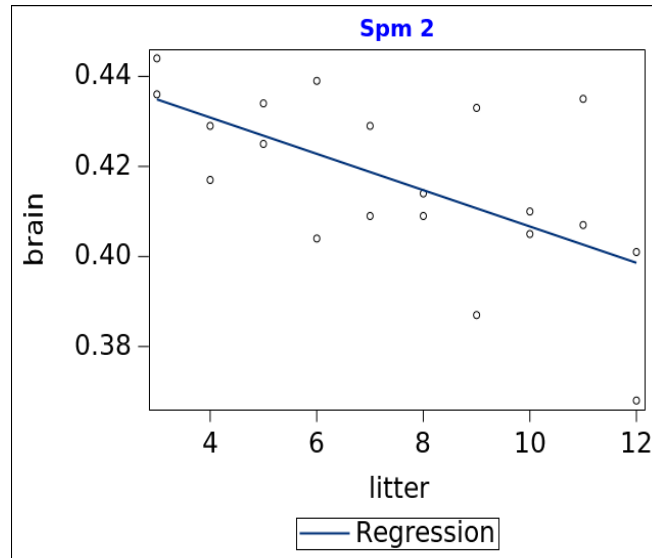
Vi ser (svarende til tegningen), at hældningskoefficienten er negativ

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

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

Vi indlægger regressionslinien på tegningen ovenfor ved at benytte **reg** i stedet for **scatter**:

```
proc sgplot data=mus;
reg Y=brain X=litter;
run;
```



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

For at svare på dette spørgsmål, inkluderede vi ovenfor en **estimate**-sætning:

```
estimate "brain for litter=5" intercept 1 litter 5;
```

og fik heraf resultatet:

Parameter	Estimate	Error	t Value	Pr > t
brain for litter=5	0.42683333	0.00456346	93.53	<.0001
95% Confidence Limits				
brain for litter=5	0.41724586	0.43642080		

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 (Root MSE).

Vi finder derfor grænserne

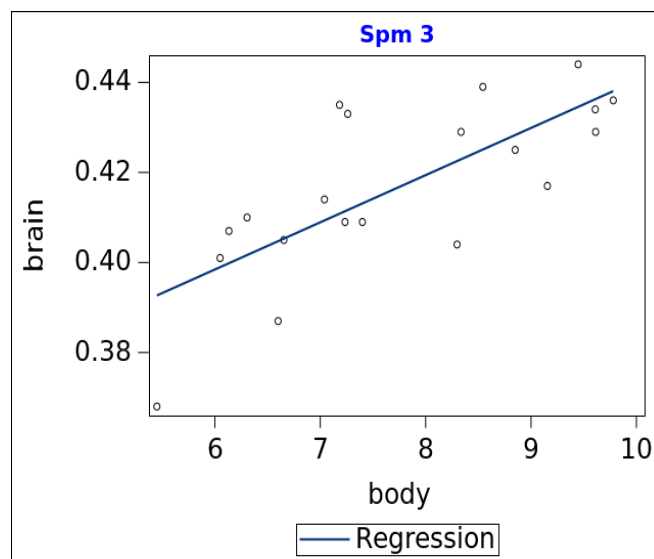
$$0.427 \pm 2 \times 0.01539 = (0.396, 0.458)$$

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

3. *Undersøg tilsvarende hjernevægtens afhængighed af kropsvægten.*

Nu ser vi tilsvarende på hjernevægtens afhængighed af kropsvægten, hvor regressionsanalysen giver nedenstående resultater. Denne gang har vi lavet analysen ved hjælp af `reg`:

```
proc reg data=mus;  
  model brain=body / clb;  
run;
```



The REG Procedure
Dependent Variable: brain

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.00387	0.00387	22.62	0.0002
Error	18	0.00308	0.00017109		
Corrected Total	19	0.00695			
Root MSE	0.01308	R-Square	0.5569		
Dependent Mean	0.41675	Adj R-Sq	0.5323		
Coeff Var	3.13861				

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	0.33557	0.01732	19.38	<.0001
body	1	0.01048	0.00220	4.76	0.0002
Variable	DF	95% Confidence Limits			
Intercept	1	0.29918	0.37195		
body	1	0.00585	0.01511		

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 kuldstørrelse og kropsvægt?

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

Korrelationerne udregnes ved at skrive

```
proc corr pearson spearman data=mus;
  var litter body;
run;
```

hvorved vi får

The CORR Procedure

2 Variables: litter body

Pearson Correlation Coefficients, N = 20

Prob > |r| under H0: Rho=0

	litter	body
litter	1.00000	-0.95472 <.0001
body	-0.95472 <.0001	1.00000

Spearman Correlation Coefficients, N = 20

Prob > |r| under H0: Rho=0

	litter	body
litter	1.00000	-0.95095 <.0001
body	-0.95095 <.0001	1.00000

Hvad enten vi benytter parametrisk (Pearson) eller nonparametrisk (Spearman) korrelation, er der helt klart en sammenhæng mellem kuld-størrelse 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 todimensional normalfordeling, bl.a. fordi der er 2 af hver kuldstørrelse.

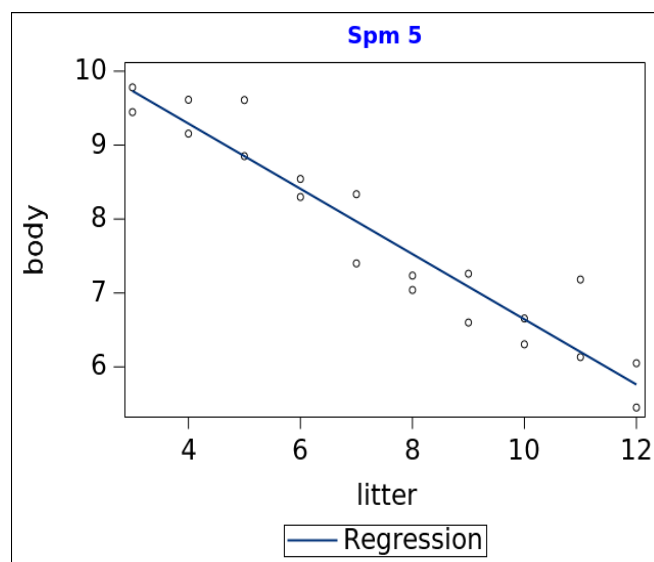
Det ser ud som om denne litter er **valgt** på **systematisk** måde, og den faktiske størrelse af korrelationskoefficienten (hvad enten den er parametrisk eller ej) kan derfor ikke tillægges nogen fornuftig mening (idet korrelationskoefficientens størrelse som bekendt afhænger af sam-plingmetoden).

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

```
proc sgplot data=mus;  
  reg Y=body X=litter;  
run;
```

som giver resultatet



Da vi igen har brug for en `estimate`-sætning, benytter vi `glm`:

```
proc glm data=mus;  
  model body=litter / solution clparm;  
  estimate "body for litter=5" intercept 1 litter 5;  
run;
```

og finder

The GLM Procedure

Dependent Variable: body

R-Square	Coeff Var	Root MSE	body Mean
0.911494	5.373459	0.416287	7.747100

Parameter	Estimate	Standard Error	t Value	Pr > t
body for litter=5	8.85020606	0.12340567	71.72	<.0001

Parameter	95% Confidence Limits	
body for litter=5	8.59094036	9.10947176

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	11.05641818	0.26027411	42.48	<.0001
litter	-0.44124242	0.03240792	-13.62	<.0001

Parameter	95% Confidence Limits	
Intercept	10.50960256	11.60323381
litter	-0.50932893	-0.37315592

Også for kropsvægten ses en negativ sammenhæng til kuldstørrelsen, 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 kuldet 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 ses udfra **estimate**-sætningen at være 8.85 g, med konfidensinterval CI=(8.59, 9.11) gram.

6. *Foretag en multipel regressionsanalyse med hjernevægt som responsvariabel og såvel kuldstørrelse 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 kuldstørrelsen
- hjernevægten er positivt relateret til kropsvægten

- hjernevægten er negativt relateret til kuldstørrelsen

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 kuldstørrelsen, kunne vi konstatere at kuldstørrelse 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 kuldstørrelse som kropsvægt som forklarende variable:

```
proc glm data=mus;
  model brain=litter body / solution clparm;
  estimate "body for litter=5, body=10" intercept 1 litter 5 body 10;
run;
```

hvorved vi får

The GLM Procedure

Number of Observations Used 20

Dependent Variable: brain

R-Square	Coeff Var	Root MSE	brain Mean
0.650391	2.868638	0.011955	0.416750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
litter	1	0.00268418	0.00268418	18.78	0.0005
body	1	0.00183587	0.00183587	12.85	0.0023

Source	DF	Type III SS	Mean Square	F Value	Pr > F
litter	1	0.00064994	0.00064994	4.55	0.0478
body	1	0.00183587	0.00183587	12.85	0.0023

Parameter	Estimate	Standard Error	t Value	Pr > t
body for litter=5, body=10	0.45472740	0.00855182	53.17	<.0001

Parameter	95% Confidence Limits	
body for litter=5, body=10	0.43668464	0.47277016

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	0.1787706594	0.07521281	2.38	0.0295
litter	0.0066712332	0.00312840	2.13	0.0478
body	0.0242600575	0.00676896	3.58	0.0023

Parameter	95% Confidence Limits	
Intercept	0.0200855075	0.3374558112
litter	0.0000708831	0.0132715834
body	0.0099787964	0.0385413186

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?*

Vores **estimate**-sætning giver os svaret 0.455 gram, med konfidensinterval (0.437, 0.473) gram.

Prediktionsintervallet for enkeltindivider er

$$0.455 \pm 2 \times 0.011995 = (0.431, 0.479)$$

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

$$8.85 \pm 2 \times 0.416 = (8.02, 9.68)$$

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 gen-

nemsnit 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.

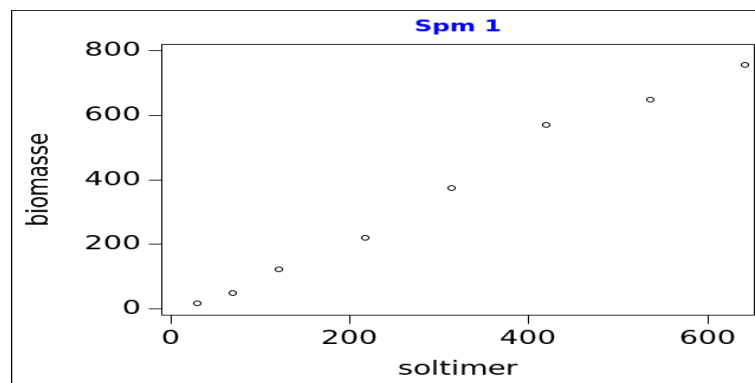
<i>nr.</i>	<i>soltimer</i>	<i>biomasse</i>
1	29.7	16.6
2	68.4	49.1
3	120.7	121.7
4	217.2	219.6
5	313.5	375.5
6	419.1	570.8
7	535.9	648.2
8	641.5	755.6

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

Data indtastes direkte i et SAS-program, og der tegnes:

```
data sol;
input soltimer biomasse;
datalines;
29.7 16.6
68.4 49.1
120.7 121.7
217.2 219.6
313.5 375.5
419.1 570.8
535.9 648.2
641.5 755.6
;
run;

proc sgplot data=sol;
scatter Y=biomasse X=soltimer;
run;
```



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.

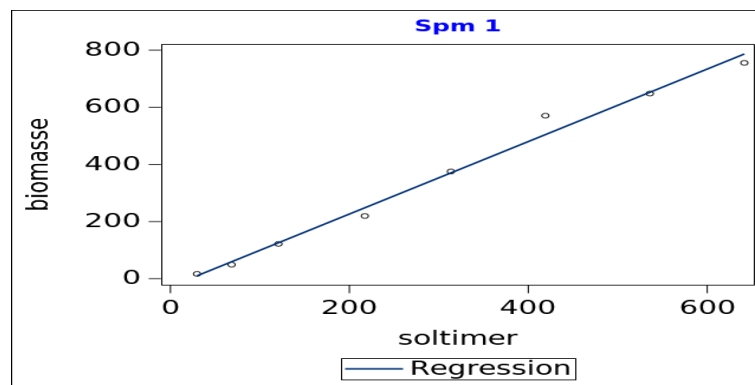
Vi skal foretage en sædvanlig lineær regression med **biomasse** (**Y**) som respons og **soltimer** (**X**) som forklarende variabel:

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

men først tegner vi lige en linie ind på plottet ved at bruge **reg** i stedet for **scatter**:

```
proc sgplot data=sol;  
scatter Y=biomasse X=soltimer;  
run;
```

hvorved vi får



Vi udfører dernæst en lineær regression, idet vi samtidig medtager en **estimate**-sætning, vi skal bruge senere:

```
proc glm plots=all data=sol;  
model biomasse=soltimer / solution clparm;  
estimate "sol=200" intercept 1 sol 200;  
run;
```

Vi får outputtet:

The GLM Procedure

Dependent Variable: biomasse

R-Square	Coeff Var	Root MSE	biomasse Mean
0.988846	9.473830	32.65037	344.6375

Parameter	Estimate	Standard Error	t Value	Pr > t
sol=200	226.280291	12.6329849	17.91	<.0001

Parameter	95% Confidence Limits	
sol=200	195.368490	257.192091

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	-27.56895187	19.84219776	-1.39	0.2141
soltimer	1.26924621	0.05503386	23.06	<.0001

Parameter	95% Confidence Limits	
Intercept	-76.12106073	20.98315698
soltimer	1.13458321	1.40390922

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. Samme P-værdi får man ved at anvende F-testet, idet $F=23.06^2=531.90 \sim F(1,6)$. 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 jo er tilvæksten i biomasse forårsaget af en enkelt solskinstime. Estimatet med tilhørende spredning (standard error) er

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

med enheder g/time.

Et 95% konfidensinterval fås som estimat $\pm$ ca. 2 gange spredningen. Nu er de 'ca. 2' jo egentlig en t-fraktil, som for store datamaterialer nærmer sig 1.96. Her har vi kun 8 observationer, og dermed 6 frihedsgrader til estimation af variansen (se også ovenstående output), og t-fraktilen er derfor en del større end 2. Vi kan slå den op i bogen s. 521 og finder værdien 2.447, og vi kan nu udregne konfidensintervallet til:

$$1.2692 \pm 2.447 \times 0.055 = (1.13, 1.40)$$

sådan ca. det samme som vi fik i outputtet.

Vi kan altså sige, at intervallet (1.13, 1.40) g/time med 95% sandsynlighed indeholder den sande tilvækst i biomasse efter en enkelt solskin-time.

Da værdien 1 ikke er indeholdt i dette interval, kan vi med det samme sige, 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:

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

som helt klart er for stor, svarende til at vi allerede ved, at hypotesen skal forkastes. P-værdien kan slås op i t-tabellen (med sølle 6 frihedsgrader), hvilket giver $0.001 < P < 0.01$.

Man kan også gøre det med SAS ved at indføje en `test`-sætning i en `proc reg`:

```
proc reg data=sol;
  model biomasse=soltimer / clb;
  test soltimer=1;
run;
```

som vil give det ekstra output

Test 1 Results for Dependent Variable biomasse

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	25516	23.94	0.0027
Denominator	6	1066.04667		

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....

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 nu

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

og den kan fittes ved at benytte option `noint` i `model`-sætningen:

```
proc glm data=sol;
model biomasse=soltimer / noint solution clparm;
estimate "sol=200" soltimer 200;
run;
```

Output bliver:

```

The GLM Procedure
Number of Observations Used      8

Dependent Variable: biomasse

R-Square      Coeff Var      Root MSE      biomasse Mean
0.994451      10.08384      34.75268      344.6375

NOTE: No intercept term is used: R-square is not corrected for the mean.

Parameter              Estimate      Standard      t Value      Pr > |t|
                        Error
sol=200                241.410723      6.81574766      35.42      <.0001

Parameter              95% Confidence Limits
sol=200                225.294041      257.527406

Parameter              Estimate      Standard      t Value      Pr > |t|
                        Error
soltimer              1.207053617      0.03407874      35.42      <.0001

Parameter              95% Confidence Limits
soltimer              1.126470206      1.287637028

```

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

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

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

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

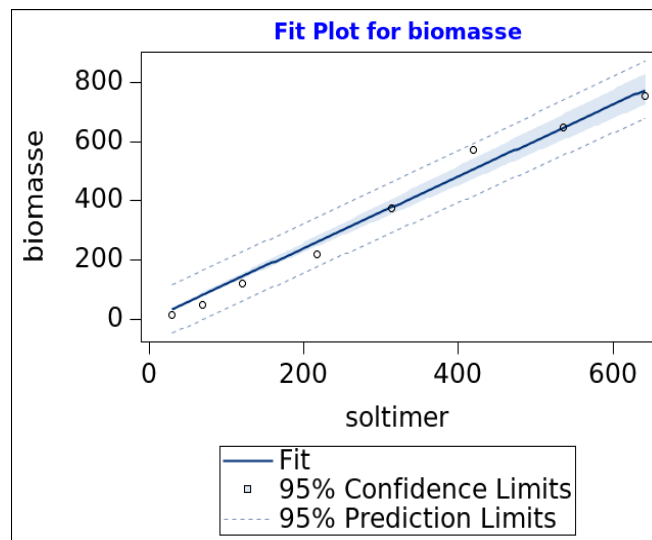
således at konfidensgrænserne bliver

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

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 havde allerede resultatet fra `estimate`-sætningen ovenfor.

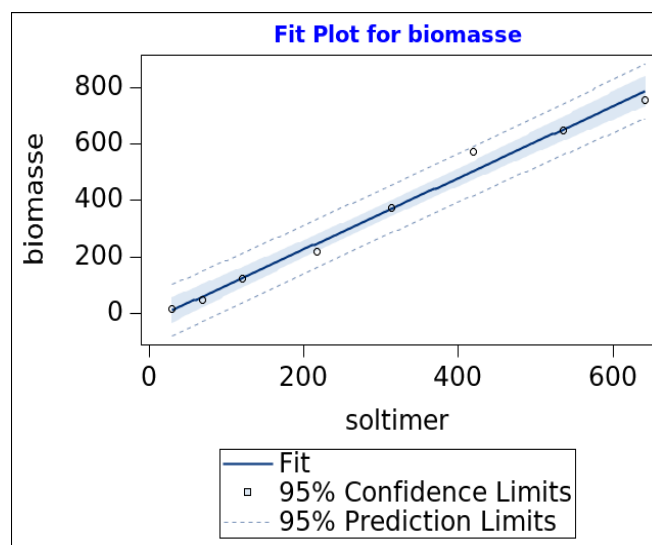
Konfidensgrænserne kommer til at se således ud:



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 der har vi også ovenfor set resultatet fra **estimate**-sætningen, nemlig et estimat på 226.28 med konfidensintervallet (195.37, 257.19).

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.