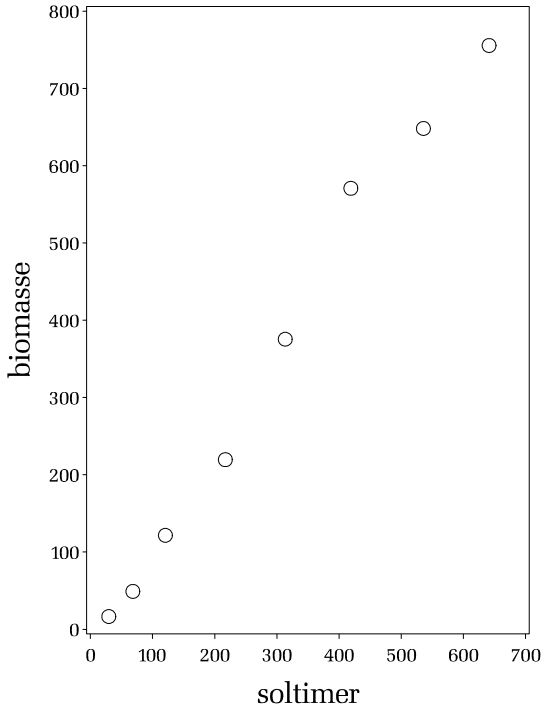
Spørgsmål 1.

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

Data indtastes direkte i `Analyst` og gemmes i `sasuser`. Variabelnavnene er her `soltimer` hhv. `biomasse`. Derefter tegnes et Scatter plot vha `Graphs/Scatter Plot/Two-dimensional`.



Figur 1: Scatter plot.

Plottet (figur 1) ser jo rimeligt lineært ud. En nøjere granskning kan evt. foretages som modelkontrol efter fit af modellen. Bemærk, at en høj korrelation (specielt hvis det er en Spearman) ikke sikrer, at sammenhængen er lineær!

Under antagelse af en lineær regressionsmodel med normalfordelte fejl, ønskes følgende spørgsmål besvaret:

Spørgsmål 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:

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

I SAS gør vi dette ved at benytte
`Statistics/Regression/Linear` og i `Plots` kan vi afkrydse
`Plot observed vs independents`, hvis vi ønsker en tegning med indlagt regressionslinie.

Vi får outputtet:

Dependent Variable: biomasse

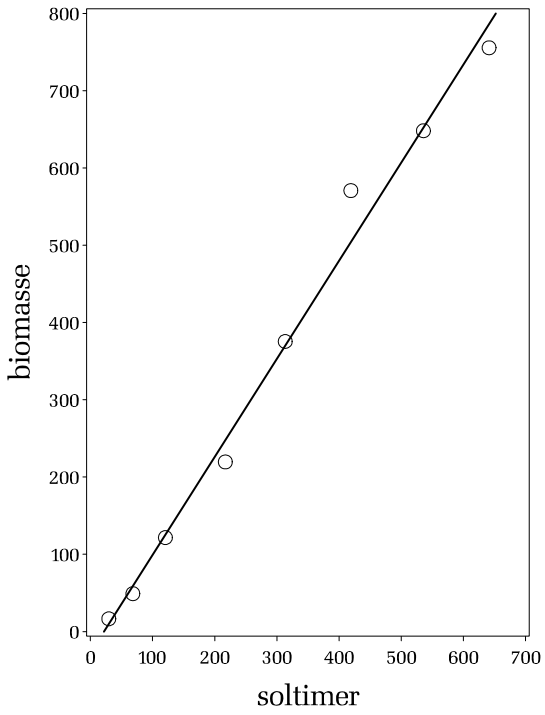
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	567033	567033	531.90	<.0001
Error	6	6396.28000	1066.04667		
Corrected Total	7	573429			

Root MSE	32.65037	R-Square	0.9888
Dependent Mean	344.63750	Adj R-Sq	0.9870
Coeff Var	9.47383		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-27.56895	19.84220	-1.39	0.2141
soltimer	1	1.26925	0.05503	23.06	<.0001



Figur 2: Scatter plot med indlagt regressionslinie.

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

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

og 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:

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

Det samme fås ved i regressionsanalyseopsætningen at afkrydse `Confidence limits for estimates`:

Parameter Estimates

Variable	DF	95% Confidence Limits	
Intercept	1	-76.12106	20.98316
soltimer	1	1.13458	1.40391

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

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 $0.001 < P < 0.01$. Mere eksakt finder man P=0.003.

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!

Spørgsmål 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 kan fittes i `Analyst` ved i opsætningen at gå ind under `Model` og afkrydse feltet `Do not include an intercept`. Output bliver:

```
Dependent Variable: biomasse

NOTE: No intercept in model. R-Square is redefined.

                          Analysis of Variance

                                    Sum of          Mean
Source                      DF       Squares        Square   F Value   Pr > F

Model                        1       1515175       1515175   1254.54   <.0001
Error                        7    8454.24118    1207.74874
Uncorrected Total            8       1523629

Root MSE              34.75268    R-Square     0.9945
Dependent Mean       344.63750    Adj R-Sq     0.9937
Coeff Var             10.08384
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
soltimer	1	1.20705	0.03408	35.42	<.0001

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.

Spørgsmål 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

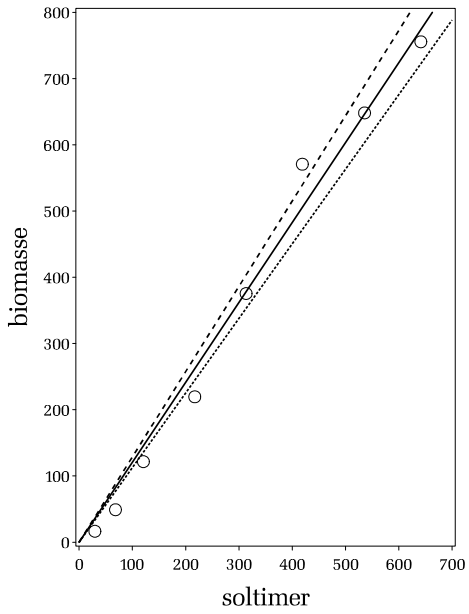
$$s.e.(200 \times \hat{\beta}) = 200 \times s.e.(\hat{\beta}) = 200 \times 0.0341 = 6.82$$

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.

Konfidensgrænserne kommer til at se således ud:



Figur 3: Konfidensgrænser i modellen uden intercept.

For modellen **med** intercept er det vanskeligt at udregne grænserne med håndkraft; her er det lettest at bruge den ovenfor beskrevne metode med `Predictions` eller ved at flytte nulpunktet hen i 200 soltimer ved at fratrække 200 fra alle soltime-observationerne, hvorved vi ville få:

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	226.28029	12.63298	17.91	<.0001
sol200	1	1.26925	0.05503	23.06	<.0001

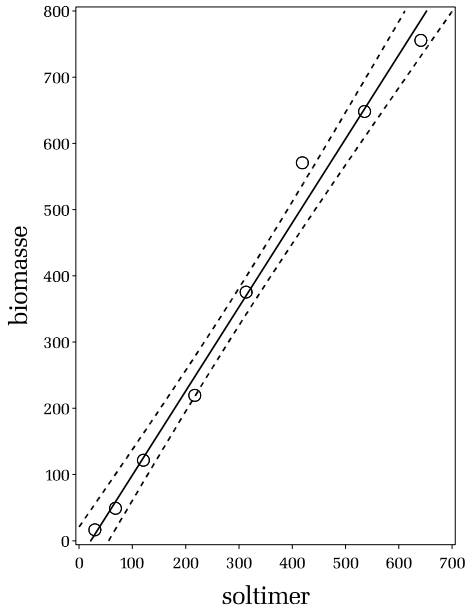
```
              Parameter Estimates

Variable      DF       95% Confidence Limits

Intercept      1      195.36849      257.19209
sol200         1        1.13458        1.40391
```

altså et estimat på 226.28 med konfidensintervallet (195.37,257.19).

Det tilhørende konfidensinterval fremgår af nedenstående figur:



Figur 4: Konfidensgrænser i modellen med intercept.

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 interceptet ud. Konfidensgrænserne er tillige væsentligt smallere, igen på grund af den øgede præcision i en model med kun en parameter.