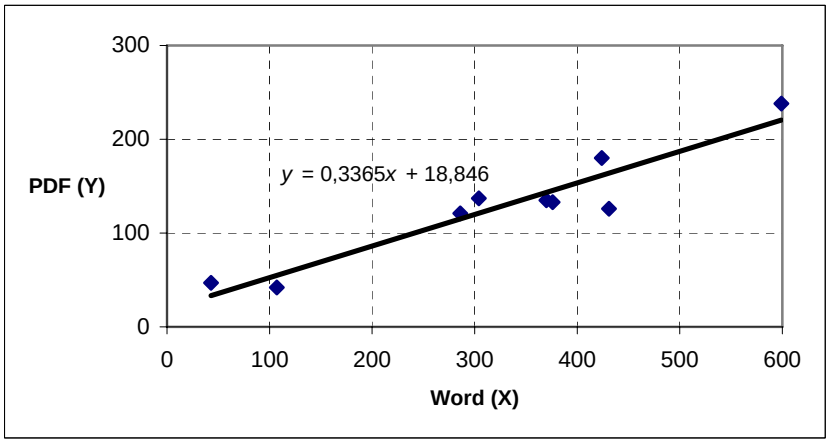
10. Korrelasjon og regresjon.

10.1. Innledning.

Jeg skriver disse forelesningsnotatene i Word. Deretter konverterer jeg dem til PDF-format før de legges ut på nettet. Dette skyldes dels at PDF-filene er mindre enn de tilsvarende Word-filene. Størrelsene i kB av Word-filene og de tilhørende PDF-filene for noen av de notatene som jeg har laget hittil er:

Word	X	43	304	370	376	599	424	286	431	107
PDF	Y	47	137	135	133	238	180	121	126	42

Samhørende verdier av X og Y kan nå oppfattes som koordinatene til et punkt. Samlingen av slike punkt kalles et ***spredningsdiagram***. Ved hjelp av Excel har jeg laget et slikt spredningsdiagram over disse punktene. Jeg har også fått Excel til å tegne inn en *lineær trendlinje*, og skrive ut likningen for den:



Likningen for trendlinja er altså

$$y = 0.3365x + 18.846 .$$

For et punkt som ligger på trendlinja, vil denne likningen fortelle oss at:

- Dersom $x = 0$ (d.v.s. at Word-fila er tom) vil likevel PDF-fila være på 18.846 kB.
- Hver gang Word-fila øker med 1 kB, vil PDF-fila øke med 0.3365 kB.

Jeg håper virkelig at du betrakter opplysningene ovenfor med en blanding av nysgjerrighet og sunn skepsis. Punktene i diagrammet kan jo oppfattes som et *tilfeldig utvalg* av Word/PDF-tallpar, hentet fra en populasjon av slike tallpar. Vi må jo forvente å få andre punkter og en annen trendlinje dersom vi trekker et annet tilfeldig utvalg fra populasjonen. Kan vi – ut fra dataene i vårt utvalg – si noe om denne *populasjonen*:

- Er det rimelig å tro at det i det hele tatt fins en lineær sammenheng mellom X og Y i populasjonen?
- Kan vi si noe om usikkerheten i verdiene for stigningstall og skjæring med Y-aksen? Er det mulig å sette opp konfidensintervall for disse parametrene? Kan det for eksempel tenkes at skjæringspunktet med Y-aksen er lik null når vi tar hensyn til alle punktene i populasjonen?
- Og ikke minst: Hvordan beregnes koeffisientene i trendlinja?

Vi vil få bruk for noen sumformler for å foreta de nødvendige beregningene. For oversiktens skyld skal jeg samle disse formlene her, og hente dem fram etter behov.

Vi har et *utvalg* av *n* tallpar $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$. Vi definerer:

Middelverdier: $\bar{x}=\frac{1}{n}\sum x_i, \quad \bar{y}=\frac{1}{n}\sum y_i,$

$$S_{xx}=\sum(x_i-\bar{x})^2=\sum x_i^2-n\cdot\bar{x}^2=\sum x_i^2-\frac{1}{n}\left(\sum x_i\right)^2$$

$$S_{yy}=\sum(y_i-\bar{y})^2=\sum y_i^2-n\cdot\bar{y}^2=\sum y_i^2-\frac{1}{n}\left(\sum y_i\right)^2$$

$$S_{xy}=\sum(x_i-\bar{x})(y_i-\bar{y})=\sum x_iy_i-n\cdot\bar{x}\cdot\bar{y}=\sum x_iy_i-\frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right)$$

Alle summene skal tas fra $i=1$ til $i=n$.

Merk at vi bruker *store S'er* i de tre siste definisjonene. Hver av de tre likningene består av selve definisjonen, samt omforminger som vanligvis gir lettere regninger.

Både kalkulatorer og dataprogram kan beregne kvadratsummer. Sjekk med ditt dataverktøy hvordan disse summene beregnes. Med Excel får jeg for mitt innledende eksempel:

$\bar{x}=326.67,\quad \bar{y}=128.78,$

$S_{xx}=229\,724,\quad S_{yy}=28\,963.6,\quad S_{xy}=77\,308.3.$

Du må ikke forveksle disse kvadratsummene med varianser eller standardavvik, som har symbol s^2 (varians) og *s* (standardavvik). Merk: *Små* bokstaver.

Nå som vi opererer med både *X*- og *Y*-verdier, må vi ha varians / standardavvik for både *X* og *Y*. Med de kvadratsummene som vi definerte ovenfor, får vi disse formlene:

Varians for *X* er $s_x^2=\frac{1}{n-1}\sum(x_i-\bar{x})^2=\frac{S_{xx}}{n-1}.$

Varians for *Y* er $s_y^2=\frac{1}{n-1}\sum(y_i-\bar{y})^2=\frac{S_{yy}}{n-1}.$

Et praktisk lite tips: Dersom du har dataverktøy som beregner varianser, kan du finne S_{xx} og S_{yy} ved å multiplisere variansene med $(n-1)$.

Vi kan også definere en ny størrelse som vi kaller ***kovariansen*** av *X* og *Y* slik:

Kovarians $s_{xy}=\frac{1}{n-1}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{S_{xy}}{n-1}.$

Dersom ditt dataverktøy beregner kovarians, kan du finne S_{xy} med $S_{xy} = (n-1)s_{xy}$.

Nå har vi de størrelsene vi trenger for å gå i gang. Vi skal ta problemene i denne rekkefølgen:

- Først skal vi definere ***korrelasjonskoeffisienten***, som er et mål for hvor nær punktene ligger en rett linje. Vi skal både beregne korrelasjonskoeffisienten for vårt utvalg, og prøve å si noe om korrelasjonskoeffisienten i populasjonen.
- Deretter skal vi vise hvordan vi kommer fram til likningen for trendlinja, eller ***regresjonslinja*** som den egentlig heter.
- Til slutt skal vi vurdere usikkerheten til koeffisientene i regresjonslinja.

10.2. Lineær korrelasjon.

Et spredningsdiagram er et nyttig hjelpemiddel til å se om punkter ligger langs en rett linje. Men vi trenger et mer presist redskap. Her er det ***korrelasjonskoeffisienten*** kommer inn i bildet. Vi definerer:

Vi har et sett av *n* tallpar $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$.
Korrelasjonskoeffisienten mellom *x* og *y* er

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}}$$

der S_{xx}, S_{yy} og S_{xy} er definert tidligere.

Vi kan vise at denne korrelasjonskoeffisienten har mange nyttige egenskaper:

- $-1 \le r \le 1$.
- Dersom $|r|$ er nær 1, ligger punktene nær en *rett* linje. Dersom $|r|$ er nær 0, ligger ikke punktene nær en rett linje (men punktene kan ligge nær eller på krumme linjer).
- Dersom $|r|$ er nær 1 og $r > 0$, ligger punktene nær en rett linje med *positivt* stigningstall. Dersom $|r|$ er nær 1 og $r < 0$, ligger punktene nær en rett linje med *negativt* stigningstall.

Merk at vi kun tester på *lineær* korrelasjon. Punktene kan ligge pent samlet nær krumme linjer (andregradsfunksjoner, eksponentialfunksjoner osv.) uten at korrelasjonskoeffisienten avslører dette.

Eksempel 10.1: Beregn korrelasjonskoeffisienten for våre Word / PDF-data fra innledningen av dette notatet.

Løsning: Benytter de kvadratsummene jeg har funnet tidligere, og finner

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} = \frac{77\,308.3}{\sqrt{229\,724 \cdot 28\,963.6}} = \underline{0.948}.$$

Vi får en verdi for *r* svært nær 1. Dette betyr at punktene ligger nær en rett linje med positivt stigningstall, noe spredningsdiagrammet også bekrefter.

Hvor nær 1 bør $|r|$ ligge for at vi kan være rimelig sikker på at det virkelig er en lineær sammenheng mellom *X* og *Y* også i *populasjonen*? For å svare på dette, kan vi benytte setningen nedenfor:

Dersom korrelasjonskoeffisienten i populasjonen er ρ, og vi beregner korrelasjons-koeffisienten *r* for et tilfeldig utvalg på *n* tallpar fra populasjonen, kan vi vise at

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}}$$

er *t*-fordelt med $n-2$ frihetsgrader.

Dette kan vi benytte til å teste om $\rho = 0$ i populasjonen, d.v.s. teste om en tilsynelatende korrelasjon bare skyldes tilfeldigheter.

Eksempel 10.2: Undersøk om den korrelasjonskoeffisienten som vi fant i eksemplet foran, er signifikant forskjellig fra 0.

Løsning: Vi setter opp en formell hypotesetest:

$H_0: \ \rho = 0.$

$H_1: \ \rho > 0$

Under H_0 får vi for våre data

$$t = \frac{r - \rho}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.948 - 0}{\sqrt{\frac{1-0.948^2}{9-2}}} = \underline{7.88}.$$

Denne *t*-verdien er så stor at vi forkaster H_0 uansett signifikansnivå. Vi slår fast at det er en lineær sammenheng mellom størrelsen av Word-fila og den tilhørende PDF-fila. Og det var jo ikke noen stor overraskelse i vårt eksempel. I andre situasjoner kan en slik test være nyttig.

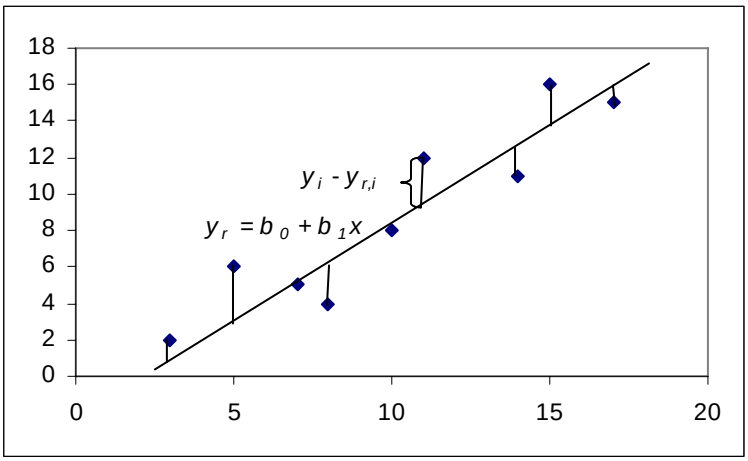
10.3. Lineær regresjon.

Vi har altså et sett av *n* tallpar $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$. I disse tallparene kalles gjerne *x* den ***uavhengige*** variable, mens *y* kalles den ***avhengige*** variable. Disse begrepene antyder at vi antar at verdien av *y* på en eller annen måte avhenger av verdien av *x* (nokså naturlig i vårt Word / PDF-eksempel).

Vi skal konstruere ei ***regresjonslinje*** med likningen

$$y_r = b_0 + b_1 x .$$

Da må vi finne koeffisientene b_0 og b_1. Vi gjør det på følgende måte: For hvert punkt (x_i, y_i) trekker du ei loddrett linje til regresjonslinja. Lengden av denne linja blir $y_i - y_{r,i}$ der $y_{r,i}$ er y-verdien til regresjonslinja for $x = x_i$. Denne loddrette avstanden fra regresjonslinja kalles punktets ***residual***. Deretter skal du bestemme b_0 og b_1 slik at kvadratsummen av residualene, $\sum (y_i - y_{r,i})^2$, blir minst mulig.



Utledningene er vist i et vedlegg. Resultatet blir:

Regresjonslinja er

$$y_r = b_0 + b_1 x$$

der stigningstallet b_1 er gitt ved

$$b_1 = \frac{S_{xy}}{S_{xx}}$$

og skjæringspunktet b_0 med y-aksen er gitt ved

$$b_0 = \bar{y} - b_1 \bar{x}$$

Eksempel 10.3: Undersøk om Excel regnet riktig da regresjonslinja (”trendlinja”) ble satt opp.

Løsning: Med våre data får vi:

$$b_1 = \frac{S_{xy}}{S_{xx}} = \frac{77\,308.3}{229\,724} = \underline{0.3365} .$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x} = 128.78 - 0.3365 \cdot 326.67 = \underline{18.856} .$$

Regresjonslinja blir

$$y_r = b_0 + b_1 x = 18.856 + 0.3365 x .$$

Det stemmer jo bra med trendlinja til Excel. Litt avvik må vi regne med som følge av avrundinger.

10.4. Regresjonslinjas variasjon.

Vi har tidligere påpekt at vårt punktsett kan oppfattes som et *tilfeldig utvalg* av alle punktsettene fra en stor populasjon. Dersom vi trekker ut et annet utvalg fra populasjonen, må vi forvente en annen regresjonslinje. Hvor store variasjoner kan vi forvente i regresjonslinje når vi velger slike tilfeldige punktsett? Mer presist: Hvor store variasjoner i b_0 og b_1 må vi forvente?

For å kunne svare på det spørsmålet, må vi legge inn et par forutsetninger:

- Punktsettene i *populasjonen* ligger langs ei regresjonslinje $y = \beta_0 + \beta_1 x$. Vi kan vise at b_0 og b_1 er beste estimat for β_0 og β_1.
- For en gitt verdi av *x* i populasjonen, er *y*-verdiene normalfordelt med standardavvik σ rundt et punkt som ligger på regresjonslinja $y = \beta_0 + \beta_1 x$.
- Standardavviket σ er like stort for alle *x*.

Legg merke til at σ ikke er standardavviket for alle *y*-ene. σ er standardavviket for *y*-verdiene til de punktene som har *samme x*-verdi, d.v.s. standardavviket til residualene.

Som illustrasjon har jeg laget et regneark der punkter trekkes tilfeldig fra en populasjon som tilfredsstiller kravene ovenfor. Merk hvordan den observerte regresjonslinja varierer!

Nå vil vi veldig gjerne vite verdien av σ. Det gjør vi dessverre (nesten) aldri. Vi må nøye oss med et estimat av σ. Vi kan vise at:

Beste estimat for σ er ***estimatets standard feil***

$$s_e = \sqrt{\frac{\sum\left(y_i - y_{r,i}\right)^2}{n-2}}$$

der

$$y_{r,i} = b_0 + b_1 x_i .$$

Når σ erstattes med s_e, må vi bruke *t*-fordeling med $n-2$ frihetsgrader.

Du ser sikkert at telleren i brøken er kvadratsummen av residualene, som vi har definert tidligere. Denne kvadratsummen er plundrete å regne ut. Med litt triksing kan vi omforme kvadratsummen til uttrykkene nedenfor, som er enklere å bruke i praksis:

$$\begin{aligned}\sum\left(y_i - y_{r,i}\right)^2 &= S_{yy} - b_1 S_{xy} \\ &= \sum y_i^2 - b_0 \cdot n \cdot \bar{y} - b_1 \sum x_i y_i\end{aligned}$$

Så til saken. Vi kan vise at:

Anta at regresjonslinja til punktene i populasjonen er

$$y = \beta_0 + \beta_1 x .$$

Beste estimat for denne regresjonslinja er

$$y_r = b_0 + b_1 x .$$

b_1 er normalfordelt med forventningsverdi β_1 og standardavvik

$$\sigma_{b_1} = \frac{\sigma}{\sqrt{S_{xx}}} .$$

b_0 er normalfordelt med forventningsverdi β_0 og standardavvik

$$\sigma_{b_0} = \sigma\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} .$$

I praksis må vi (nesten) alltid erstatte σ med estimatets standard feil s_e som er definert ovenfor.

La oss se hva disse formlene fører til i vårt Word / PDF-eksempel. Vi fant regresjonslinja

$$y_r = 18.856 + 0.3365x$$

der *x* er størrelsen av Word-fila mens y_r er størrelsen av den tilhørende PDF-fila dersom punktet ligger på regresjonslinja. Tallet $b_0 = 18.856$ må tolkes som antall kB data som legges inn i PDF-fila utenom selve konverteringen av Word-fila, mens $b_1 = 0.3365$ angir hvor mange kB PDF-fila øker med for hver kB Word-fila øker. Men disse tallene gjelder bare for vårt tilfeldige utvalg av Word / PDF-kombinasjoner. Kan vi si noe om verdiene av de tilsvarende koeffisientene i *populasjonen* av Word / PDF-kombinasjoner? Her er et par problemstillinger:

Eksempel 10.4: Anta at den regresjonslinja vi har funnet:

$$y_r = 18.856 + 0.3365x$$

er et tilfeldig utvalg av regresjonslinjer som kan beregnes for en populasjon med regresjonslinje

$$y = \beta_0 + \beta_1 x .$$

a) Finn et 95 % konfidensintervall for stigningstallet β_1 i populasjonen.

b) Er det sikkert at $\beta_0 > 0$? Test med 5 % signifikans.

Løsning: Vi starter med å finne estimatet for standardavviket σ rundt regresjonslinja:

$$\hat{\sigma} = s_e = \sqrt{\frac{\sum (y - y_r)^2}{n-2}} = \sqrt{\frac{S_{yy} - b_1 S_{xy}}{n-2}} = \sqrt{\frac{28963.6 - 0.3365 \cdot 77308.3}{9-2}} = \underline{20.53} .$$

a) Beste estimat for standardavviket for β_1 er

$$\hat{\sigma}_{b_1} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{20.53}{\sqrt{229724}} = \underline{0.043} .$$

Feilmarginen i konfidensintervallet er

$$E = t_{\alpha/2} \cdot \hat{\sigma}_{b_1} = 2.365 \cdot 0.043 = \underline{0.1017}$$

der $t_{\alpha/2}$ er funnet av *t*-tabell med 7 frihetsgrader. Et 95 % konfidensintervall for β_1 er da

$$\left[b_1 - E,\ b_1 + E\right] = \left[0.3365 - 0.1017,\ 0.3365 + 0.1017\right] \approx \underline{\underline{\left[0.235,\ 0.438\right]}}.$$

b) Vi formulerer problemer som en hypotesetest:

$H_0:\ \beta_0 = 0$.

$H_1:\ \beta_0 > 0$.

Beste estimat for standardavviket til β_0 er

$$\hat{\sigma}_{b_0} = \hat{\sigma} \cdot \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}} = 20.53 \cdot \sqrt{\frac{1}{9} + \frac{326.67^2}{229\,724}} = \underline{15.58}.$$

Testobservatoren blir

$$t = \frac{b_0 - \beta_0}{\hat{\sigma}_{b_0}} = \frac{18.856 - 0}{15.58} = \underline{1.21}.$$

Av et *t*-tabell med 7 frihetsgrader ser vi at kritisk verdi er $t = 1.895$. Vi er ikke i det kritiske området, og kan ikke forkaste H_0. Våre data gir derfor ikke grunnlag for å påstå at det legges inn noen data i PDF-fila utenom den konverterte Word-fila.

La oss avslutte med en problemstilling til. Anta at vi har ei Word-fil på x_0 kB. Kan vi si noe om hvor stor den tilhørende PDF-fila blir? Mer presist: Kan vi sette opp et konfidensintervall for størrelsen av PDF-fila?

Vi har sett at det er knyttet usikkerhet til hvor *populasjonens* regresjonslinje $y = \beta_0 + \beta_1 x$ egentlig går. I tillegg er det spredning *rundt* denne regresjonslinja. Vi kan imidlertid vise at:

For et tallpar $(x_0,\ y)$ der x_0 er et fast tall, er beste estimat for forventningsverdien til *y* gitt ved $\hat{y} = y_r(x_0) = b_0 + b_1 \cdot x_0$
mens standardavviket til *y* gitt ved

$$\sigma_{y|x_0} = \sigma \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}}.$$

I praksis må vi (nesten) alltid erstatte σ med estimatet s_e, og deretter bruke *t*-tabell med $n - 2$ frihetsgrader. En forenklet utledning av formlene i de to siste rammene er tatt med i et vedlegg.

Eksempel 10.5: Vi fortsetter med dataene fra vårt eksempel. Anta at vi har ei Word-fil på $x_0 = 500$ kB. Sett opp et 90 % konfidensintervall for størrelsen av den tilhørende PDF-fila.

Løsning: Forventningsverdien til *y* er

$$\hat{y} = b_0 + b_1 \cdot x_0 = 18.856 + 0.3365 \cdot 500 = \underline{187.1}.$$

Estimat for standardavviket til y er

$$\hat{\sigma}_{y|x_0} = \hat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 20.53 \cdot \sqrt{1 + \frac{1}{9} + \frac{(500 - 326.67)^2}{229724}} = \underline{22.88}.$$

Feilmarginen i konfidensintervallet er

$$E = t_{\alpha/2} \cdot \hat{\sigma}_{b_1} = 1.895 \cdot 22.88 = \underline{43.4}$$

der $t_{\alpha/2}$ er funnet av t-tabell med 7 frihetsgrader. Et 90 % konfidensintervall for y er da

$$\left[\hat{y} - E,\ \hat{y} + E\right] = \left[187.1 - 43.4,\ 187.1 + 43.4\right] \approx \underline{\underline{\left[143.7,\ 230.5\right]}}.$$

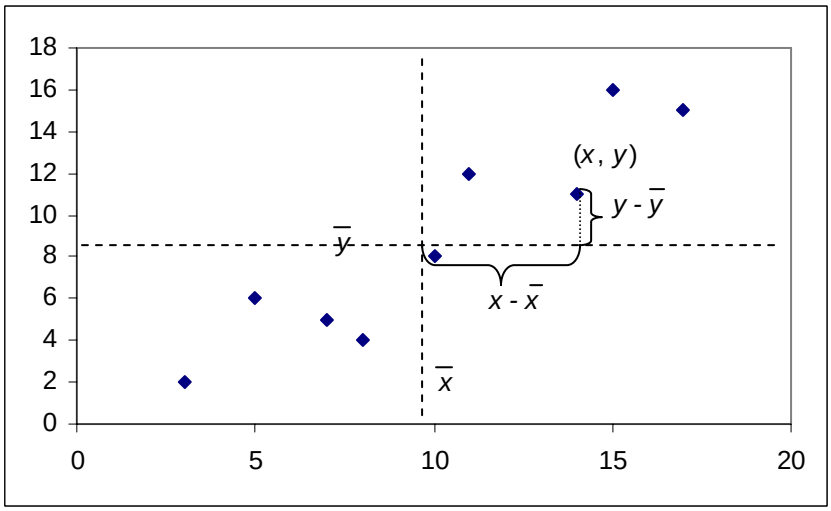
Dette vil si at med 90% konfidens kan vi si at når Word-fila er på 500 kB, vil den tilhørende PDF-fila ligge mellom 143.7 kB og 230.5 kB.

10.5. Noen tilleggsbetraktninger.

I avsnittene foran har jeg konsentrert med om de "matnyttige" sidene ved dette temaet. Nå er det på tide å se nærmere på noen andre sider, som forhåpentlig vil bidra til å øke forståelsen for denne delen av statistikken. Noe av dette krever ferdigheter i matematikk, men skulle ikke være verre enn at du bør kunne henge med.

10.5.1. Mer om kovarians og korrelasjon.

Figuren nedenfor viser noen punkter i et spredningsdiagram:



Du ser at for det avmerkede punkt med koordinater $(x,\ y)$ er både $(x - \bar{x})$ og $(y - \bar{y})$ positive. Da blir også $(x - \bar{x})(y - \bar{y})$ positiv. Det samme gjelder for alle punkter der $(x > \bar{x}) \wedge (y > \bar{y})$. Men det gjelder også for alle punkter der $(x < \bar{x}) \wedge (y < \bar{y})$. Av figuren ser du at nesten alle punktene er plassert slik at $(x - \bar{x})(y - \bar{y}) > 0$. Da er også

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) > 0.$$

Dersom punktene ligger nær en linje med negativt stigningstall, vil $(x-\bar{x})$ og $(y-\bar{y})$ ha motsatte fortegn for de fleste punktene. For disse punktene blir $(x-\bar{x})(y-\bar{y})<0$, slik at S_{xy} blir negativ. Dette forklarer at

$$r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{xy}}}$$

er positiv når linja har positivt stigningstall, og negativ når linja har negativt stigningstall.

Vi kan skrive

$$r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{xy}}}=\frac{\frac{1}{n-1}S_{xy}}{\sqrt{\frac{1}{n-1}S_{xx}\cdot\frac{1}{n-1}S_{xy}}}=\frac{s_{xy}}{\sqrt{s_x^{\,2}\cdot s_y^{\,2}}}=\frac{s_{xy}}{s_x\cdot s_y}$$

der s_{xy} er kovariansen til *x* og *y* mens s_x og s_y er standardavvikene til henholdsvis *x* og *y*. Noen lærebøker definerer korrelasjonskoeffisienten som

$$r=\frac{s_{xy}}{s_x\cdot s_y}.$$

10.5.2. Mer om residualene.

Du husker sikkert at størrelsen $(y_i-y_{r,i})$ angir avviket fra regresjonslinja, og kalles *residual*. Vi beregnet koeffisientene til regresjonslinja ved å kreve at $\sum(y_i-y_{r,i})^2$ skal være minst mulig. Denne kvadratsummen kalles gjerne "**S**um of **S**quared **E**rrors", forkortet **SSE**, slik at

$$\text{SSE}=\sum(y_i-y_{r,i})^2$$

er et naturlig mål for spredningen rundt regresjonslinja. Dette gjelder også for andre regresjonslinjer enn vår lineære.

På samme måte er

$$S_{yy}=\sum(y_i-\bar{y})^2$$

et naturlig mål for den *totale* spredningen av *y*-verdiene. Vi kan nå definere en størrelse

$$R^2=1-\frac{\text{SSE}}{S_{yy}}=1-\frac{\sum(y_i-y_{r,i})^2}{\sum(y_i-\bar{y})^2}.$$

Dersom det ikke er spredning rundt regresjonslinja, blir $\sum(y_i-y_r)^2=0$ slik at $R^2=1$. Dersom spredningen rundt regresjonslinja er lik den totale spredningen, er

$$\sum(y_i-y_{r,i})^2=\sum(y_i-\bar{y})^2$$

slik at $R^2=0$. Vi ser at R^2 kan brukes som mål for hvor stor del av den totale spredningen av *y*-verdier som skyldes spredning rundt regresjonslinja.

For lineær regresjon kan vi vise at R^2 er kvadratet av korrelasjonskoeffisienten *r*, slik denne er definert tidligere i notatet. R^2 kan derfor oppfattes som en generalisert korrelasjons-koeffisient som gjelder for andre typer regresjon enn lineær regresjon.