ECON 2130
HG, mars 2010

Notat til kapittel 4 i Løvås

Om enkel lineær regresjon I

1 Innledning

Enkel regresjonsanalyse dreier seg om å studere sammenhengen mellom en responsvariabel, *y*, og en forklaringsvariabel, *x*, basert på et datamateriale som består av *n* observasjonspar av (x, y): $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Anta vi er interessert i sammenhengen mellom tidene på 1500m og 5000m for skøyteløpere som er i tilstrekkelig god form til å kunne delta i store mesterskap som EM, VM og lignende. Data er hentet fra europamesterskapet (EM) i Heerenveen i 2004 og gitt i tabell 1.

(Data kan lastes ned i en Excel-fil på http://folk.uio.no/haraldg/)

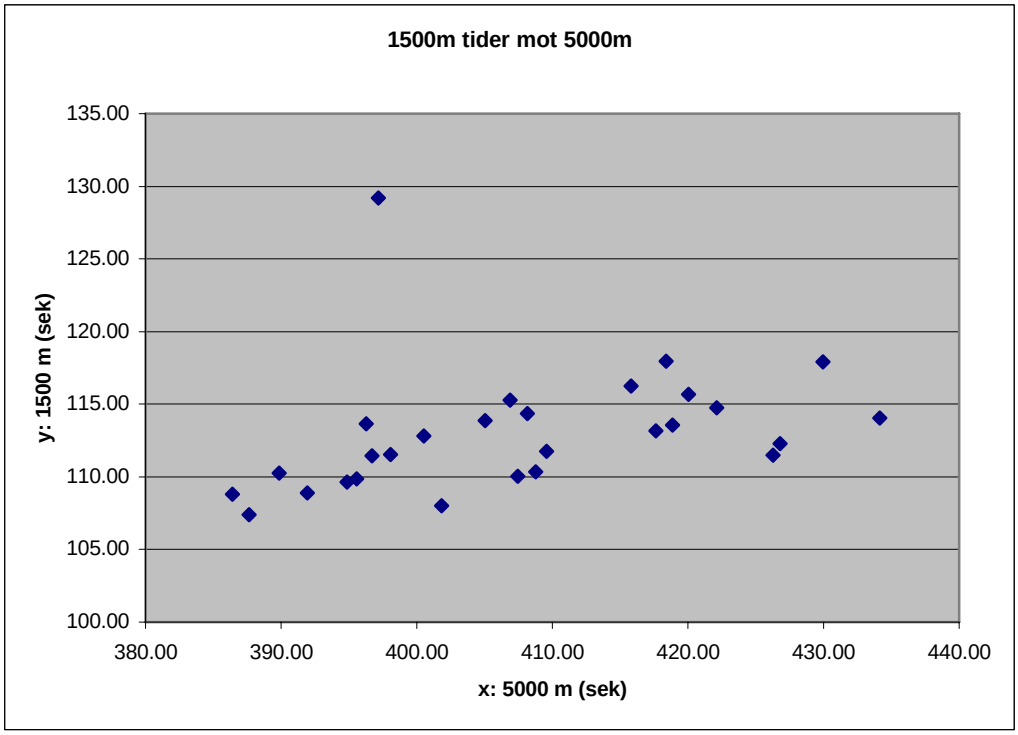
Tabell 1 *Resultater fra 1500m og 5000m for menn fra EM i skøyter Heerenveen 2004*

Obs		5000m		1500m		Obs		5000m		1500m	
nr.		Tid	Sekunder	Tid	Sekunder	nr.		Tid	Sekunder	Tid	Sekunder
1	Tuitert NED	6:27.63	387.63	1:47.41	107.41	17	Poutala FIN	7:06.81	426.81	1:52.27	112.27
2	Verheijen NED	6:26.43	386.43	1:48.80	108.80	18	Rosendahl FIN	7:06.25	426.25	1:51.49	111.49
3	Uytdehaage NED	6:31.93	391.93	1:48.90	108.90	19	Vreugdenhil BEL	6:58.85	418.85	1:53.56	113.56
4	Romme NED	6:29.88	389.88	1:50.24	110.24	20	Zoller AUT	7:14.14	434.14	1:54.05	114.05
5	Skobrev RUS	6:35.55	395.55	1:49.84	109.84	21	Makovetskij BLR	7:02.10	422.10	1:54.75	114.75
6	Lalenkov RUS	6:41.84	401.84	1:48.01	108.01	22	Veldkamp BEL	6:48.17	408.17	1:54.36	114.36
7	Fabris ITA	6:34.87	394.87	1:49.64	109.64	23	Vtípil CZE	7:00.04	420.04	1:55.67	115.67
8	Röjler SWE	6:36.68	396.68	1:51.46	111.46	24	Grozea ROU	6:55.81	415.81	1:56.25	116.25
9	Friesinger GER	6:48.76	408.76	1:50.34	110.34	25	Bosker SUI	6:58.39	418.39	1:57.96	117.96
10	Sætre NOR	6:36.25	396.25	1:53.65	113.65	26	Pedos UKR	7:09.93	429.93	1:57.94	117.94
11	Detyshev RUS	6:38.06	398.06	1:51.54	111.54	27	Valtonen FIN	6:57.62	417.62	1:53.16	113.16
12	Andersen NOR	6:47.43	407.43	1:50.02	110.02	28	Mazur POL	6:46.91	406.91	1:55.28	115.28
13	Zygmunt POL	6:40.54	400.54	1:52.80	112.80						
14	Anesi ITA	6:49.58	409.58	1:51.77	111.77						
15	Schneider GER	6:45.05	405.05	1:53.85	113.85						
16	Ervik NOR	6:37.15	397.15	2:09.20	129.20						

La x_i betegne 5000m-tiden i sekunder, og y_i 1500m-tiden i sekunder for løper nr. *i*, $i = 1, 2, \ldots, 28$. Datamaterialet vårt (i utvalget) består altså av $n = 28$ observasjonspar, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.
For få en ide om hva slags sammenheng som kan være aktuell, kan man lage et spredningsdiagram (scatter plot) der *y*-ene plottes mot *x*-ene.

> [For å få Excel til å tegne dette plottet bør vi lage to kolonner ved siden av hverandre, der x_i -ene ligger i den ene og y_i -ene i den andre kolonnen. Deretter, marker de to kolonnene og velg scatter fra graf-rutinene.]

Figur 1

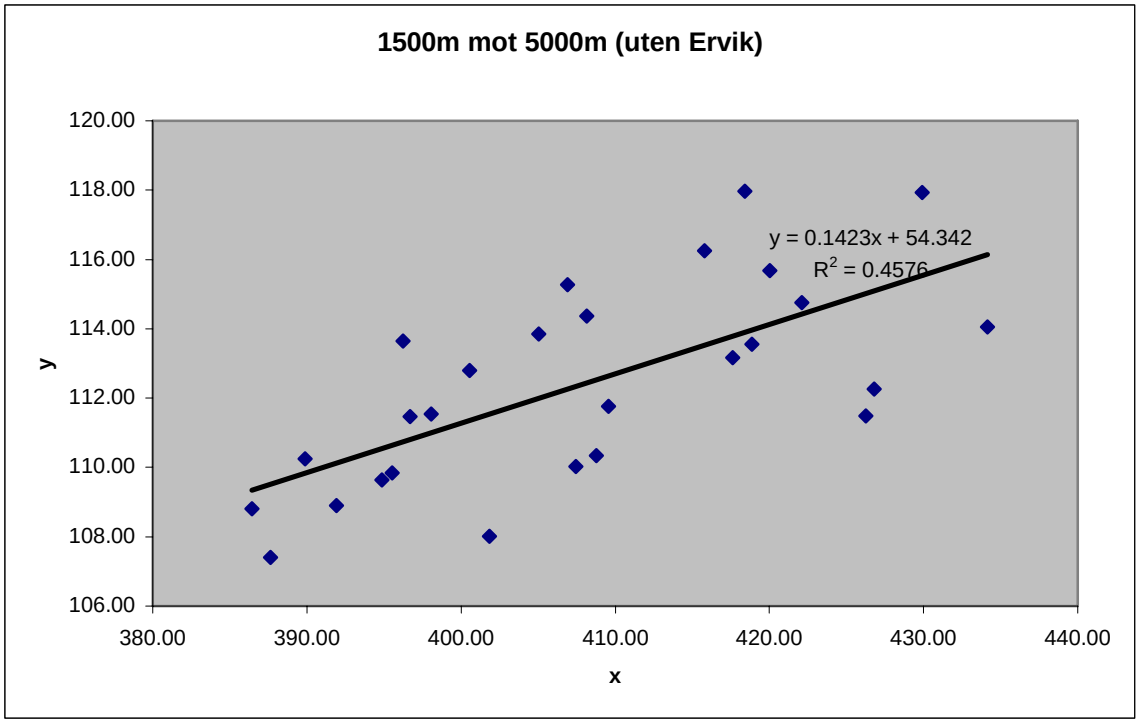


Merk at Excel valgte origo i punktet (380, 100) istedenfor i (0, 0).

Vi legger også merke til et enslig isolert punkt langt over de øvrige. Det viser seg å gjelde Eskil Ervik som falt på 1500m. Ettersom denne observasjonen ikke er representativ for det vi er interessert i (nemlig skøytetider for løpere som holder seg på beina), kan vi trygt fjerne dette observasjonsparet fra data. Vårt datamateriale består dermed av $n = 27$ observasjonspar.

Figur 1 tyder på at en eventuell sammenheng mellom y_i -ene og x_i -ene synes å være av lineær type. I figur 2 har jeg plottet dataene på nytt (uten Eskil Ervik) med en innlagt (minste kvadraters) trendlinje som uttrykk for denne sammenhengen.

Figur 2



For å få fram dette plottet i Excel, laget jeg først spredningsdiagrammet som over. Deretter høyreklikket jeg på et av punktene i diagrammet slik at punktene ble markert og valgte "add trendline" fra menyen som kom fram. I "options" på samme meny spesifiserte jeg at Excel skal skrive ut ligningen for den rette trendlinjen og R^2 som er et mål på hvor stor del av den totale variasjonen av y_i-ene i data er forklart av trend linjen. Siden $R^2 = 0.4576$, betyr det at 45.76% av total-variasjonen av y_i-ene i data er forklart av trendlinja – som indikerer en viss sammenheng.

Trendlinja er et eksempel på en "minste kvadraters regresjonslinje", som er bestemt som den linja som i en viss forstand best beskriver punktene i spredningsplottet.

I dette notatet skal vi først og fremst forklare hvordan disse størrelsene er beregnet. Analysen gjelder kun datamaterialet selv og innebærer (foreløpig) ingen tolkning om populasjonen dataene er trukket fra. For å kunne tolke resultatene i forhold til populasjonen, trenger vi apparatet etablert i kapittel 6 og 7 i Løvås.

2 Noen relevante utvalgsstørrelser

La $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ være *n* observasjonspar av to variable *x* og *y* (som kan være hva som helst – ikke nødvendigvis stokastiske). Da kan vi definere og regne ut (i utvalget) visse størrelser som blant annet er viktig i regresjonsanalyse (jfr. Løvås kap. 7):

Gjennomsnitt:

i) $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ ii) $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Empiriske varianser (også kalt utvalgsvarianser eller sampelvarianser):

iii) $s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ iv) $s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

Empirisk kovarians (også kalt utvalgskovarians eller sampelkovarians):

v) $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$

Empirisk korrelasjonskoeffisient (også kalt utvalgs-korrelasjonskoeffisienten eller sampel-korrelasjonskoeffisienten):

vi) $r = r_{xy} = \frac{s_{xy}}{\sqrt{s_x^2}\sqrt{s_y^2}} = \frac{s_{xy}}{s_x s_y}$

Merk at disse størrelsene, som alltid kan beregnes ut fra data, er bygget opp på samme måte som tilsvarende størrelser definert i populasjonen, der vanlige gjennomsnitt erstattes av forventninger. Hvis X og Y er to stokastiske variable i en populasjon, defineres populasjonsgjennomsnittene ved forventningsverdier, $\mu_X = E(X)$, $\mu_Y = E(Y)$, (populasjons)variansene ved, $\sigma_X^2 = E\left[\left(X - E(X)\right)^2\right]$, $\sigma_Y^2 = E\left[\left(Y - E(Y)\right)^2\right]$, og (populasjons)kovariansen ved, $\text{cov}(X,Y) = E\left[\left(X - E(X)\right)\left(Y - E(Y)\right)\right]$.

(Populasjons)korrelasjonskoeffisienten defineres ved $\rho = \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)}\sqrt{\text{var}(Y)}}$.

I det spesielle tilfellet at $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ kan betraktes som uavhengige observasjoner av (X,Y), vil utvalgs-størrelsene i)-vi) kunne betraktes som anslagsverdier for de tilsvarende populasjonsstørrelsene (intuitivt begrunnet ved de store talls lov og mer presist begrunnet ved statistisk teori som etableres delvis i dette kurset og senere i Stat2).

Egenskaper ved korrelasjonskoeffisientene.
Korrelasjonskoeffisienten, enten det gjelder r eller ρ, er et tall mellom -1 og 1 og måler i hvilken grad sammenhengen mellom x og y (X og Y i populasjonen) kan beskrives ved en rett linje. Ekstremverdiene -1 og 1 svarer til en situasjon der alle observasjonene ligger eksakt på en rett linje. I så fall finnes det konstanter a og b slik at $y_i = a + bx_i$ for alle $i = 1, 2, \ldots, n$ i data, eller $Y = a + bX$ for alle mulige observasjoner av X og Y i populasjonen. Når det gjelder r vil disse egenskapene bli klargjort i dette notatet.

Regneksempel. For å illustrere beregningene i et mindre materiale trakk jeg ut rent tilfeldig[1] fem av observasjonsparene i tabell 1 (minus Ervik). De observasjonsnumrene i tabell 1 som ble trukket ut var, obs.nr. 11, 4, 17, 25 og 23[2] i tabell 1, som jeg kalte $i = 1, 2, 3, 4, 5$ i tabell 2 nedenfor.

Tabell 2 Mini-utvalg på fem trukket fra 27 observasjonspar.

i	x_i	y_i	$x_i - \overline{x}$	$y_i - \overline{y}$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$	$(x_i - \overline{x})(y_i - \overline{y})$
1	398.06	111.54	-12.58	-2.00	158.1558	3.9840	25.1017
2	389.88	110.24	-20.76	-3.30	430.8115	10.8636	68.4118
3	426.81	112.27	16.17	-1.27	261.5983	1.6028	-20.4763
4	418.39	117.96	7.75	4.42	60.1245	19.5718	34.3037
5	420.04	115.67	9.40	2.13	88.4352	4.5540	20.0681
sum	2053.18	567.68			999.1253	40.5761	127.4090
Gj.snitt	410.636	113.536					

De fem sentrale størrelsene kan nå lett beregnes:

$\overline{x} = 410.636 \qquad \overline{y} = 113.536$

$s_x^2 = 999.1253 / 4 = 249.7813$

$s_y^2 = 40.5761 / 4 = 10.1440$

$s_{xy} = 127.4090 / 4 = 31.8523$

Korrelasjonskoeffisienten i miniutvalget på 5 blir da

$$r = \frac{s_{xy}}{s_x s_y} = \frac{31.8523}{\sqrt{249.7813}\sqrt{10.1440}} = 0.633.$$

Den tilsvarende korrelasjonen i hele materialet (27 observasjonspar) ble 0.676. Sjekk selv dette ved å bruke Excel. Rutinen "Covariance" i modulen "Data analysis" gir deg s_x^2, s_y^2, s_{xy} svakt modifisert (se fotnote 6 side 15). Rutinen "Correlation", også i "Data analysis", gir deg *r* direkte.

De fem størrelsene, $\overline{x}, \overline{y}, s_x^2, s_y^2, s_{xy}$, er alt vi trenger for å beregne en enkel regresjon-analyse. Det er en kjedelig jobb å beregne disse fem størrelsene med kalkulator, med stor sjanse for å regne feil, så denne jobben er best å gjøre med computer. Når de først er beregnet, vil alle

[1] Jeg lot Excel trekke ut fem løpere for meg ved å bruke "Random number generation" med uniform fordeling fra modulen "Data analysis".
[2] Dvs. løperne, Detyshev (RUS), Romme (NED), Poutala (FIN), Bosker (SUI) og Vtípil (CZE) hhv.

andre formler som er aktuelle for en enkel regresjon med bare en forklaringsvariabel, *x*, lett kunne beregnes med kalkulator.

Om man likevel trenger å beregne hele regresjonsanalysen med kalkulator, kan følgende beregningsformler være nyttige:

Regel 1 (Utledet i appendiks)

Hvis $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ er *n* observasjonspar, gjelder

(a) $(n-1)s_x^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$

(b) $(n-1)s_y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$

(c) $(n-1)s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\,\bar{y}$

3 Minste kvadraters regresjonslinje i utvalget

Vi har *n* observasjoner av variablene[3] *x* og *y* , $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, og ønsker å forklare mest mulig av y_i-ene ved hjelp av x_i-ene og en rett linje, $\hat{y} = a + bx$, der jeg skriver $\hat{y}$ istedenfor *y*-en for ikke å blande sammen med *y*-en som observeres. Med andre ord, for hvert observasjonspunkt, (x_i, y_i), kan vi skrive

$$y_i = a + bx_i + d_i = \hat{y}_i + d_i \quad \text{for } i = 1, 2, \ldots, n$$

der $\hat{y}_i = a + bx_i$ representerer den "forklarte" delen av y_i, og $d_i = y_i - \hat{y}_i$ den "uforklarte" delen av y_i. (Merk at vi alltid kan skrive $y_i = \hat{y}_i + y_i - \hat{y}_i = \hat{y}_i + d_i$.) Se figur 3.

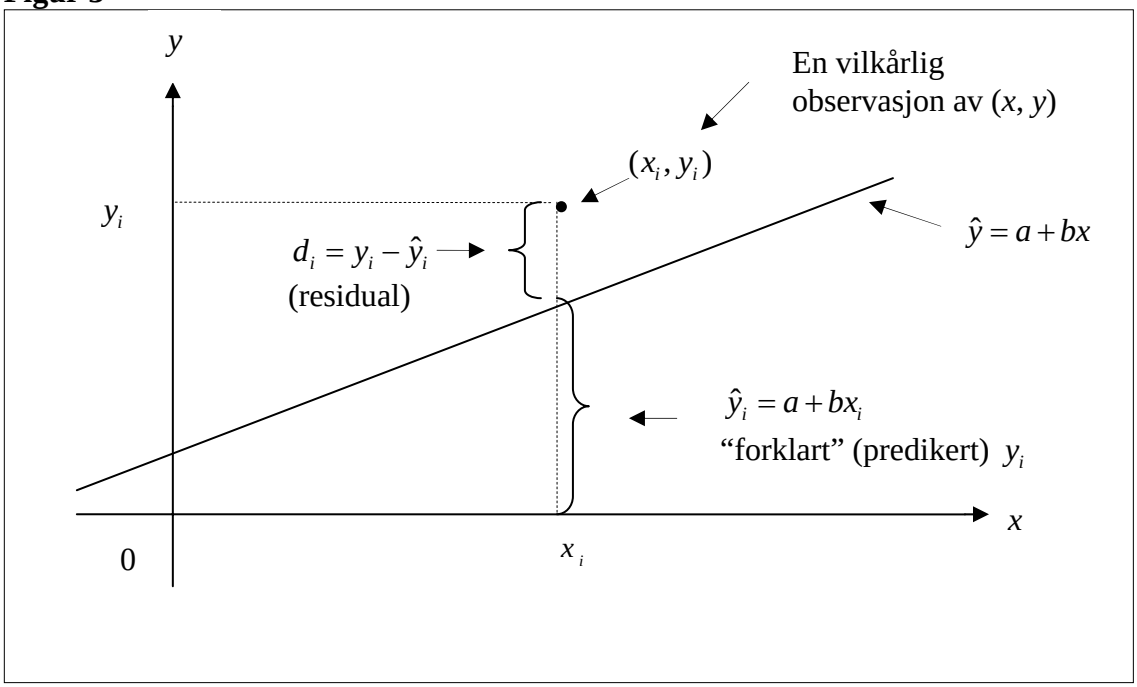
Jeg bruker anførselstegn rundt det suggestive uttrykket "forklare" siden "forklaring" egentlig er et for sterkt uttrykk i denne sammenhengen. Et mer vanlig uttrykk i litteraturen er "*predikert* y_i" for $\hat{y}_i$. Den "uforklarte" delen av y_i, d_i kalles oftest for "*residual*".

Oppgaven er nå å velge linja $\hat{y} = a + bx$ - dvs. å velge koeffisientene *a* og *b* – slik at "forklaringen" blir best mulig. Eller, sagt på en annen måte, slik at residualene, d_i, som måler de loddrette avstandene til linja fra observasjonspunktene, minimeres i en eller annen forstand. Vi ser av figur 3 at for punkter som ligger over linja ($y_i > \hat{y}_i$), så blir $d_i > 0$, mens

[3] Merk at jeg bruker små bokstaver for *x* og *y* for å indikere at disse kan være hvilke som helst variable – ikke nødvendigvis stokastiske. For eksempel, *x* kunne være en innsatsfaktor i en produktfunksjon med verdier valgt av forskeren, mens *y* er størrelsen på produktet. I så fall er kun *y* å betrakte som stokastisk.

$d_i < 0$ for punkter som ligger under linja ($y_i < \hat{y}_i$). Det nytter altså ikke å minimere summen av avstandene til linja, $\sum_i d_i$, siden de negative avstandene vil oppheve de positive. I stedet velger man å se på de kvadrerte avstandene, d_i^2 , som fjerner fortegnene. (Man kunne naturligvis også se på absoluttverdiene, $|d_i|$, men det gir en vesentlig mer komplisert løsning.)

Figur 3



Definisjon 1

En minste kvadraters regresjonslinje[4] (MKV), $\hat{y} = a + bx$, for *y* med hensyn på *x* for dataene, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, bestemmes slik at

$$Q = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

blir minst mulig.

[4] Uttrykket "regresjon" har historisk opprinnelse og ingen relevant betydning for eksemplene våre. Uttrykket ble først brukt i visse anvendelser i genetikk og har blitt hengende ved siden.

Dette er et veldefinert minimeringsproblem som har en entydig løsning (se **regel 2**). Det er flere måter å finne minimum av $Q = Q(a,b)$ på. Den vanligste er å sette de deriverte av *Q* med hensyn på *a* og *b* lik null siden de deriverte av *Q* må være null i minimumspunktet. Da får vi to ligninger til å bestemme *a* og *b* (husk at den deriverte til en sum er lik summen av de deriverte):

$$\frac{\partial Q}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\frac{\partial}{\partial a}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}(-2)(y_i - a - bx_i) = (-2)\sum_{i=1}^{n} d_i$$

$$\frac{\partial Q}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\frac{\partial}{\partial b}(y_i - a - bx_i)^2 = \sum_{i=1}^{n}(-2x_i)(y_i - a - bx_i) = (-2)\sum_{i=1}^{n} x_i d_i$$

Dermed, ved å sette de to deriverte lik 0, får vi to ligninger til å bestemme *a* og *b*:

(1) $\sum_{i=1}^{n} d_i = 0$ (eller $\sum_{i=1}^{n}(y_i - a - bx_i) = 0$)

(2) $\sum_{i=1}^{n} x_i d_i = 0$ (eller $\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$)

Løsningen er gitt ved regel 2 (se appendiks for detaljer):

Regel 2

Mkv regresjonslinje for y med hensyn på x for data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, er gitt ved linja $\hat{y} = a + bx$, der

$$a = \overline{y} - b\overline{x} \quad \text{og} \quad b = \frac{s_{xy}}{s_x^2}$$

Algebraen for å finne løsningen er ikke spesielt vanskelig, men er ikke pensum å kunne beherske. Den er derfor skrevet ut i appendiks som frivillig lesning for interesserte studenter. Finnes også i Løvås i appendiks B5 (side 450).

Regel 2 viser en interessant relasjon mellom stigningstallet, *b*, i mkv-linja og den empiriske korrelasjonskoeffisienten, *r*. Den forteller noe om *r* som vi tidligere bare har begrunnet intuitivt. Vi kan nemlig skrive

$$b = \frac{s_{xy}}{s_x^2} = \frac{s_{xy}}{s_x s_x}\cdot\frac{s_y}{s_y} = \frac{s_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x}$$

Sammenhengen mellom korrelasjonskoeffisienten, r, og b, er altså gitt ved

(3) $b = r\frac{s_y}{s_x}$

Dermed, siden s_x og s_y ikke kan være negative, må *b* og *r* ha samme fortegn. Dette betyr at en positiv korrelasjonskoeffisient, *r*, er det samme som at mkv regresjonslinja for *y* mhp *x* har positiv stigning. Hvis $r < 0$, må regresjonslinja helle nedover ($b < 0$), og hvis $r = 0$, er regresjonslinja flat ($b = 0$).

Regneeksempel fortsatt:
Siden vi har allerede har beregnet de fem grunnleggende størrelsene, $\overline{x}, \overline{y}, s_x^2, s_y^2, s_{xy}$, er det fort gjort å beregne mkv-linja:

$$b = \frac{s_{xy}}{s_x^2} = \frac{127.409}{999.1253} = 0.1275$$

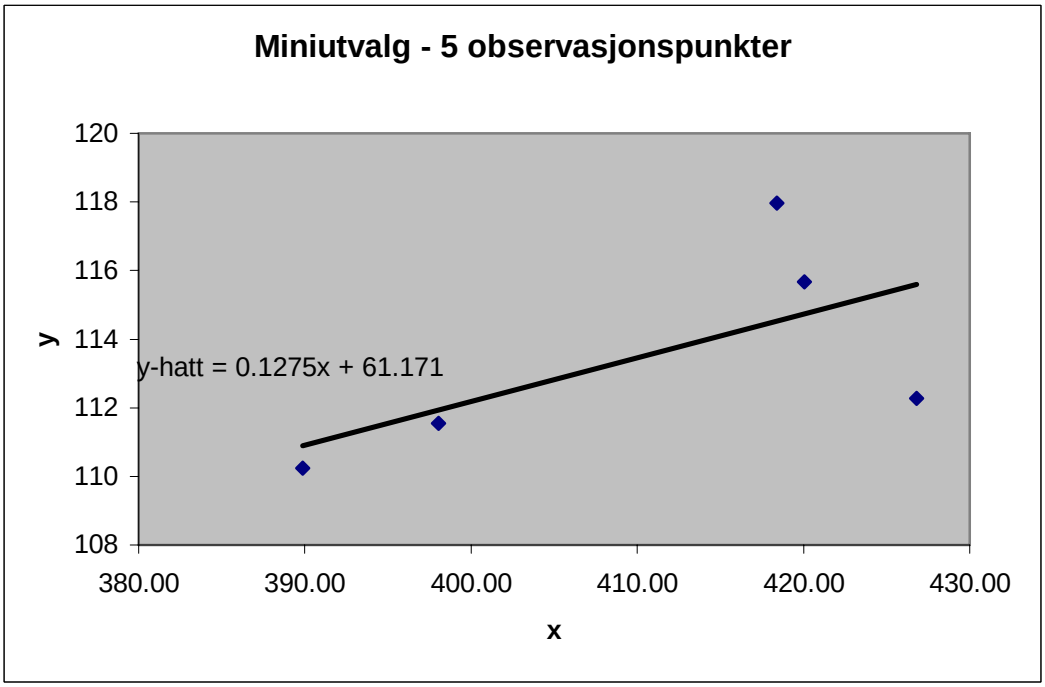
$$a = \overline{y} - b\overline{x} = 113.536 - (0.1275)(410.636) = 61.171$$

Den beregnete mkv-linja blir således

$$\hat{y} = 61.171 + (0.1275) \cdot x$$

I figur 4 har jeg latt Excel tegne inn mkv-linja in spredningsdiagrammet for miniutvalget.

Figur 4



Det kan også være instruktivt å se hvor mye av hver enkelt y_i blir "forklart" (predikert) av x_i. Vi kan nå beregne både $\hat{y}_i$ og d_i, som er vist i tabell 3:

Tabell 3

Obs. nr.	x_i	y_i	Predikert y $\hat{y}_i = a + bx_i$	Residual $d_i = y_i - \hat{y}_i$
1	398.06	111.54	111.92	-0.38
2	389.88	110.24	110.88	-0.64
3	426.81	112.27	115.59	-3.32
4	418.39	117.96	114.52	3.44
5	420.04	115.67	114.73	0.94

Vi ser at observasjon 1,2 og 5 synes rimelig bra "forklart", mens observasjon 3 og 4 er dårligere "forklart".

Vi forlater regneeksempelet og stiller det naturlige spørsmålet om en fornuftig måte å lage et samlet mål på hvor mye *x* "forklarer" av y i datamaterialet. Vi tar da utgangspunkt i et uttrykk for total variasjon av y_i-ene i materialet og spør hvor stor andel av denne totale variasjonen er "forklart" av x_i-ene via de predikerte y_i-ene.

Som uttrykk for total variasjon i y_i-ene brukes nå kvadratsummen

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

der notasjonen SS_T står for det engelske "sum of squares total" og er veldig vanlig i statistisk og økonometrisk litteratur. Merk at $SS_T = (n-1)s_y^2$ og inneholder samme informasjon som variansmålet s_y^2 (siden *n* er et fast tall for et gitt datamateriale). Tilsvarende definerer vi (total)variasjonen av den "forklarte" (predikerte) delen av y_i, nemlig $\hat{y}_i$:

$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{\hat{y}})^2$ "sum of squares of regression" måler totalvariasjonen av den "forklarte" (predikerte) delen. Kalles ofte for "forklart variasjon".

Tilsvarende definerer vi en kvadratsum for variasjonen av den "uforklarte" delen, d_i.

$SS_E = \sum_{i=1}^{n} (d_i - \bar{d})^2 = \sum_{i=1}^{n} d_i^2$ (siden (1) impliserer at $\bar{d} = 0$). Står for "sum of squares of error" som måler totalvariasjonen av "den uforklarte delen" (residualene).

Disse variasjonsmålene er bundet sammen ved følgende fundamentale setning (regel 3) bevist i appendiks:

Regel 3

(a) $SS_T = SS_R + SS_E$

Regneformler:

(b) $SS_T = (n-1)s_y^2$

(c) $SS_E = Q_{\min} = \sum_{i=1}^{n} d_i^2 = (n-1)s_y^2(1-r^2)$ (der $Q_{\min}$ er minmumsverdien for Q)

Vårt mål på andelen av den totale variasjonen i y (i data) "forklart" av x via mkv-linja blir nå

Definisjon 2 Mål på forklart variasjon av y i data: $\frac{SS_R}{SS_T}$

som i følge regel 3(a) må være et tall mellom 0 og 1. Ved å bruke regneformlene i regel 3, får vi

$$\frac{SS_R}{SS_T} = \frac{SS_T - SS_E}{SS_T} = 1 - \frac{SS_E}{SS_T} = 1 - \frac{(n-1)s_y^2(1-r^2)}{(n-1)s_y^2} = 1-(1-r^2) = r^2$$

Vi har dermed bevist

Regel 4

Andel forklart variasjon i y av mkv-linja, $\hat{y} = a + bx$, er gitt ved

$$\frac{SS_R}{SS_T} = r^2$$

der r er (den empiriske) korrelasjonskoeffisienten for data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$

Merk at regel 4 gir en ny tolkning av den empiriske korrelasjonskoeffisienten r mellom to variable x og y– nemlig at r^2 kan tolkes som et mål på hvor mye av variasjonen av y i data blir forklart av x (i data) om vi prøver å beskrive mest mulig av y_i-ene ved en rett linje, $\hat{y} = a + bx$. Om vi omvendt forsøker å forklare x ved y i data ved en mkv regresjonslinje, $\hat{x} = a_1 + b_1 y$, vil vi på grunn av symmetrien i r få samme svar: $100r^2\%$ av variasjonen i x_i-ene blir forklart ved den rette linja $\hat{x} = a_1 + b_1 y$, der $b_1 = s_{xy}/s_y^2 = r \cdot (s_x / s_y)$ og $a_1 = \overline{x} - b_1\overline{y}$ (ved formlene ovenfor der x-ene og y-ene bytter plass).

Merk at r kan alltid regnes ut når vi har data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, men kan ikke alltid tolkes som en korrelasjonskoeffisient – for eksempel i en situasjon der x_i-ene er gitte tall (som valgte verdier av en innsatsfaktor) valgt ut av en forsøksleder, mens y_i-ene er tilsvarende verdier av en respons (output). I en slik situasjonen har verdien vi får ved å regne

ut *r* ingen naturlig tolkning som en korrelasjonskoeffisient. Men tolkningen av r^2 i regel 4 er fortsatt meningsfull. Det gir nemlig ofte god mening å forsøke å forklare noen observerte verdier av produktet, *y*, ved noen utvalgte verdier av innsatsfaktoren, *x*, ved en rett linje – spesielt i situasjoner der de valgte x_i -ene ikke varierer for mye, som er under forsøksleders kontroll.

Regneeksempel fortsatt. I miniutvalget på 5 observasjonspunkter fant vi en korrelasjonskoeffisient på $r = 0.633$ ($\Rightarrow r^2 = 0.400689$) slik at 40% av variasjonen i 1500m–tidene i miniutvalget blir forklart av 5000m–tidene via mkv regresjonslinje. Vi ville ha fått samme svar (40%) om vi omvendt hadde forsøkt å forklare 5000m-tidene ved 1500m-tidene.

I det opprinnelige utvalget på 27 observasjonspunkter (uten Ervik) ble $r = 0.676$ ($r^2 = 0.456976$), slik 45.7% av variasjonen i 1500m-tidene i data blir forklart av 5000m-tidene.

Noen generelle egenskaper ved den empiriske korrelasjonskoeffisienten.
Til slutt er det verdt å nevne at noen generelle egenskaper ved den empiriske korrelasjonskoeffisienten, *r*, basert på data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, som ble nevnt i begynnelsen av avsnitt 2, følger direkte av regel 3:

Siden $SS_E = \sum_{i=1}^{n} d_i^2$ ikke kan være negativ, følger av regel 3(c) at $1 - r^2 \geq 0$, dvs. $r^2 \leq 1$ må være oppfylt. Vi ser også av samme ligning at ekstremtilfellet, $r^2 = 1$ (dvs. $r = \pm 1$) impliserer at $SS_E = \sum_{i=1}^{n} d_i^2 = 0$. Siden alle leddene i summen er ikke-negative, er dette kun mulig hvis alle $d_i = 0, \quad i = 1, 2, \ldots, n$ er oppfylt. Siden $d_i = y_i - a - bx_i$, må i så fall alle (x_i, y_i) oppfylle, $y_i = a + bx_i, \quad i = 1, 2, \ldots, n$ - dvs. alle observasjonspunktene ligger (eksakt) på en rett linje.

4 Noen ord om tolkning av resultatene i avsnitt 2 og 3.

Vi ser at vi nå har to forskjellige måter å uttrykke graden av lineær sammenheng mellom to variable, *x* og *y*, basert på data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$:

(i) Beregne (den empiriske) korrelasjonskoeffisienten, *r*.

(ii) Beregne mkv-regresjonslinje av y med hensyn på x, basert på data, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Spørsmål 1: Så hva er den beste måten? Svaret på det avhenger av problemstillingen. Merk at *r* er helt symmetrisk bygget opp og gir samme svar uansett om vi ønsker å forklare *y* ved hjelp av *x* eller, omvendt, om vi ønsker å forklare *x* ved hjelp av *y*. Metode (i) er fornuftig i noen (symmetriske) situasjoner. Metode (ii), derimot, er mer skreddersydd for den (kanskje mer vanlige) asymmetriske situasjonen der vi har en avhengig variabel (responsvariabel), *y*, vi ønsker å forklare ved hjelp av en forklaringsvariabel, *x*, (jamfør, for eksempel, den andre

merknaden etter regel 4). Under (ii) vil sammenhengen faktisk se forskjellig ut, avhengig av om vi prøver å forklare *y* ved hjelp av *x*, eller om vi prøver å forklare x ved hjelp av *y*.

I regneeksemplet prøvde vi (i miniutvalget) å forklare y (1500m-tid) ved x (5000m-tid), kanskje motivert ut fra at 5000m-løpet går dagen før 1500m-løpet for menn. Analysen resulterte i mkv-regresjonslinje,

(a) $\hat{y} = 61.171 + (0.1275) \cdot x$

På den annen side, er det ikke noe som hindrer oss i å være interessert i det omvendte problemet, nemlig å forklare 5000m-tiden ved prestasjonen på 1500m. Vi kan fort beregne mkv-regresjonslinje for *x* med hensyn på *y* ved simpelthen å bytte om x og y i alle formlene (merk at r i så fall ikke endrer seg (hvorfor)?). Vi har alt vi trenger i tallene under tabell 2 i avsnitt 2, og får (sjekk selv!) mkv-linja for x med hensyn på y:

(b) $\hat{x} = 54.1328 + (3.140)\, y$

Merk at stigningstallet i (b) ikke er lik 1/(stigningstallet i (a)), som ville vært tilfelle dersom alle observasjonspunktene hadde ligget eksakt på en rett linje. De to svarene er altså forskjellige (når $|r| < 1$) og således avhengig av problemstillingen.

Spørsmål 2: Hvordan tolkes resultatene ovenfor i forhold til populasjonen data er hentet (trukket) fra?

Svaret avhenger sterkt av vår statistiske modell for populasjonen - dvs. av hvilke forutsetninger vi er villige til å postulere om populasjonen. Hvis vi ikke er villig til å forutsette noe om populasjonen, vil analysen ovenfor være helt tom – dvs. ikke si noe som helst om populasjonen vi er interessert i. Med andre ord, uten noen ideer om populasjonen har vi ikke noe grunnlag for å tolke resultatene ovenfor utover datamaterialet selv.

Så hva er populasjonen her? Dette er ikke noe trivielt spørsmål. Spesielt spørsmålet om avgrensning kan være problematisk. Hvilke skøyteløpere skal være representert i populasjonen? Bare de 27 som deltok ved EM i Heerenveen i 2004, eller andre også? Den vanlige måten å nærme seg dette problemet på, er i første omgang å hoppe bukk over problemet med avgrensning og rett og slett erstatte populasjonen med en statistisk modell. En statistisk modell for en slik populasjon omfatter grovt sett to ting, (1) en liste over hvilke stokastiske (og andre typer) variable som våre data antas å være observasjoner av, og (2) et sett av forutsetninger vi er villig til å gjøre om sannsynlighetsfordelingene for de stokastiske variablene som inngår.

I det spesielle tilfellet at $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ kan betraktes som uavhengige og rent tilfeldige observasjoner av et stokastisk variabel-par (X, Y), som diskutert i første del av avsnitt 2, vil populasjonen være representert ved de stokastiske variablene *X* og *Y* sammen med de forutsetningene vi er villig til å forutsette om sannsynlighetsfordelingen til (X, Y). Er dette en rimelig modell for vår situasjon? Tvilsomt![5] For det første er det ikke rimelig å anta

[5] For miniutvalget på 5 vi brukte som regneeksempel, kunne dette opplegget passe. Populasjonen vi trekker fra er da gitt ved de opprinnelige 27 observasjonspunktene, og utvalget er trukket representativt – dvs. slik at alle mulige utvalg på 5 fra de 27 er like sannsynlige.

at skøyteløperne er tilfeldig trukket ut – de er vel snarere behendig valgt ut av forskjellige lands skøyteforbund. For det andre virker det tvilsomt å anta at alle tidene oppnådd i data stammer fra samme fordeling i og med at det er store individuelle forskjeller i yteevne mellom de forskjellige skøyteløperne. Et alternativ som tar hensyn til disse innvendingene kan være den enkle regresjonsmodellen som Løvås setter opp i avsnitt 7.3.1. I den modellen erstattes *Y* med *n* stokastiske variable, $Y_1, Y_2, \ldots, Y_n$, en for hver skøyteløper, og observasjonen y_i oppfattes som en observasjon av Y_i. På den annen side oppfattes x_i-ene (5000m-tidene) som gitte tall (som kan være rimelig her siden vi ønsker å forklare Y_i-ene ved x_i-ene). Sammenhengen uttrykkes ved å postulere (der α kalles "alfa" og β kalles "beta")

$$Y_i = \alpha + \beta x_i + e_i \quad \text{for } i = 1, 2, \ldots, n$$

der $e_1, e_2, \ldots, e_n$ antas å være stokastisk uavhengige restledd som alle forutsettes å ha forventning , $E(e_i) = 0$, og samme varians, $\text{var}(e_i) = \sigma^2$. Dette impliserer at Y_i har forventning $E(Y_i) = \alpha + \beta x_i + E(e_i) = \alpha + \beta x_i$ som altså varierer med 5000m-tiden x_i. Vi ser dermed at fordelingen til Y_i varierer med *i* siden forventningen varierer med *i*. Y_i er altså en stokastisk variabel som måler 1500m-tiden for en løper som dagen før har oppnådd en 5000m-tid på x_i. Sannsynlighetsfordelingen for Y_i antas å være et uttrykk for utallige tilfeldigheter som kan spille inn ved et slikt 1500m-løp som dagsform, isens kvalitet, osv.

Denne modellen, som altså antas å representere populasjonen dataene er trukket fra, inneholder tre "ukjente" parametre, α, β og σ^2 i den forstand at deres sanne verdier er ukjente.

Modellen er den enkleste varianten av en regresjonsmodell, og den eneste som behandles i dette kurset.

Det viser seg (jfr. kapittel 7 i Løvås) at hvis denne modellen kan anses som en akseptabel beskrivelse av populasjonen dataene, $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, er trukket fra, så vil mkv-regresjonslinjen ovenfor vise seg være det beste estimatet for (den ukjente) regresjonslinja, $y = \alpha + \beta x$ i populasjonen, og vi kan tolke dataene i forhold til denne modellen, ved bruk av teorien i kapittel 6 og 7. Vi vil komme tilbake til dette i et senere notat når relevant teori fra kapittel 6 er etablert.

Øvelse i Excel

Du trenger modulen "Data analysis" for å løse oppgaven. Sjekk at Data analysis ligger på data-menyen. Hvis ikke, må den legges til ("add in"): I siste Excel versjon: Start fra "office button" (en sirkel øverst til venstre på Excel-arket). Klikk så på "excel options" helt nederst på menyen som kommer fram. Og videre:

office button → excel options → add-ins → marker "Analysis toolpack" → Klikk "Go.." → → merk av "Analysis toolpack" → klikk OK.

(I eldre Excel: Fra menyen: tools → add-ins → merk av "Analysis toolpack" → klikk OK.)

1) Last ned skøytedataene fra http://folk.uio.no/haraldg/.

2) Lag to kollonner ved siden av hverandre x-ene (5000m) og y-ene (1500m), målt i sekunder (uten Eskil Ervik).

3) Beregn $\bar{x}, \bar{y}, s_x^2, s_y^2, s_{xy}$ [Fås fra "Descriptive Statistics" eller fra "Covariance" [6], begge rutiner i "Data analysis"]

4) Beregn a, b, SS_T, SS_R, SS_E fra de fem verdiene i 3).

5) Kjør Excels regresjonsrutine ("regression" i "Data analysis") og identifiser a, b, SS_T, SS_R, SS_E i utskriften.

6) Reproduser figur 2 (i avsnitt 1)

[6] En liten modifikasjon: "Covariance" gir ikke eksakt s_x^2, s_y^2, s_{xy}, men s_x^2, s_y^2, s_{xy} multiplisert med $(n-1)/n$ som innebærer at Excel der deler summene på *n* istedenfor $n-1$. For å få fram "våre" s_x^2, s_y^2, s_{xy}, bør man altså multiplisere tallene i "Covariance" med $n/(n-1)$. s_x^2 og s_y^2 fra "Descriptive Statistics" derimot er definert som hos oss.

Appendiks (frivillig lesning)

Det kreves ikke i dette kurset at man behersker algebraen bak formlene ovenfor. Det som kreves er at man kan bruke formlene og stort sett skjønner hva de står for. På den annen side er ikke algebraen verre enn at studenter på forutsatt nivå for Econ 2130 bør kunne følge med på den, og noen vil (forhåpentligvis) være nysgjerrige over å vite hvordan formlene har oppstått. Uansett slipper man ikke unna en grundigere trening i denne algebraen i senere (økonometri-) kurs.

Algebraen er preget av summer. For dem som føler seg utrygge med summe-manipulasjoner har jeg oppsummert de viktigste reglene i avsnitt **A1**.

A1 Noen regler for summer

(**a**) Hvis *uttrykk* i summen $\sum_{i=1}^{n} uttrykk$ selv er en sum bestående av flere ledd, gjelder summetegnet kun for det første leddet i *uttrykk* – altså til første pluss eller minus dukker opp i uttrykket. Hvis man ønsker at summetegnet skal omfatte mer enn bare første ledd, må man bruke parentes.

[**Eksempel:** $\sum_{i=1}^{n} 2x_i - 3$ betyr $(2x_1 + 2x_2 + \cdots + 2x_n) - 3$. Med andre ord, -3 hører ikke med under summetegnet. Om man ønsker at -3 skal være med under summetegnet, må man bruke parentes:

$$\sum_{i=1}^{n}(2x_i - 3) = 2x_1 - 3 + 2x_2 - 3 + \cdots + 2x_n - 3 = (2x_1 + 2x_2 + \cdots + 2x_n) - 3n \ ]$$

(**b**) Hvis c er en konstant, er $\sum_{i=1}^{n} c = nc$

$$[\ \sum_{i=1}^{n} c = c + c + \cdots + c = nc\ ]$$

(**c**) En felles faktor i en sum kan settes utenfor summen: $\sum_{i=1}^{n} cx_i = c\sum_{i=1}^{n} x_i$

$$[\ cx_1 + cx_2 + \cdots + cx_n = c \cdot (x_1 + x_2 + \cdots + x_n)\ ]$$

(**d**) Hvis a, b, c, d er konstanter, gjelder

$$\sum_{i=1}^{n}(a + bx_i + cy_i + dz_i) = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} y_i + d\sum_{i=1}^{n} z_i$$

[Ses ved å skrive ut summen til venstre og ordne om på leddene. Vi vet jo at endring av rekkefølgen av leddene i en sum ikke endrer summen, som for eksempel, $3 - 5 = 3 + (-5) = -5 + 3$.]

A2 Bevis for regel 1 – 3

Regel 1:

Det er nok å vise (c) $(n-1)s_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - n\,\overline{x}\,\overline{y}$, siden (a) og (b) følger av (c) ved å sette $x_i = y_i$ for alle *i* i (c): Vi bruker **A1** og får

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n}(x_i y_i - x_i \cdot \overline{y} - \overline{x} \cdot y_i + \overline{x} \cdot \overline{y}) \overset{A1(d)}{=} \sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i - \overline{x}\sum_{i=1}^{n} y_i + n\overline{x} \cdot \overline{y}$$

Nå er $\sum_{i=1}^{n} x_i = n\overline{x}$ (av definisjonen av gjennomsnitt) og $\sum_{i=1}^{n} y_i = n\overline{y}$, og vi får

$$\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \overline{y} \cdot n\overline{x} - \overline{x} \cdot n\overline{y} + n\overline{x} \cdot \overline{y} = \sum_{i=1}^{n} x_i y_i - n\overline{y} \cdot \overline{x}$$

Bevis slutt.

Regel 2:
Oppgaven er å løse ligningene (1) og (2) med hensyn på *a* og *b*, der

(1) $\sum_{i=1}^{n} d_i = 0$ (eller $\sum_{i=1}^{n}(y_i - a - bx_i) = 0$)

(2) $\sum_{i=1}^{n} x_i d_i = 0$ (eller $\sum_{i=1}^{n} x_i(y_i - a - bx_i) = 0$)

Av (1) og **A1(d)** får vi

$$0 = \sum_{i=1}^{n}(y_i - a - bx_i) \overset{A1(d)}{=} \sum_{i=1}^{n} y_i - na - b\sum_{i=1}^{n} x_i = n\overline{y} - na - nb\overline{x}$$

Deler vi med *n* på begge sider, får vi $0 = \overline{y} - a - b\overline{x}$, som gir

(4) $a = \overline{y} - b\overline{x}$

Av (2) og **A1(d)** får vi

(5) $0 = \sum_{i=1}^{n} x_i(y_i - a - bx_i) = \sum_{i=1}^{n}(x_i y_i - ax_i - bx_i^2) \overset{A1(d)}{=} \sum_{i=1}^{n} x_i y_i - a\sum_{i=1}^{n} x_i - b\sum_{i=1}^{n} x_i^2$

Av regel 1 finner vi $\sum_{i=1}^{n} x_i y_i = (n-1)s_{xy} + n\bar{x}\bar{y}$ og $\sum_{i=1}^{n} x_i^2 = (n-1)s_x^2 + n\bar{x}^2$. Ved i tillegg utnytte at $\sum_{i=1}^{n} x_i = n\bar{x}$, får vi ved innsetting i (5)

$$
\begin{aligned}
0 &= (n-1)s_{xy} + n\bar{x}\bar{y} - an\bar{x} - b\left((n-1)s_x^2 + n\bar{x}^2\right) \overset{(4)}{=} \\
&= (n-1)s_{xy} + n\bar{x}\bar{y} - \left(\bar{y} - b\bar{x}\right)n\bar{x} - b\left((n-1)s_x^2 + n\bar{x}^2\right) = \\
&= (n-1)s_{xy} + n\bar{x}\bar{y} - n\bar{x}\bar{y} + nb\bar{x}^2 - b(n-1)s_x^2 - nb\bar{x}^2 = (n-1)s_{xy} - b(n-1)s_x^2 = \\
&= (n-1)(s_{xy} - bs_x^2)
\end{aligned}
$$

som viser at

(6) $b = \dfrac{s_{xy}}{s_x^2}$ er løsningen.

Strengt tatt er det nødvendig å vise at løsningen gitt ved (4) og (6) faktisk bestemmer et minimum for *Q*. Jeg hopper over den delen og henviser til Sydsæter for dette (for å slippe å bringe inn matrisen av annenderiverte av *Q* og dens determinant. Om man likevel vil gjøre det, er det ikke vanskelig å finne de annen-deriverte og se at både determinanten og hoveddiagonalelementene er positive, som er en tilstrekkelig betingelse for minimum her.) **Bevis slutt.**

Regel 3:

(a) $SS_T = SS_R + SS_E$

Resultatet følger igjen nokså direkte av relasjonene (1) og (2). Det første vi merker oss, er at y_i-ene og $\hat{y}_i$-ene må ha samme gjennomsnitt, $\bar{\hat{y}} = \bar{y}$. Dette følger av (1) siden $0 = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{y}_i$, eller $\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i$. Deler vi begge sider på *n*, får vi $\bar{\hat{y}} = \bar{y}$.

Dermed kan vi forenkle SS_R til

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 .$$

Ved å legge til og trekke fra $\hat{y}_i$, får vi for alle *i*

$$y_i - \bar{y} = \hat{y}_i - \bar{y} + y_i - \hat{y}_i = \hat{y}_i - \bar{y} + d_i$$

som gir

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + d_i^2 + 2(\hat{y}_i - \bar{y})d_i$$

Ved å ta summen av begge sider får vi (jfr **A1**)

$$SS_T = SS_R + SS_E + 2\sum_{i=1}^{n}(\hat{y}_i - \bar{y})d_i$$

Beviset vil være fullført om vi kan vise at den siste summen er null. Dette følger av (1) og (2) ved:

$$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})d_i \overset{A1(d)}{=} \sum_{i=1}^{n}\hat{y}_i d_i - \bar{y}\sum_{i=1}^{n} d_i \overset{(1)}{=} \sum_{i=1}^{n}\hat{y}_i d_i = \sum_{i=1}^{n}(a+bx_i)d_i \overset{A1(d)}{=} \sum_{i=1}^{n} ad_i + \sum_{i=1}^{n} bx_i d_i =$$

$$= a\sum_{i=1}^{n} d_i + b\sum_{i=1}^{n} x_i d_i \overset{(1),(2)}{=} 0$$

og beviset for (a) er fullført.

(b) følger av definisjonen på s_y^2.

Bevis for (c):

Ved å legge til og trekke fra $\bar{y}$ i d_i og sette inn for *a* fra (4), får vi for hver *i*

$$d_i = y_i - a - bx_i = y_i - \bar{y} + \bar{y} - (\bar{y} - b\bar{x}) - bx_i = y_i - \bar{y} + b\bar{x} - bx_i =$$
$$= y_i - \bar{y} - b(x_i - \bar{x})$$

Dermed

$$Q_{\min} = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n}\left(y_i - \bar{y} - b(x_i - \bar{x})\right)^2 =$$
$$\overset{A1}{=} \sum_{i=1}^{n}(y_i - \bar{y})^2 + b^2\sum_{i=1}^{n}(x_i - \bar{x})^2 - 2b\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) =$$
$$= (n-1)s_y^2 + b^2(n-1)s_x^2 - 2b(n-1)s_{xy} =$$
$$\overset{(6)}{=} (n-1)\left[s_y^2 + \frac{s_{xy}^2}{s_x^4}s_x^2 - 2\frac{s_{xy}}{s_x^2}s_{xy}\right] = (n-1)\left[s_y^2 + \frac{s_{xy}^2}{s_x^2} - 2\frac{s_{xy}^2}{s_x^2}\right] =$$
$$= (n-1)\left[s_y^2 - \frac{s_{xy}^2}{s_x^2}\right] = (n-1)\left[s_y^2 - \frac{s_{xy}^2}{s_x^2 s_y^2}s_y^2\right] = (n-1)\left[s_y^2 - r^2 s_y^2\right]$$

Altså $SS_E = Q_{\min} = (n-1)s_y^2(1-r^2)$

Bevis slutt.