ECON 2130
HG, april 2010

Notat til kapittel 7 i Løvås

Om enkel lineær regresjon II

Merk: Det kan lønne seg først å lese avsnitt 4 i regresjon-I-notatet på nytt.

Regresjonsmodellen.

La Y være en stokastisk variabel (som vi kaller responsvariabelen) og x en forklaringsvariabel som oppfattes som ikke-stokastisk. Vi tenker oss da at paret (x, Y) observeres sammen slik at Y kun observeres sammen med gitte x-verdier. I eksemplet fra regresjons-I-notatet med skøytetider fra EM i Heerenveen 2004, kunne x representere 5000m-tiden (i sekunder), og Y representere 1500m-tiden (i sekunder) en vilkårlig valgt skøyteløper ville oppnå dersom vedkommende løp 1500m dagen etterpå. Denne skøyteløperen er ikke nødvendigvis en av dem som deltok i Heereveen i 2004, men tenkes tilfeldig valgt fra populasjonen av skøyteløpere mer generelt i 2004 som var i stand til å løpe 5000m på x sekunder. (Jfr. litt diskusjon av populasjonsbegrepet i avsnitt 4 i regr.-I-notatet.)

I den enkle (og eneste) regresjonsmodellen i pensum antas Y normalfordelt, og sammenhengen mellom x og Y uttrykkes via forventningen til Y[1]

(1) $E(Y) = \mu(x) = \alpha + \beta x$ for vilkårlig x i et passende definisjonsintervall.

I tillegg antas at variansen til Y for gitt x-verdi er konstant uansett hva x er:

(2) $\text{var}(Y) = \sigma^2$ uansett hva x er.

Merk at vi kan skrive denne modellen kort som $Y \sim N(\mu(x), \sigma) = N(\alpha + \beta x,\ \sigma)$ for x i definisjonsintervallet.

α kalles for *konstantleddet*, β for *regresjonskoeffisienten* og $\mu(x)$ for *regresjonsfunksjonen*. β tolkes ofte som "effekten av en enhets endring i x på Y".

> [Begrunnelse: La x_1 og $x_1 + 1$ være to x-verdier med en enhets forskjell. La Y' og Y'' være tilsvarende responser. Forventet endring i Y blir da
> $E(Y'' - Y') = E(Y'') - E(Y') = \mu(x_1 + 1) - \mu(x_1) = \alpha + \beta(x_1 + 1) - (\alpha + \beta x_1) = \beta$]

[1] Symbolkollisjon: Merk at konstantleddet α i (1) ikke må forveksles med den α som dukker opp i konfidensgrader og signifikansnivå hos Løvås.

(3) NB! *Merk at hvis den "sanne" verdien* $\beta = 0$ *i denne modellen, vil leddet* βx *forsvinne fra* $\mu(x)$ *og fordelingen for Y være den samme uansett hva x er. I så fall er regresjonsfunksjonen flat, og det er ingen sammenheng mellom x og Y. Modellen omfatter altså muligheten av at det ikke er noen sammenheng mellom x og Y.*

En alternativ og ekvivalent formulering av modellen er

(4) $Y = \alpha + \beta x + e$

der *x* er en gitt ikke-stokastisk størrelse og *e* er en stokastisk variabel (kalt *feilledd* eller *restledd*[2]) som er normalfordelt med forventning 0 og varians σ^2, dvs. $e \sim N(0, \sigma)$.

Vanligvis vil modellen referere til en eller annen populasjon av interesse (jfr. skøyteeksemplet). I relasjon til en slik populasjon (som ikke kan observeres i sin helhet) vil parametrene, α, β og σ og således også regresjonsfunksjonen $\mu(x) = \alpha + \beta x$, som oftest stå for *ukjente* populasjonsstørrelser som vi er interessert i å si noe om basert på data trukket fra populasjonen.

Modell for data (enkel univariat regresjonsmodell).
Data består av gjentatte observasjoner av (x, Y), $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Uansett hvordan x_i-ene er oppstått (stokastisk eller valgte tall), oppfatter vi dem i modellen vår som faste tall som om de skulle vært valgt på forhånd. $y_1, y_2, \ldots, y_n$, derimot, oppfatter vi som observasjoner av stokastiske variable, $Y_1, Y_2, \ldots, Y_n$, som oppfyller:

(5) $Y_i = \mu(x_i) = \alpha + \beta x_i + e_i$ for $i = 1, 2, \ldots, n$

der $e_1, e_2, \ldots, e_n$ er uavhengige og identisk normalfordelte stokastiske variable kalt *restledd* (Løvås: "feilledd") med forventning 0 og varians σ^2 (kort $e_i \sim N(0, \sigma)$, $i = 1, 2, \ldots, n$). Denne modellen er ekvivalent med å si at $Y_1, Y_2, \ldots, Y_n$ er uavhengige og normalfordelte med forventning, $E(Y_i) = \mu(x_i) = \alpha + \beta x_i$ og konstant varians, $\text{var}(Y_i) = \sigma^2$ (hvorfor?).

Parametrene α, β, σ tolkes nå nettopp som de ukjente populasjonsstørrelser vi er interessert i å si noe om (estimere eller teste hypoteser om) basert på data.

Merk at (5) omfatter både kjente (observerbare) og ukjente størrelser i og med at modellen knyttes til en populasjon som ikke kan observeres i sin helhet. Størrelser vi kan beregne ut fra data kalles i en slik sammenheng for "observerbare". Dette gjelder kun x_i-ene og Y_i-ene i (5). Det stokastiske restleddet, imidlertid, $e_i = Y_i - \alpha - \beta x_i$, er en ikke-observerbar stokastisk

[2] Og ikke *residual* som Løvås feilaktig kaller det øverst på side 272. Residual står for estimert (predikert) restledd som er noe annet, nemlig $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$ (se nedenfor).

variabel siden den avhenger av de ukjente populasjonsstørrelsene, α og β (selv om vi har observerte verdier for x_i, Y_i, vet vi ikke hvilken verdi den tilsvarende e_i har fått).

I tillegg til å estimere α, β, σ, vil vi også ofte være interessert i å estimere forventet Y ($\mu(x_0)$) for en eller annen utvalgt verdi $x = x_0$ av interesse (som ikke nødvendigvis må være blant x_i-ene i data). I så fall er vi interessert i $\mu(x_0) = \alpha + \beta x_0$. For eksempel kunne vi være interessert å estimere forventet tid (gjennomsnittlig tid i populasjonen) på 1500m for skøyteløpere som oppnår 6:30 min (390 sek) på 5000m, nemlig $\mu(390) = \alpha + \beta \cdot 390$.

Minste kvadraters estimatorer $\hat{\alpha}, \hat{\beta}$ for α og β, er definert som funksjoner av x_i-ene og Y_i-ene som minimerer kvadratsummen $Q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2$ med hensyn på α og β. Dette er akkurat det samme minimeringsproblemet som er gjennomført i regr.-I-notatet og løsningen er gitt i regel 2 i det notatet (side 8), der den eneste forskjellen er at de små y_i-ene (står for observerte tall) er byttet ut med de tilsvarende stokastiske Y_i-ene.

De fem størrelsene fra regr.-I-notatet som regresjonsbergeningen bygger på blir nå:

(6)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i, \quad s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2, \quad S_y^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \quad S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})$$

Her er $\bar{x}, s_x^2$ å oppfatte som konstanter, mens $\bar{Y}, S_y^2$ og S_{xy} er stokastiske variable.

Av formlene i regel 2 i regr.-I-notatet pluss litt kjedelig algebra basert på regel 4.12 og 4.17 i Løvås, kan vi formulere følgende regel

Regel 1

(a) Minste kvadraters (mkv) estimatorer for $\alpha, \beta, \mu(x_0)$ er gitt ved

$$\hat{\beta} = \frac{S_{xy}}{s_x^2}, \quad \hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x} \quad \text{og} \quad \hat{\mu}(x_0) = \hat{\alpha} + \hat{\beta} x_0$$

(b) $\hat{\alpha}, \hat{\beta}$ og $\hat{\mu}(x_0)$ er alle forventningsrette.

(c) Variansene er gitt ved

$$\text{var}(\hat{\beta}) = \frac{\sigma^2}{(n-1)s_x^2}, \quad \text{var}(\hat{\alpha}) = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right) \quad \text{og} \quad \text{var}(\hat{\mu}(x_0)) = \sigma^2\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}\right)$$

Merk at formlene som Løvås gir nederst på for variansene til $\hat{\alpha}$ og $\hat{\beta}$ er litt anderledes formulert. Sjekk selv at uttrykkene i Løvås må være lik uttrykkene i regel 1(c)

Siden mkv-estimatorene er forventningsrette, vil standardfeilene være lik standardavviket som vi ser avhenger av populasjons-standardavviket, σ. Vi trenger derfor å estimere denne.. Utgangspunktet er en estimator for σ^2:

Motivasjon. Om restleddet, $e_i = Y_i - \alpha - \beta x_i$, vet vi at $E(e_i) = 0$ og $\sigma^2 = \text{var}(e_i) = E(e_i^2) - \big(E(e_i)\big)^2 = E(e_i^2)$. Dermed (hvorfor?) $E\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) = \sigma^2$. Om vi hadde kunnet observere e_i-ene ut fra data (hvilket vi ikke kan), ville vi således kunne benytte $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$ som forventningsrett estimator for σ^2. $\tilde{\sigma}^2$ er imidlertid ubrukelig siden den ikke er observerbar. Men, siden vi kan estimere (predikere) det ikke-observerbare restleddet, e_i, ved den observerbare residualen, $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$, er det naturlig å prøve å erstatte e_i med $\hat{e}_i$ i $\tilde{\sigma}^2$. Litt algebra (som vi ikke tar her) viser nå at $E\left(\sum_{i=1}^{n} \hat{e}_i^2\right) = (n-2)\sigma^2$. Dermed (hvorfor?) får vi en forventningsrett estimator for σ^2 ved $\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2$ - som er den som benyttes. Vi trenger en regneformel for denne. På grunn av måten mkv-estimatorene er konstruert på, innser vi at residualen, $\hat{e}_i$, simpelthen er den stokastiske variabelen bak residualen, d_i, definert i regr.-I-notatet. Av side 10 i notatet får vi derfor $\sum_{i=1}^{n} \hat{e}_i^2 = SS_E$, der SS_E nå står for den stokastiske variabelen vi får ved å bytte ut alle små y_i-er med store Y_i-er. Ved å benytte regel 3 i regr.-I-notatet og sette inn for $r = S_{xy} \big/ (\sqrt{s_x^2 S_y^2})$, får vi

$$SS_E = (n-1)S_y^2\left(1 - \frac{S_{xy}^2}{s_x^2 S_y^2}\right) = (n-1)\left(S_y^2 - \frac{S_{xy}^2}{s_x^2}\right) = (n-1)\left(S_y^2 - \hat{\beta}^2 s_x^2\right)$$

Av dette får vi

Regel 2

(a) En forventningsrett estimator for σ^2 er gitt ved

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} \hat{e}_i^2 = \frac{n-1}{n-2}\left(S_y^2 - \hat{\beta}^2 s_x^2\right) \qquad \text{(kalt } s^2 \text{ i Løvås)}$$

der $\hat{e}_i = Y_i - \hat{\alpha} - \hat{\beta} x_i$, $i = 1, 2, \ldots, n$, er residualene fra regresjonen.

(b) Populasjons-standardavviket, σ, estimeres vanligvis ved

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} \qquad \text{(kalt } s \text{ i Løvås)}$$

Inferens.

La θ stå for en av de ukjente populasjonsstørrelsene, $\alpha, \beta, \mu(x_0)$, og $\hat{\theta}$ for mkv estimatoren. Siden $\hat{\theta}$ er forventningsrett, er standardfeilen lik standardavviket, $\sqrt{\text{var}(\hat{\theta})}$. Estimert standardfeil som Løvås betegner med $SE(\hat{\theta})$, fås ved å erstatte den ukjente σ med estimatoren, $\hat{\sigma} = \sqrt{SS_E/(n-2)}$, i **regel 2.**

Tabell 1 ($\hat{\sigma}$ er lik *s* i Løvås)

θ	Mkv estimator $\hat{\theta}$	Estimert st.feil $SE(\hat{\theta})$
α	$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$	$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}$
β	$\hat{\beta} = \frac{S_{xy}}{s_x^2}$	$\frac{\hat{\sigma}}{s_x\sqrt{n-1}}$
$\mu(x_0)$	$\hat{\mu}(x_0) = \hat{\alpha} + \hat{\beta}x_0$	$\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}}$

Tester og konfidensintervall bygger nå på det fundamentale teoremet (bevist i videregående teori) formulert i regel 3, og følger det samme mønsteret beskrevet i notatene *oversikt over tester og konfidensintervall*.

Regel 3

(a) Under modellen beskrevet i (5) gjelder for alle $n > 2$

$$W = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \sim t_{n-2}\text{-fordelt} \quad \text{(t-fordelt med } n-2 \text{ frihetsgrader).}$$

(b) Hvis $n \geq 30$ omtrent, er W tilnærmet $N(0,1)$-fordelt selv om Y_i-ene ikke er normalfordelte.

Konfidensintervall. La $t_{n-2,r}$ betegne *r*-kvantilen i t_{n-2}-fordelingen (definert ved $P(W > t_{n-2,r}) = r$).

Av regel 3(a) får vi et eksakt[3] $1-\alpha$ konfidensintervall for θ

(7) $\hat{\theta} \pm t_{n-2,\alpha/2} SE(\hat{\theta})$

[3] Denne alfaen må ikke forveksles med konstantleddet alfa i regresjonsfunksjonen $\mu(x) = \alpha + \beta x$ i populasjonen (Løvås' notasjon)

dvs $P\left(\hat{\theta} - t_{n-2,\,\alpha/2} SE(\hat{\theta}) \le \theta \le \hat{\theta} + t_{n-2,\,\alpha/2} SE(\hat{\theta})\right) = 1 - \alpha$.

Tester. La som i notatet *oversikt over tester* θ_0 betegne en kjent hypotetisk verdi av θ. Om den sanne verdien θ er lik θ_0 eller ikke vet vi ikke. Som i oversiktsnotatet vil vi bruke samme testobservator for diverse tester om θ:

Testobservator: $T = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$

Under regel 3(a) vil T være t_{n-2}-fordelt som W kun i det spesielle tilfellet at $\theta = \theta_0$, noe som er tilstrekkelig for å bestemme den kritiske verdien (som vil være en kvantil i t_{n-2}-fordelingen) og p-verdien.

Tabell 2

Test-situasjon	H_0	H_1	Testobservator	α-nivå test: Forkast H_0 hvis	$P-verdi$ (t_o er observert verdi av T)
1	$\theta \le \theta_0$	$\theta > \theta_0$	$T = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$	$T > t_{n-2,\,\alpha}$	$P_{\theta=\theta_0}(T > t_o)$
2	$\theta \ge \theta_0$	$\theta < \theta_0$	$T = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$	$T < -t_{n-2,\,\alpha}$	$P_{\theta=\theta_0}(T < t_o)$
3	$\theta = \theta_0$	$\theta \ne \theta_0$	$T = \dfrac{\hat{\theta} - \theta_0}{SE(\hat{\theta})}$	$T < -t_{n-2,\,\alpha/2}$ eller $T > t_{n-2,\,\alpha/2}$	$2 \cdot P_{\theta=\theta_0}(T > \lvert t_o \rvert)$

Merk at hvis $n \ge 30$ omtrent kan vi bytte ut t-fordelingskvantilene med tilsvarende kvantiler i $N(0,1)$ og oppnå en test med tilnærmet nivå α. Likeledes under beregningen av p-verdier, kan vi anta T er tilnærmet $N(0,1)$-fordelt. Dette gjelder selv om Y_i-ene ikke er normalfordelte.

Regneeksempel (Skøytedataene fra EM Heerenveen 2004)

Antall observasjonspar: $n = 27$ (uten Ervik).

x svarer til tiden (i sekunder) på 5000m (ikke-stokastisk) og Y står for tiden (i sekunder) på 1500m (stokastisk).

De fem grunnleggende størrelsene, som gjør resten av regresjonsberegningene relativt enkle med kalkulator, får vi fra Excel[4]:

Tabell 3

[4] Jfr. Excel-øvelsen på slutten av Regr.-I-notatet.

	$\bar{x}$	$\bar{Y}$	s_x^2	S_y^2	S_{xy}
Observert verdi	407,98	112,41	186,0527	8,24001	26,48637

Anta vi i tillegg til regresjonsparametrene er interessert i forventet tid på 1500m hvis 5000m tiden (dagen før) var 6:30 min blank (= 390 sek), nemlig $\mu(390) = \alpha + \beta \cdot 390$.

Vi finner observert verdi: $SS_E = (n-1)\left(S_y^2 - \frac{S_{xy}^2}{s_x^2} \right) \rightarrow 116{,}2050$

Estimatet på σ^2 blir dermed:

(8) $\hat{\sigma}_{obs}^2 = 4{,}64820$

Vi finner nå lett (med kalkulator eller Excel):

Tabell 4

Estimator $\hat{\theta}$	Estimat $\hat{\theta}_{obs}$	St. feil $SE(\hat{\theta})_{obs}$
$\hat{\alpha}$	54,3302	12,6535
$\hat{\beta}$	0,1424	0,0310
$\hat{\mu}(390)$	109,85 sek (1:50 min)	0,6948

Vi ønsker 95% konfidensintervall. Vi trenger da kvantilen $t_{25,\,0{,}025} = 2{,}060$ (fra tabell D5 i Løvås).

Tabell 5

	95% konfidensintervall	
Parameter θ	Nedre grense $\hat{\theta} - (2{,}06)SE$	Øvre grense $\hat{\theta} + (2{,}06)SE$
α	28,264	80,396
β	0,0785	0,2062
$\mu(390)$	108,42 sek (1:48 min)	111,28 sek (1:51 min)

Vi ønsker nå å teste om det er sammenheng mellom *x* og *Y* i det hele tatt, nemlig om det er sterk evidens i data for at $\beta \neq 0$. Det betyr i det generelle oppsettet at θ nå er β og $\theta_0 = 0$. Testobservatoren blir her

$$T = \frac{\hat{\theta} - \theta_0}{SE(\hat{\theta})} = \frac{\hat{\beta}}{SE(\hat{\beta})}$$

og tabell 4 gir observert verdi av testobservatoren:

$$t_0 = T_{obs} = \frac{0{,}1424}{0{,}0310} = 4{,}59$$

Tosidig hypotese: $H_0 : \beta = 0$ mot $H_1 : \beta \neq 0$
Valgt signifikansnivå: for eksempel 5%.
Kritiske verdier: $\pm t_{25,\,0{,}025} = \pm 2{,}06$
Testkriterium (5% signifikansnivå): Forkast H_0 hvis $T_{obs} < -2{,}06$ eller $T_{obs} > 2{,}06$
Konklusjon: Forkast H_0 siden $T_{obs} = 4{,}59$,
(eller: "det er sterk evidens i data for at $\beta \neq 0$").

I dette tilfellet er det kanskje rimeligere å formulere problemet som ensidig ut fra allmenn viten (?) om skøyteløpere som deltar i all-round mesterskap: nemlig at hvis det er sammenheng, så må den antakelig være positiv. I så fall ville vi foretrekke å teste den ensidige hypotesen under istedenfor:

Ensidig hypotese: $H_0 : \beta \leq 0$ mot $H_1 : \beta > 0$
Valgt signifikansnivå: for eksempel 5%.
Kritisk verdi: $t_{25,\,0{,}05} = 1{,}708$
Testkriterium (5% signifikansnivå): Forkast H_0 hvis $T_{obs} > 1{,}708$
Konklusjon: Forkast H_0 siden $T_{obs} = 4{,}59$,
(eller: "det er sterk evidens i data for at $\beta > 0$").