Multippel lineær regresjon, kort repetisjon

Multippel lineær regresjon er en utvidelse av enkel lineær regresjon ved at det er flere enn én regressorvariabel (prediktor). Med *k* variabler kan modellen uttrykkes som

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

$\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ **er ukjente parametrer, regresjonskoeffisienter i regresjon**

$x_{i1}, x_{i2}, \ldots, x_{ik}$ **er gitte, kjente størrelser, binære eller kvantitative**

ε_i **er uavhengige** $N(0, \sigma^2)$

$i = 1, 2, \ldots, n$

Støyleddene ε_i **skal ha samme fordeling, og vi kunne godt ha satt** $\varepsilon_i = \varepsilon$ **for alle *i*. Modellen ovenfor er generell, og den er basis for alle modeller i *lineære modeller* (LM).**

Merk at linearitetskravet ligger på koeffisientene $\beta_0, \beta_1, \ldots, \beta_k$ og ikke på x-ene. Dette innebærer at regressorene som opptrer i modellen, ikke nødvendigvis er x-ene selv, men kan være en hvilken som helt funksjon av dem, slik at den generelle lineære modellen kan uttrykkes som

$$Y_i = \beta_0 + \beta_1 \varphi_1(x_{i1}) + \beta_2 \varphi_2(x_{i2}) + \ldots + \beta_k \varphi_k(x_{ik}) + \varepsilon_i$$

der $\varphi_k(x_{ik})$ er en funksjon av x_{ik}. Dette ses uten videre ved å definere

$z_{ik} = \varphi_k(x_{ik})$**, som gir modellen**

$$Y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_k z_{ik} + \varepsilon_i$$

Eksempel:
Forventningen til *Y* i modellen $Y = \beta_0 + \beta_1 x^2 + \varepsilon$ uttrykker en krum linje i *XY*-planet. Settes $z = x^2$, fås modellen $Y = \beta_0 + \beta_1 z + \varepsilon$ uttrykt som en rett linje i *ZY*-planet.

Multippel regresjonsmodell på matriseform

I multippel lineær regresjon settes dataene inn i en tallmatrise med én søyle for løpenummer, én søyle for responsvariabel *Y* og *k* søyler for regressorene. Antall rekker er antall observasjoner, n.

i	Y_i	x_{i1}	x_{i2}	.	.	x_{ik}
1	Y_1	x_{11}	x_{12}	.	.	x_{1k}
2	Y_2	x_{21}	x_{22}	.	.	x_{2k}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n	Y_n	x_{n1}	x_{n2}	.	.	x_{nk}

Responsene $Y_1, Y_2, \ldots, Y_n$ uttrykkes som en *søylevektor* Y. Tallmatrisen med regressorverdiene samt en søyle med enere i første søyle kaller vi *designmatrisen* X. Støyeddene uttrykkes som søylevektoren ε. I tillegg definerer vi en søylevektor β med elementene $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$, parametrene som skal estimeres.

Vi får nå vektor- og matriseuttrykkene:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & x_{1k} \\ 1 & x_{21} & x_{22} & . & . & x_{2k} \\ 1 & . & . & . & . & . \\ 1 & . & . & . & . & . \\ 1 & x_{n1} & x_{32} & . & . & x_{nk} \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ . \\ . \\ \beta_k \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ \varepsilon_n \end{pmatrix}$$

Modellen kan da skrives på matriseform

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

Under visse forutsetninger og med "egnet" metode kan denne likningen løses med hensyn på β og gi de ukjente parametrene $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$.

Estimering av parametrer

Vi har tidligere lært at regresjonskoeffisientene (parametrene) $\beta_0, \beta_1, \ldots, \beta_k$ blir estimert ved *minste kvadratsums metode* (MK-estimering), som er basert på at kvadratsummen av differensene mellom observert respons Y og estimert Y minimeres (geometriske betraktninger).

Estimatorene er da *forventningsrette* uten hensyn til sannsynlighetsfordelingen til observasjonene.

I LM med mer kompliserte modeller og designmatriser kan det hende at MK-estimering ikke gir noen løsning. I slike tilfeller benyttes *maximum likelihood-estimering* (MLE)

Forenklet sagt går det ut på å bestemme parametrene slik at sannsynligheten for de observasjonene vi faktisk har, blir maksimert.

For å kunne benytte denne metoden, må vi i motsetning til MK-estimering kjenne formelen til den sannsynlighetsfordeling observasjonene har.

Eksempel på ML-estimering

Enkel lineær regresjonsmodell: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, ε_i **er uavhengige** $N(0, \sigma^2)$.

Medfører at $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$.

Sannsynlighetstettheten for en enkelt observasjon Y_i:

$$f_Y(y_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

Vi har her tre ukjente parametrer - β_0, β_1 og σ som skal estimeres. Matematisk sett er det lettere å håndtere σ^2 enn σ. Derfor oppfatter vi variansen σ^2 som parameteren istedenfor standardavviket σ.

***Likelihood* defineres som simultantettheten av alle observasjonene oppfattet som funksjon av de ukjente parametrene. Ettersom observasjoner antas uavhengige, får vi:**

$$L(\beta_0, \beta_1, \sigma^2) = f_Y(y_1) \cdot f_Y(y_2) \cdots f_Y(y_n) = \prod_{i=1}^{n} f_Y(y_i)$$

$$= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}\left(Y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

$$= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 x_i\right)^2\right]$$

Merk at likelihood ved kontinuerlige fordelinger ikke er noen sannsynlighet.

Likelihood skal maksimeres m.h.p. parametrene, men det er enklere å maksimere log-likelihood definert som ln(Likelihood). Resultatet blir det samme.

Log-likelihood og maksimering

$$l(\beta_0, \beta_1, \sigma^2) = ln(L) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \beta_0 - \beta_1 x_i\right)^2$$

skal maksimeres med hensyn på β_0, β_1 og σ^2. Dette gjøres ved at de partiellderiverte settes lik null, og det derav følgende sett på tre likninger løses. Derved får vi estimatorene for β_0, β_1 og σ^2:

$$\tilde{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2}$$ **samme som ved MKE**

$$\tilde{\beta}_0 = \bar{Y} - \beta_1 \bar{x}$$ **samme som ved MKE**

$$\widetilde{\sigma^2} = \frac{\sum(Y_i - \tilde{\beta}_0 - \tilde{\beta}_1 x_i)^2}{n} = \frac{SS_E}{n} = \frac{(n-2)SS_E}{(n-2)n} = \frac{n-2}{n}\widehat{\sigma^2}$$

$\widehat{\sigma^2}$ **fra MKE var forventningsrett, følgelig er** $\widetilde{\sigma^2}$ **forventningsskjev (biased)**

Lineære kontraster i variansanalyse

For tester om forventningene i enveis ANOVA er det hensiktsmessig å benytte *kontraster*.

En kontrast L er en *lineærkombinasjon* av forventningsverdiene $\mu_1\ \mu_2,\cdots,\mu_k$ der koeffisientene summeres opp til null:

$$L=\sum_{i=1}^{k}c_i\mu_i \quad \text{og} \quad \sum_{i=1}^{k}c_i=0$$

En forventningsrett estimator for L er:

$$\hat{L}=\sum_{i-1}^{k}c_i\hat{\mu}_i=\sum_{i=1}^{k}c_i\bar{Y}_{i\cdot}$$

Siden Y_{ij} er uavhengige med konstant varians σ^2, blir variansen til $\hat{L}$:

$$Var(\hat{L})=\sum_{i=1}^{k}c_i^2\cdot\frac{\sigma^2}{n_i}=\sigma^2\sum_{i=1}^{k}\frac{c_i^2}{n_i}$$

Fra ANOVA-tabellen får vi variansestimatoren $\widehat{\sigma^2}=MS_E$, og vi kan da få en estimator for variansen til $\hat{L}$ basert på samtlige observasjoner.

Eksempel:
Anta at vi har 4 grupper og at vi skal teste

$$H_0:\mu_1=\mu_2\ \ (\mu_1-\mu_2=0)\ \text{ mot }\ H_1:\mu_1\neq\mu_2$$

Dette er et eksempel på parvis sammenlikning. Vi lager kontrasten L med koeffisientene $c_1=1,\ c_2=-1,\ c_3=0,\ c_2=0$

$$L=1\cdot\mu_1+(-1)\cdot\mu_2+0\cdot\mu_3+0\cdot\mu_4=\mu_1-\mu_2$$

Testen blir: $L=0$ mot $L\neq 0$

$$\hat{L}=1\cdot\hat{\mu}_1+(-1)\cdot\hat{\mu}_2+0\cdot\hat{\mu}_3+0\cdot\hat{\mu}_4=\hat{\mu}_1-\hat{\mu}_2$$

Testobservator:

$$T=\frac{\hat{\mu}_1-\hat{\mu}_2}{\sqrt{MS_E\left(1/n_1+1/n_2\right)}}$$

, som under H_0 er t-fordelt med $(N-4)$ frihetsgrader.

Dette er i realiteten en to-utvalgs t-test, men merk av variansen til observasjonene er estimert ved MS_E fra ANOVA-tabellen og at antall frihetsgrader derved blir høyere enn ved bare å konsentrere seg om de to gruppene som sammenliknes.

Dersom vi for eksempel ønsker å teste gjennomsnittet av μ_1 og μ_2 mot gjennomsnittet av μ_3 og μ_4, danner vi kontrasten

$$L=\frac{\mu_1+\mu_2}{2}-\frac{\mu_3+\mu_4}{2}$$

Her er $c_1=1/2,\ c_2=1/2,\ c_3=-1/2,\ c_4=-1/2$

Testobservator:

$$T=\frac{\hat{\mu}_1+\hat{\mu}_2-\hat{\mu}_3-\hat{\mu}_4}{\frac{1}{2}\sqrt{MS_E\left(1/n_1+1/n_2+1/n_3+1/n_4\right)}}$$

Dette viser hvordan man ved hjelp av kontraster kan utføre mer komplisert testing i ANOVA og samtidig benytte all informasjon dataene.