Regresjon og konfundering – notat til STK2120

Ørulf Borgan februar 2016

I dette notatet vil vi se litt nærmere på konfundering (engelsk: "confounding") ved bruk av lineær regresjon. Notatet er et supplement til stoffet om lineær regresjon i avsnittene 12.7 og 12.8 i boka til Devore & Berk (D&B). Når ikke annet er sagt bruker vi notasjonen i denne boka.

Problemstilling og resultat

Konfundering er en generell problemstilling ved multippel lineær regresjon når vi har korrelerte forklaringsvariabler. Men for å ikke gjøre framstillingen mer tekninsk enn nødvendig, vil vi her nøye oss med å se på situasjonen med to forklaringsvariabler. Vi antar derfor i dette notetet at de stokastiske variablene $Y_1, \ldots, Y_n$ er gitt ved modellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i, \tag{1}$$

der ϵ_i-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Fra side 712 i D&B vet vi at minste kvadraters estimatorer $\hat{\beta}_1$ og $\hat{\beta}_2$ er forventningsrette for β_1 og β_2.

Den problemstillingen vi vil ta for oss her, er hva som skjer hvis vi tilpasser en lineær regresjonsmodell der vi bare tar med den første av de to forklaringsvariablene. Vi antar altså at modellen er gitt ved (1), men at vi tilpasser modellen

$$Y_i = \gamma_0 + \gamma_1 x_{i1} + \epsilon_i^*. \tag{2}$$

Fra side 626 i D&B får vi da estimatoren

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n (x_{i1} - \overline{x}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2} = \frac{\sum_{i=1}^n (x_{i1} - \overline{x}_1) Y_i}{\sum_{i=1}^n (x_{i1} - \overline{x}_1)^2}, \tag{3}$$

der $\overline{x}_1 = \sum_{i=1}^n x_{i1}/n$ og $\overline{Y} = \sum_{i=1}^n Y_i/n$. Vi vil vise nedenfor at vi har sammenhengen

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \, r_{12} \sqrt{\frac{s_{22}}{s_{11}}}, \tag{4}$$

der

$$s_{11} = \sum_{i=1}^n (x_{i1} - \overline{x}_1)^2, \quad s_{22} = \sum_{i=1}^n (x_{i2} - \overline{x}_2)^2, \quad \text{og} \quad r_{12} = \frac{\sum_{i=1}^n (x_{i1} - \overline{x}_1)(x_{i2} - \overline{x}_2)}{\sqrt{s_{11} \, s_{22}}}. \tag{5}$$

Merk at r_{12} er (den empiriske) korrelasjonen mellom de to forklaringsvariablene og at kvadratroten i (4) er forholdet mellom de (empiriske) standardavvikene for de to forklaringsvariablene.

Når modellen er gitt ved (1), har vi at

$$E(\hat{\gamma}_1) = \beta_1 + \beta_2 \, r_{12} \sqrt{\frac{s_{22}}{s_{11}}}.$$

Hvis de to forklaringsvariblene er korrelerte, vil altså estimatoren (3) *ikke* estimere den faktiske effekten av den første forklaringsvariablen (dvs. β_1), men en størrelse som avhenger av effekten av begge forklaringsvariablene. Vi sier da at effekten av den første forklaringsvariabelen er konfundert av den andre forklaringsvariabelen. Merk at avhengig av fortegnet til korrelasjonen r_{12}, kan konfundering føre både til at vi overestimeter effekten av den første forklaringsvariabelen (dvs. $E(\hat{\gamma}_1) > \beta_1$) og at vi underestimerer effekten (dvs. $E(\hat{\gamma}_1) < \beta_1$).

Et eksempel

Til illustrasjon ser vi på følgende eksempel: Ved et computer science institutt ved et amerikansk universitet har en registrert data om studentenes karakterer de tre første semestrene, samt karakterer fra high school og poeng fra SAT-testen (som kreves ved opptak ved mange amerikanske universiteter). Formålet var å studere i hvor stor grad karakterene fra high school og poengene for SAT-testen kan brukes til å forutsi karakterene de tre føste semestrene ved universitetet. På kurssiden er det gitt en nærmere beskrivelse av dataene og hvordan du kan lese dataene inn i en dataramme i R.

I dette eksempelet vil vi nøye oss med å se på hvordan gjennomsnittskarakten de tre første semestrene ved universitetet (Y) avhenger av poengene fra matematikkdelen av SAT-testen (x_1) og gjennomsnittskarakteren i matematikk fra high school (x_2).

Vi tilpasser først en modell der vi bruker poengene fra matematikkdelen av SAT-testen som eneste forklaringsvariabel:

```
> fit1.kar=lm(kar~satm,data=karakterer)
> summary(fit1.kar)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.2835556  0.3524349   3.642 0.000337
satm        0.0022706  0.0005859   3.875 0.000140

Residual standard error: 0.756 on 222 degrees of freedom
Multiple R-squared:  0.06336,   Adjusted R-squared:  0.05914
F-statistic: 15.02 on 1 and 222 DF,  p-value: 0.0001402
```

Vi finner da at det er en signifikant effekt av poengene fra matematikkdelen av SAT-testen.

Vi tilpasser så en modell der vi bruker både poengene fra matematikkdelen av SAT-testen og gjennomsnittskarakteren i matematikk fra high school som forklaringsvariabler:

```
> fit12.kar=lm(kar~satm+hsm,data=karakterer)
> summary(fit12.kar)

             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.6657425  0.3434918   1.938   0.0539 .
satm        0.0006105  0.0006112   0.999   0.3190
hsm         0.1930048  0.0322237   5.990 8.45e-09 ***

Residual standard error: 0.7028 on 221 degrees of freedom
Multiple R-squared:  0.1942,    Adjusted R-squared:  0.1869
F-statistic: 26.63 on 2 and 221 DF,  p-value: 4.365e-11
```

Vi ser nå at estimatet for effekten av poengene fra matematikkdelen av SAT-testen er redusert til mindre enn en tredel av det det vi fikk for den første modellen og effekten ikke lenger er signifikant. Så det estimatet vi fikk for effekten av poengene fra matematikkdelen av SAT-testeni den første modellen er konfundert av matematikkarakterene fra high school.

Bevis for (4)

For å for å forenkle beviset vil vi anta at forklaringsvariablene er sentreret, dvs. at $\sum_{i=1}^n x_{i1} = \sum_{i=1}^n x_{i2} = 0$. Dette gir ikke noe tap av generalitet, siden vi alltid kan sentrere forklaringsvariablene (ved trekke fra gjennomsnittet av dem), og det vil bare endre estimatorene for kostantleddene i (1) og (2).

Når vi har sentrert forklaringsvariablene kan vi gi estimatoren (3) på formen

$$\hat{\gamma}_1 = \frac{\sum_{i=1}^n x_{i1} Y_i}{\sum_{i=1}^n x_{i1}^2}, \tag{6}$$

mens sammenhnegen vi skal vise [dvs. (4)] kan gis som

$$\hat{\gamma}_1 = \hat{\beta}_1 + \hat{\beta}_2 \frac{s_{12}}{s_{11}}, \tag{7}$$

der vi nå har [jf. (5)]

$$s_{11} = \sum_{i=1}^n x_{i1}^2, \quad s_{22} = \sum_{i=1}^n x_{i2}^2 \quad \text{og} \quad s_{12} = \sum_{i=1}^n x_{i1} x_{i2}.$$

For å vise (7) innfører vi vektor og matrisenotasjon:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \boldsymbol{\epsilon} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \text{og} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

Vi kan da skrive modellen (1) på formen $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ og minste kvadraters estimator er gitt ved

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y}. \tag{8}$$

For å finne eksplisitte uttrykk for $\hat{\beta}_1$ og $\hat{\beta}_2$ ser vi nærmere på høyre side av (8). Vi merker oss først at

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n & 0 & 0 \\ 0 & s_{11} & s_{12} \\ 0 & s_{12} & s_{22} \end{pmatrix}.$$

Determinanten til denne matrisen er $\Delta = n\left(s_{11}s_{22} - s_{12}^2\right)$, og den inverse er gitt ved

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \frac{1}{\Delta} \begin{pmatrix} \Delta/n & 0 & 0 \\ 0 & ns_{22} & -ns_{12} \\ 0 & -ns_{12} & ns_{11} \end{pmatrix}. \tag{9}$$

Videre har vi at

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n x_{i1} Y_i \\ \sum_{i=1}^n x_{i2} Y_i \end{pmatrix}. \tag{10}$$

Av (8), (9) og (10) finner vi nå at

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} \overline{Y} \\ \frac{s_{22}\sum_{i=1}^n x_{i1}Y_i - s_{12}\sum_{i=1}^n x_{i2}Y_i}{s_{11}s_{22}-s_{12}^2} \\ \frac{s_{11}\sum_{i=1}^n x_{i2}Y_i - s_{12}\sum_{i=1}^n x_{i1}Y_i}{s_{11}s_{22}-s_{12}^2} \end{pmatrix}. \tag{11}$$

Vi har altså at $\hat{\beta}_0 = \overline{Y}$ og

$$\hat{\beta}_1 = \frac{s_{22}\sum_{i=1}^n x_{i1}Y_i - s_{12}\sum_{i=1}^n x_{i2}Y_i}{s_{11}s_{22} - s_{12}^2}, \tag{12}$$

$$\hat{\beta}_2 = \frac{s_{11}\sum_{i=1}^n x_{i2}Y_i - s_{12}\sum_{i=1}^n x_{i1}Y_i}{s_{11}s_{22} - s_{12}^2}. \tag{13}$$

Av (12), (13) og (6) får vi nå at

$$\begin{aligned} \hat{\beta}_1 + \hat{\beta}_2 \frac{s_{12}}{s_{11}} &= \frac{s_{22}\sum_{i=1}^n x_{i1}Y_i - s_{12}\sum_{i=1}^n x_{i2}Y_i}{s_{11}s_{22} - s_{12}^2} + \frac{s_{11}\sum_{i=1}^n x_{i2}Y_i - s_{12}\sum_{i=1}^n x_{i1}Y_i}{s_{11}s_{22} - s_{12}^2}\left(\frac{s_{12}}{s_{11}}\right) \\ &= \frac{s_{22}\sum_{i=1}^n x_{i1}Y_i - (s_{12}^2/s_{11})\sum_{i=1}^n x_{i1}Y_i}{s_{11}s_{22} - s_{12}^2} \\ &= \frac{\sum_{i=1}^n x_{i1}Y_i}{s_{11}} \\ &= \hat{\gamma}_1 \end{aligned}$$

som er det vi ville vise [jf. (7)].