Enveis variansanalyse og lineær regresjon – notat til STK1110

Ørulf Borgan oktober 2019

Formålet med dette notatet er å beskrive sammenhengen mellom enveis variansanalyse og multippel lineær regresjon og å diskutere hvordan `R` kan brukes til å tilpasse modeller i variansanalysen. Notatet er et supplement til det som står om enveis variansanalyse i avsnittene 11.1 og 11.3 i boka til Devore & Berk (D&B). Når ikke annet er sagt bruker vi notasjonen i denne boka.

1. Modellene

Vi begynner med å oppsummere de modellene for enveis variansanalyse og multippel lineær regresjon.

Enveis variansanalyse

Modellen for enveis variansanalyse kan gi som (jf. avsnitt 11.3 i D&B)

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}\,; \quad j = 1, 2, \ldots, J_i\,; \quad i = 1, 2, \ldots, I\,; \tag{1}$$

der $\sum_{i=1}^{I} \alpha_i = 0$ og ϵ_{ij}-ene er uavhengige og $N(0, \sigma^2)$-fordelte. Vi lar $n = \sum_{i=1}^{I} J_i$ betegne det totale antall observasjoner.

Multippel lineær regresjon

Modellen for multippel lineær regresjon er gitt ved (jf. avsnitt 12.7 i D&B)

$$Y_l = \beta_0 + \beta_1 x_{l1} + \cdots + \beta_k x_{lk} + \epsilon_l\,; \quad l = 1, 2, \ldots, n\,; \tag{2}$$

der $x_{l1}, x_{l2}, \ldots, x_{lk}$ er kjente tall og ϵ_l-ene er uavhengige og $N(0, \sigma^2)$-fordelte.

Vi bruker her indeksen l for å angi individ i stedet for i som brukes i avsnitt 12.7 i D&B for å unngå sammenblanding med indeksen i i (1).

2. Enveis variansanalyse

Vi ser så på modellen (1) for enveis variansanalyse. I denne modellen angir X_{ij} observasjonen for individ j i gruppe i. Vi vil skrive modellen som en lineær regresjonsmodell, dvs. på formen (2). I (2) angir Y_l observasjonen for individ l, mens $x_{l1}, x_{l2}, \ldots, x_{lk}$ er forklaringvariabler som angir ulike egenskaper ved individet.

For å skrive (1) på formen (2) ser vi først på sammenhengen mellom X_{ij}-ene og Y_l-ene. Vi får denne sammenhengen ved å la de J_1 første Y_l-ene være observasjonene fra gruppe 1, de neste J_2 Y_l-ene være observasjonen fra gruppe 2, osv. Mer presist har vi at (husk at $n = \sum_{i=1}^{I} J_i$):

$$\begin{array}{cccc} Y_1 = X_{11} & Y_2 = X_{12} & \cdots & Y_{J_1} = X_{1J_1} \\ Y_{J_1+1} = X_{21} & Y_{J_1+2} = X_{22} & \cdots & Y_{J_1+J_2} = X_{2J_2} \\ Y_{J_1+J_2+1} = X_{31} & Y_{J_1+J_2+2} = X_{32} & \cdots & Y_{J_1+J_2+J_3} = X_{3J_3} \\ \cdots & \cdots & \cdots & \cdots \\ Y_{n-J_I+1} = X_{I1} & Y_{n-J_I+2} = X_{I2} & \cdots & Y_n = X_{IJ_I} \end{array} \tag{3}$$

For å angi hvilken gruppe en observasjon Y_l hører til, må vi innføre passende forklaringsvariabel. Før vi gjør det, merker vi oss at restriksjonen $\sum_{i=1}^{I} \alpha_i = 0$ gjør at bare $I-1$ av α_i-ene kan variere fritt. Vi vil la de $I-1$ første α_i-ene variere fritt, mens den siste er gitt ved

$$\alpha_I = -\sum_{i=1}^{I-1} \alpha_i \tag{4}$$

For gruppe I kan derfor (1) gis som

$$X_{Ij} = \mu - \sum_{i=1}^{I-1} \alpha_i + \epsilon_{Ij}\,; \quad j = 1, 2, \ldots, J_I \tag{5}$$

Vi innfører nå forklaringsvariablene

$$x_{li} = \begin{cases} 1 & \text{hvis individ } l \text{ hører til gruppe } i \\ -1 & \text{hvis individ } l \text{ hører til gruppe } I \\ 0 & \text{ellers} \end{cases} \tag{6}$$

for $i = 1, \ldots, I-1$. Hvis vi setter disse inn i formelen (2) for lineær regresjon (med $k = I-1$), får vi

$$Y_l = \beta_0 + \beta_i + \epsilon_l \tag{7}$$

hvis individ l hører til gruppe i der $i = 1, \ldots, I-1$, mens

$$Y_l = \beta_0 - \beta_1 - \beta_2 - \cdots - \beta_{I-1} + \epsilon_l \tag{8}$$

hvis individ l hører til gruppe I. Ved å sammenholde (7) med (1) og (8) med (5), ser vi at de to modellene er like når vi lar parametrene $\mu, \alpha_1, \ldots, \alpha_{I-1}$ i (1) svare til parametrene $\beta_0, \beta_1, \ldots, \beta_{I-1}$ i (7) og (8). [Sammenhengen mellom ϵ_{ij}-ene i (1) og ϵ_l-ene i (7) og (8) er tilsvarende som sammenhengen mellom X_{ij}-ene og Y_j-ene i (3).]

Vi har dermed vist at ved å innføre forklaringsvariablene (6) kan modellen (1) for enveis variansanalyse skrives som en multippel lineær regresjonsmodell (2).

3. Bruk av `R` for enveis variansanalyse

For å vise hvordan vi kan bruke `R` for enveis variansanalyse, ser vi på eksempel 11.5 i D&B. Eksempelet gjelder en studie der $n = 20$ rotter ble delt inn i $I = 4$ grupper med $J = 5$ rotter i hver gruppe. Rottene fikk en viss mengde etanol avhengig av hvilken gruppe de var i, og en målte lengden av REM søvn de neste 24 timene. Se side 568 i D&B for en mer utførlig beskrivelse av eksempelet.

Vi kan lese dataene inn i `R` ved kommandoene:

```
path="http://www.uio.no/studier/emner/matnat/math/STK1110/data/exmp11-05.txt"
exmp11.5=read.table(path,sep=",",header=T)
```

For å gjøre en enveis variansanalyse kan vi bruke kommandoen `aov`:

```
> fit.aov=aov(sleep~factor(treat),data=exmp11.5)
```

Da få vi variansanalysetabellen (jf. side 569 i D&B):

```
> summary(fit.aov)
              Df Sum Sq Mean Sq F value   Pr(>F)
factor(treat)  3   5882    1961   21.09 8.32e-06
Residuals     16   1487      93
```

For å få estimater for μ og α_i-ene i (1) kan vi ikke bruke `aov`-kommandoen. Da må vi i stedet bruke `lm`-kommandoen for lineær regresjon. Men før vi bruker denne kommandoen til å gjøre en enveis variansanalyse, må vi gi `R` beskjed om at α_i-ene i (1) skal tilfredsstille restriksjonen $\sum_{i=1}^{I} \alpha_i = 0$. Det gjør vi ved kommandoen:

```
> options(contrasts=c("contr.sum","contr.poly"))
```

Vi kan så tilpasse modellen (1) med kommandoen:

```
> fit.lm=lm(sleep~factor(treat),data=exmp11.5)
```

Da får vi resultatet:

```
> summary(fit.lm)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      55.375      2.156  25.685 1.96e-14
factor(treat)1   23.905      3.734   6.402 8.77e-06
factor(treat)2    6.165      3.734   1.651   0.1182
factor(treat)3   -7.455      3.734  -1.996   0.0632

Residual standard error: 9.642 on 16 degrees of freedom
Multiple R-squared: 0.7982,     Adjusted R-squared: 0.7603
F-statistic: 21.09 on 3 and 16 DF,  p-value: 8.325e-06
```

Merk at det ovenfor er gitt estimater for μ og de tre første α_i-ene. Estimatet for den siste α_i-en får vi ved å bruke relasjonen (4) med $I = 4$.

Etter at vi har tilpasset en variansanalysemodell med `lm`-kommandoen, kan vi få variansanalysetabellen med `anova`-kommandoen:

```
> anova(fit.lm)
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(treat)  3 5882.4 1960.79  21.092 8.325e-06
Residuals     16 1487.4   92.96
```

Vi ser at dette er den samme tabellen som vi fikk med `aov`-kommandoen.

I `lm`-kommandoen (og `aov`-kommandoen) skrev vi `factor(treat)` for å gi `R` beskjed om at `treat` er en kategorisk variabel. Det som skjer når vi bruker `factor(treat)` i `lm`-kommandoen er at `R` lager 3 forklaringsvariabler slik det er gitt i (6). Du kan gi kommandoen `model.matrix(fit.lm)` for å se at det er dette `R` gjør.

4. En alternativ parameterisering av variansanalysemodeller

I modellen for enveis variansanalyse:

$$X_{ij} = \mu + \alpha_i + \epsilon_{ij}\,; \quad j = 1, 2, \ldots, J_i\,; \quad i = 1, 2, \ldots, I\,; \tag{1}$$

er der vanlig å bruke restriksjonen

$$\sum_{i=1}^{I} \alpha_i = 0 \tag{9}$$

slik vi har gjort i avsnittene 2 og 3 ovenfor og slik det er beskrevet i avsnitt 11.3 i D&B. Et alternativ til denne restriksjonen er å sette

$$\alpha_1 = 0 \tag{10}$$

Når vi bruker restriksjonen (10) vil μ i (1) være forventningsverdien for gruppe 1, mens α_i for $i = 2, \ldots, I$ vil være forventet forskjell mellom gruppene $2, \ldots, I$ og gruppe 1 (som da kalles referansegruppen). Om vi velger restriksjonen (9) eller restriksjonen (10) har ingen betydning for variansanalysetabellen. Men estimatene for μ og α_i-ene blir ikke de samme for de to restriksjonene.

Svarende til restriksjonen (10) har vi forklaringsvariablene

$$x_{l,i-1} = \begin{cases} 1 & \text{hvis individ } l \text{ hører til gruppe } i \\ 0 & \text{ellers} \end{cases} \tag{11}$$

for $i = 2, \ldots, I$. Merk at dette svarer til det som står om koding av kategoriske variable ved lineær regresjon på sidene 696-699 i D&B.

Hvis vi setter forklaringsvariablene (11) inn i (2) med $k = I - 1$, ser vi at den lineære regresjonsmodellen er lik variansanalysemodellen (1) når vi bruker restriksjonen (10).

For å vise hvordan vi kan bruke `R` for å få estimater for μ og α_i-ene når vi bruker restriksjonen (10), ser vi på eksempelet i avsnitt 3. Vi må først gi `R` beskjed om at vi vil bruke restriksjonen (10). Det gjør vi ved kommandoen:

```
> options(contrasts=c("contr.treatment","contr.poly"))
```

Merk at opsjonen `options(contrasts=c("contr.treatment","contr.poly"))` er "default", så vi behøver bare gi kommandoen hvis vi tidligere i samme `R`-sesjon har gitt kommandoen `options(contrasts=c("contr.sum","contr.poly"))`.

Vi kan så tilpasse modellen (1) på tilsvarende måte som i avsnitt 3:

```
> fit.lm=lm(sleep~factor(treat),data=exmp11.5)
> summary(fit.lm)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      79.280      4.312  18.386 3.49e-12
factor(treat)2  -17.740      6.098  -2.909   0.0102
```

```
factor(treat)3  -31.360      6.098  -5.143 9.83e-05
factor(treat)4  -46.520      6.098  -7.629 1.02e-06

Residual standard error: 9.642 on 16 degrees of freedom
Multiple R-squared: 0.7982,     Adjusted R-squared: 0.7603
F-statistic: 21.09 on 3 and 16 DF,  p-value: 8.325e-06
```

Merk at vi nå får estimater for μ og de tre siste α_i-ene, mens den første α_i-en er satt til null ved restriksjonen (10).