

# Hypotesetesting ved lineær regresjon – notat til STK2120

Ørulf Borgan februar 2016

I dette notatet vil vi se litt nærmere på testing av hypoteser for den lineære regresjonsmodellen. Notatet er et supplement til det som står om hypotesetesting ved 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.

Vi antar at de stokastiske variablene  $Y_1, \dots, Y_n$  er gitt ved

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{ik} + \epsilon_i \quad (1)$$

der  $\epsilon_i$ -ene er uavhengige og  $N(0, \sigma^2)$ -fordelte. I avsnittene 12.7 og 12.8 i D&B er det beskrevet to tester for den lineære regresjonsmodellen (1).

## 1. Test for at én regresjonskoeffisient er null

For å teste nullhypotesen

$$H_0 : \beta_j = 0 \quad (2)$$

mot alternativet at  $\beta_j \neq 0$ , bruker vi testobservatoren (jf. sidene 688, 689 og 712 i D&B)

$$T = \frac{\hat{\beta}_j}{S_{\hat{\beta}_j}} \quad (3)$$

der  $S_{\hat{\beta}_j}$  er standardfeilen til  $\hat{\beta}_j$ . Hvis nullhypotesen (2) er sann, er testobservatoren  $t$ -fordelt med  $n - (k + 1)$  frihetsgrader. Vi får altså en test med nivå  $\alpha$  hvis vi forkaster nullhypotesen så sant  $|T| > t_{\alpha/2, n-(k+1)}$ .

## 2. Test for at alle regresjonskoeffisientene er null

For å teste nullhypotesen

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_k = 0 \quad (4)$$

mot alternativet at minst én av  $\beta_j$ -ene ikke er lik null, tar vi utgangspunkt i kvadratsummen SSR for regresjon og residualkvadratsummen SSE. Hvis vi lar  $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k$  betegne minste kvadraters estimatorer for modellen (1) og setter

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_k x_{ik} \quad (i = 1, 2, \dots, n)$$

så er kvadratsummene gitt ved

$$\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \quad (5)$$

og

$$\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 \quad (6)$$

For å teste nullhypotesen (4) bruker vi testobservatoren (jf. sidene 687 og 709 i D&B)

$$F = \frac{\text{SSR}/k}{\text{SSE}/[n - (k + 1)]} \quad (7)$$

Hvis nullhypotesen er sann, er testobservatoren  $F$ -fordelt med  $k$  og  $n - (k + 1)$  frihetsgrader. Vi får altså en test med nivå  $\alpha$  hvis vi forkaster nullhypotesen så sant  $F > F_{\alpha, k, n - (k + 1)}$ .

### 3. Test for at noen regresjonskoeffisienter er null

Nullhypotesene (2) og (4) utgjør ytterpunktene av mulige hypoteser om  $\beta_j$ -ene. I stedet for å teste om én  $\beta_j$  er lik null eller om alle  $\beta_j$ -ene er lik null, kan vi være interessert i å teste om noen av  $\beta_j$ -ene er lik null. Konkret vil vi se hvordan vi kan teste nullhypotesen

$$H_0 : \beta_{m+1} = \beta_{m+2} = \dots = \beta_k = 0 \quad (8)$$

mot alternativet at minst én av  $\beta_{m+1}, \beta_{m+2}, \dots, \beta_k$  ikke er lik null.

For å teste nullhypotesen (8) tar vi utgangspunkt i residual kvadratsummen SSE gitt ved (6) og den residual kvadratsummen  $\text{SSE}_0$  vi får når vi tilpasser den lineære regresjonsmodellen under forutsetning at nullhypotesen (8) er sann.

La  $\beta_0^*, \beta_1^*, \dots, \beta_m^*$  være miste kvadraters estimatorer for den lineære regresjonsmodellen

$$Y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_k x_{im} + \epsilon_i \quad (9)$$

og la

$$Y_i^* = \beta_0^* + \beta_1^* x_{i1} + \dots + \beta_m^* x_{im}$$

være de tilsvarende tilpassede verdiene. Da er

$$\text{SSE}_0 = \sum_{i=1}^n (Y_i - Y_i^*)^2 \quad (10)$$

For å teste nullhypotesen (8) bruker vi testobservatoren

$$F = \frac{(\text{SSE}_0 - \text{SSE})/(k - m)}{\text{SSE}/[n - (k + 1)]} \quad (11)$$

Hvis nullhypotesen er sann, er testobservatoren  $F$ -fordelt med  $k - m$  og  $n - (k + 1)$  frihetsgrader. Vi får altså en test med nivå  $\alpha$  hvis vi forkaster nullhypotesen så sant  $F > F_{\alpha, k - m, n - (k + 1)}$ .

Testen i avsnitt 1 er et spesialtilfelle av den testen vi ser på her. For vi kan vise at hvis  $m = k - 1$ , så er  $F = T^2$  der  $T$  er testobservatoren (3) for testing av nullhypotesen  $H_0 : \beta_k = 0$ . Også testen i avsnitt 2 er et spesialtilfelle av testen i dette avsnittet. For hvis nullhypotesen (4) er sann, er  $Y_i^* = \beta_0^* = \bar{Y}$  for alle  $i$ , slik at  $\text{SSE}_0 = \sum_{i=1}^n (Y_i - \bar{Y})^2 = SST$  og dermed  $\text{SSE}_0 - \text{SSE} = SST - \text{SSE} = SSR$  (jf. side 708 i D&B).

#### 4. Bruk av R for testing av (8)

Vi vil vise hvordan vi kan bruke R til å teste nullhypoteser av formen (8). 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 første semestrene ved universitetet. For hver av 224 studenter har vi registrert:

- **id**: Student nummer.
- **kar**: Gjennomsnittskarakter de tre første semestrene ved universitetet. Ved beregningen av gjennomsnittet gir A 4 poeng, B 3 poeng, C 2 poeng og D 1 poeng.
- **hsm**: Gjennomsnittskarakter i matematikk fra high school. Ved beregningen av gjennomsnittet gir A 10 poeng, A- 9 poeng, B+ 8 poeng, osv.
- **hse**: Gjennomsnittskarakter i engelsk fra high school.
- **satm**: Poeng fra matematikkdelen av SAT-testen.
- **satv**: Poeng fra den verbale delen av SAT-testen.

Vi leser dataene inn i en dataramme som vi kaller **karakterer** slik det er beskrevet på kurssidene. Dataene er som følger:

```
> karakterer
  id  kar hsm hse satm satv
1   1 3.32 10 10  670  600
2   2 2.26  6  5  700  640
3   3 2.35  8  8  640  530
4   4 2.08  9  7  670  600
5   5 3.38  8  8  540  580
6   6 3.29 10  8  760  630
7   7 3.21  8  7  600  400
8   8 2.00  3  6  460  530
9   9 3.18  9  8  670  450
10  10 2.34  7  6  570  480
11  11 3.08  9  6  491  488
12  12 3.34  5  7  600  600

.....

215 215 1.80  8  9  620  600
216 216 1.52  9 10  520  570
217 217 3.40  6  9  480  480
218 218 2.86  9  8  640  470
219 219 3.32 10 10  640  560
220 220 2.07  9  6  600  440
221 221 0.85  7  9  510  480
222 222 1.86  7  7  356  350
223 223 2.59  5  7  630  470
224 224 2.28  9  9  559  488
```

Vi vil bruke en lineær regresjonsmodell med `kar` som respons og `hsm`, `hse`, `satm` og `satv` som forklaringsvariable. Vi vil teste nullhypotesen om at SAT-poengene `satm` og `satv` ikke har noen betydning for karakterene ved universitetet. Vi gjør det ved å tilpasse en regresjonsmodell med alle de fire forklaringsvariablene [jf. (1)] og en regresjonsmodell som bare har forklaringsvariablene `hsm` og `hse` [jf. (9)]:

```
> fit.all=lm(kar~hsm+hse+satm+satv,data=karakterer)
> fit.hs=lm(kar~hsm+hse,data=karakterer)
```

For å teste nullhypotesen om at SAT-poengene ikke har noen betydning kan vi bruke `anova`-kommandoen [jf. (8) og (11)]:

```
> anova(fit.hs,fit.all)
Analysis of Variance Table
```

```
Model 1: kar ~ hsm + hse
```

```
Model 2: kar ~ hsm + satm + hse + satv
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	221	108.16				
2	219	107.26	2	0.89801	0.9168	0.4013

Vi ser at nullhypotesen ikke blir forkastet, så det ser ut til at SAT-poengene ikke har noen betydning for å forutsi karakterene ved universitetet (når vi har gitt high school karakterene).