Faculty of Life Sciences

Multipel lineær regression og tosidet variansanalyse

Claus Ekstrøm
E-mail: ekstrom@life.ku.dk

Program

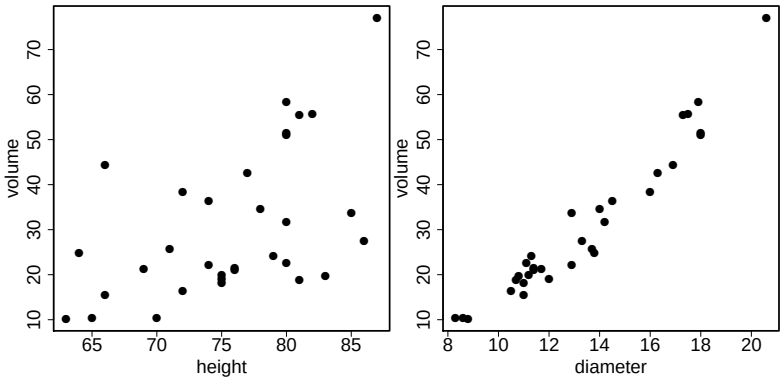
- Multipel lineær regression
- Tosidet variansanalyse
  - Flersidet variansanalyse
- Sammenhæng mellem regressions- og variansanalyse

Eksempel — volumen af kirsebærtræer

Tree	Diameter	Height	Volume	Tree	Diameter	Height	Volume
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Lineær regression

Simpel lineær regression kan beskrive sammenhængen mellem 2 variable:



Lineær regression

Regression af volumen på højde:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835 **
Height        1.5433     0.3839   4.021 0.000378 ***
```

Regression af volumen på diameter:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Girth         5.0659     0.2474   20.48  < 2e-16 ***
```

Men hvad hvis nu begge forklarende variable har en betydning for volumen?

Multipel regression

Den multiple lineære regressionsmodel med d forklarende variable er givet ved

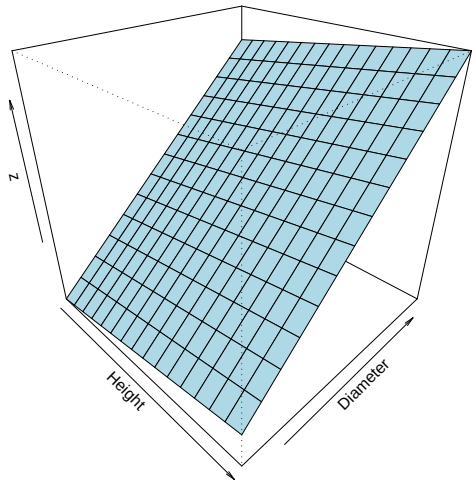
$$y_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_d x_{id} + e_i, \quad i = 1, \ldots, n,$$

hvor $e_i \sim N(0, \sigma^2)$.
Bemærk at den har samme form som simpel lineær regression, med ekstra forklarende variable tilføjet.
Tre middelværdiparametre:

- α skæring (intercept) med y-aksen når $x_{i1} = \cdots = x_{id} = 0$
- β_1 og β_2 er partielle hældninger, dvs. ændringen i y hvis de øvrigt variable "fastfryses".

Samt selvfølgelig variansen σ^2.

Grafisk fremstilling af multipel regression



Estimation og tests for multipel lineær regression

Har allerede lært *alt* det, som vi skal bruge

Vi kan bruge hele maskineriet fra de tidligere uger til estimation (least squares), test af hypoteser (t-tests), konfidensintervaller og modelkontrol.

I R håndteres multipel lineær regression ved at tilføje yderligere led til `lm`.

For eksempel:
`lm(Volumen ~ Height + Girth)`

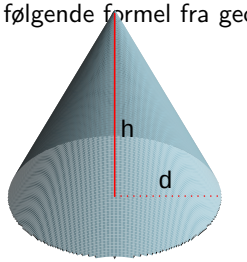
Transformation

Hvis vi tror at vi kan modellere træstammen som en kegle med diameter d og højde h kan vi bruge følgende formel fra geometrien

$$v = \frac{\pi}{12} \cdot h \cdot d^2,$$

og vi kan gøre denne formel lidt mere fleksibel, hvis vi erstatter konstanterne med parametre:

$$v = c \cdot h^{\beta_1} \cdot d^{\beta_2},$$



Med log-transformering får vi

$$\log v_i = \alpha + \beta_1 \log h_i + \beta_2 \log d_i + e_i, \quad i = 1, \ldots, n$$

Polynomial regression

Som et specialtilfælde af multipel lineær regression har vi polynomial regression af orden k

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_k x_i^k + e_i, \quad i = 1, \ldots, n,$$

Kan beskrive komplicerede sammenhænge mellem én variabel og en anden.
Kvadratisk regression er polynomial regression af orden 2

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + e_i, \quad i = 1, \ldots, n,$$

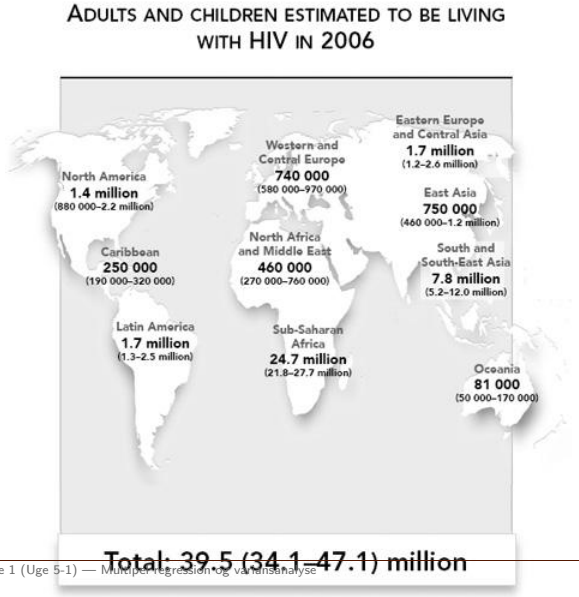
Bemærk, at det er samme forklarende variabel, x, der indgår begge steder som hhv. x_i og x_i^2.
Håndteres i R ved hjælp af `lm`:

```
x2 <- x^2       # Definer ny variabel
lm(y ~ x + x2)
```

eller

```
lm(y ~ x + I(x^2))
```

Prævalens AIDS



Tosidet variansanalyse

Tosidet variansanalyse udvider ensidet ANOVA til mere end en forklarende variabel:

$$y_i = \alpha_{g(i)} + \beta_{h(i)} + e_i, \quad i = 1, \ldots, n,$$

hvor α og β er parametrene hørende til de to kategoriske variabel, og g samt h definerer "grupperne" for de to variable.
Tosidet (og flersidet) variansanalyse håndteres i R ved hjælp af `lm`.
Eksempel:

```
lm(y ~ x1 + x2)
```

hvor man skal huske at `x1` og `x2` skal være defineret som faktorer!
Ellers

```
lm(y ~ factor(x1) + factor(x2))
```

Bemærk, at vi nu har flere interessante hypoteser, vi kan teste (med `drop1()` i R)!

Eksempel — dyrkning af kål

Fire marker er opdelt i 4 parceller hvor der blev dyrket kål. Interesseret i at undersøge tilførsel af nitrogen som calciumnitrat (C), ammoniumsulfat (A), nitrat (N) eller som kontrol (K).

Udbytte	Mark 1	Mark 2	Mark 3	Mark 4
C	70.3	72.5	79.0	86.2
A	75.5	63.0	65.4	67.7
N	85.2	80.5	83.6	92.3
K	36.7	39.6	45.5	50.5

Inferens i tosidet variansanalyse

Har allerede lært alt det, som vi skal bruge

Vi kan bruge hele maskineriet fra de tidligere uger til estimation (least squares), test af hypoteser (F-tests), konfidens- og prædiktionsintervaller og modelkontrol.

Det eneste, der er lidt anderledes er antallet af parametre/frihedsgrader, men det kan programmerne let regne ud for en.

Alt er lineær regression

Lineær regression og variansanalyse er samme model

Det er muligt at specificere kategoriske variable ved hjælp af modellen for multipel lineær regression.
I sidste ende betyder det, at vi kan lave modeller, som både har kategoriske og kvantitative forklarende variable.

For at skrive variansanalyse som regression kan vi benytte dummy variable

$$x_{ij}^k = \begin{cases} 1 & \text{hvis observation } i \text{ tilhører kategori } j \text{ for variabel } k \\ 0 & \text{ellers} \end{cases}.$$

Dagens hovedpunkter

- Multipel lineær regression
  - Hvad kan det?
  - Fortolkning, estimation og test af hypoteser
- Flersidet variansanalyse
  - Hvad kan det?
  - Fortolkning, estimation og test af hypoteser
- Alt er "lineær regression"