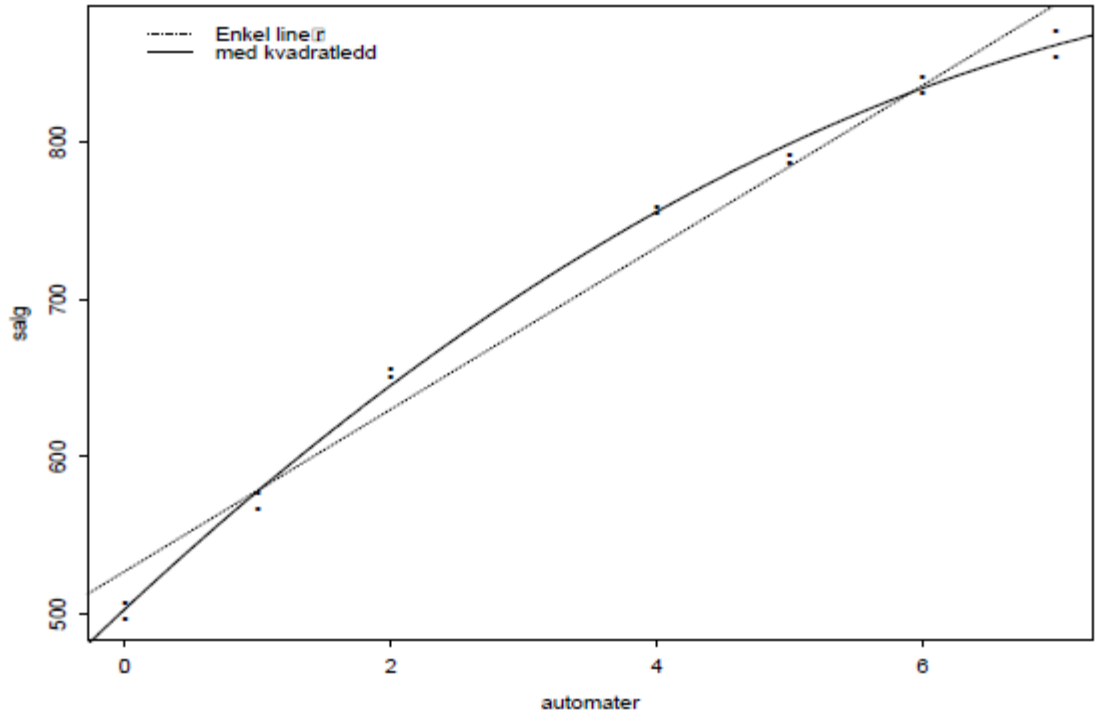
Multippel regresjon og anova

STK2120 våren 2012

Lin. regr. med kvadratledd

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

Kaffesalg: Bedre tilpasning med 2.gradsledd.



Multippel lineær regresjon

Responser ("avhengige" variable) $y_i, i = 1, \ldots, n$

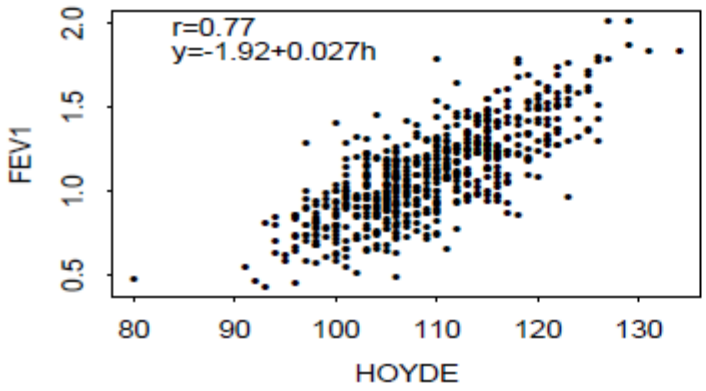
Forklaringsvariable (uavhengige variable, kovariater): $x_{i1}, x_{i2}, \ldots, x_{i,p-1}$

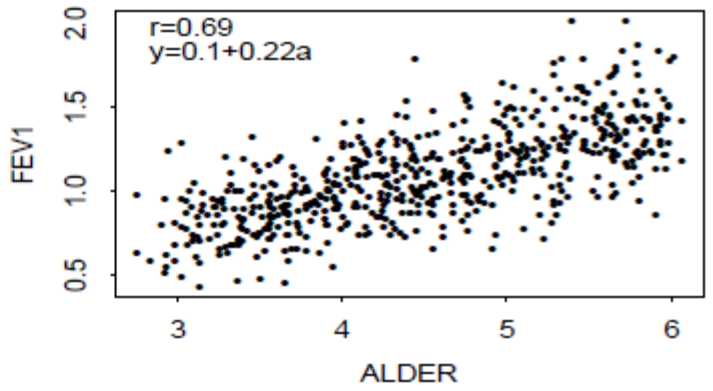
Eksempel kaffesalg: $x_{i1} = x_i, x_{i2} = x_i^2$

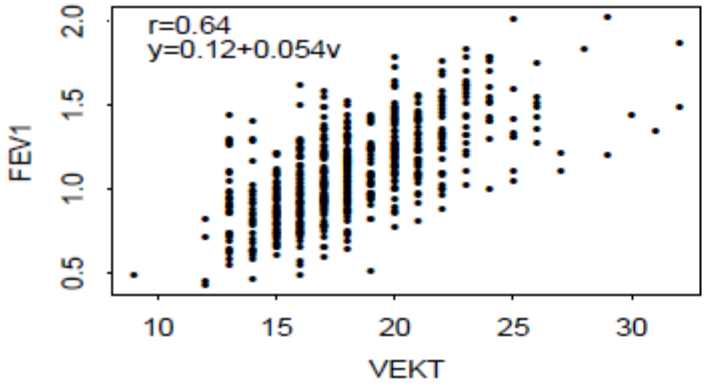
Eksempel lungefunksjon, y_i =FEV1

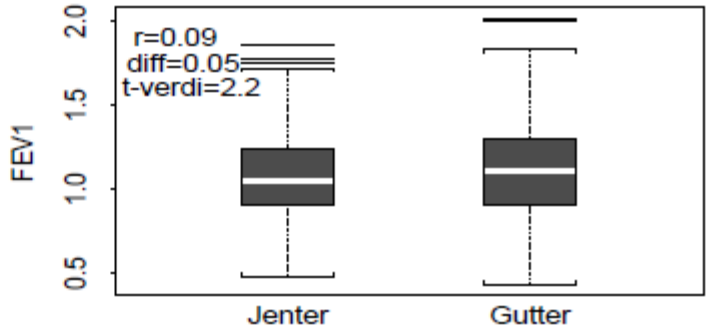
- x_{i1} = høyde
- x_{i2} = alder
- x_{i3} = vekt
- x_{i4} = kjønn (=0 for gutter og =1 for jenter)

Eksempel 2: Lungefunksjon









Multippel lineær regresjon

Modell:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \varepsilon_i$$

der $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$ og uavhengige.

Total p regresjonsparametre $\beta_0, \beta_1, \ldots, \beta_{p-1}$ (samt σ)

Alternativt kan vi spesifisere modellen

- $\mathrm{E}[y_i]$ er lineær i x_{ij}
- Uavhengige y_i
- Konstant varians $\mathrm{Var}[y_i] = \mathrm{Var}[\varepsilon_i] = \sigma^2$
- Normfordelte y_i

Formål multippel regresjon

- I forhold til enkle lineær regresjon:
  kan ta hensyn til ikke-lineariteter

- Multifaktorielle fenomener

  - Mer presis prediksjon

  - Kan løse opp i "sammenblandende" effekter
    Eks: Er det høyde, alder eller vekt (eller alle tre kovariater) som "forklarer" variasjonen i lungefunksjon?

Minste kvadrater mult. regr.

Kvadratsum, med $\beta^\top = (\beta_0, \beta_1, \ldots, \beta_{p-1})$,

$$S(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_{p-1} x_{i,p-1})^2$$

finnes ved å løse p lineære ligninger med p ukjente

$$\frac{\partial S(\beta)}{\partial \beta_j}\Big|_{\beta=\hat{\beta}} = 0, \quad j = 0, 1, \ldots, p-1$$

der $\hat{\beta}^\top = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_{p-1})$ er minste kvadraters estimatoren (MKE).

Minste kvadrater mult. regr.

Eksplisitt er de derivert av $S(\beta)$ gitt ved

$$\frac{\partial S(\beta)}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \mu_i)$$

$$\frac{\partial S(\beta)}{\partial \beta_j} = -2 \sum_{i=1}^{n} x_{ij}(y_i - \mu_i) \quad , j = 1, 2, ..., p-1$$

der $\mu_i = \mathrm{E}[y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$

Matrise-representasjon

Vektor av responser og forventninger

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \qquad \text{og} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}$$

Matrise av forklaringsvariable, **designmatrisen**,

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix}$$

Matriserepresentasjon

Med $\mathbf{Y}$, μ og $\mathbf{X}$ kan vi uttrykke systemet av partiellderiverte av $S(\beta)$ ved

$$\frac{\partial S}{\partial \beta} = -2\mathbf{X}^\top[\mathbf{Y} - \mu]$$

der $\mathbf{X}^\top$ er den transponerte matrisen til $\mathbf{X}$.

Vi kan dessuten uttrykke

$$\mu = \mathbf{X}\beta$$

Dermed kan vi skrive ligningene for å bestemme MKE β ved

$$\mathbf{X}^\top[\mathbf{Y} - \mathbf{X}\hat{\beta}] = 0$$

eller ekvivalent ved "normal-ligningene"

$$\mathbf{X}^\top \mathbf{Y} = \mathbf{X}^\top \mathbf{X} \hat{\beta}$$

Såsant $\mathbf{X}^\top \mathbf{X}$ er inverterbar finner vi MKE på matriseform ved

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}$$

Dette kommer vi tilbake til i 12.8!

Parametrisering med enveis variansanalyse

Sammenligning av forventning mellom J grupper:

Modell: Anta at individ i er i gruppe j. Da er $Y_i \sim \mathrm{N}(\mu_j, \sigma^2)$ La for $j = 1, \ldots, J$

$$x_{ij} = \begin{cases} 1 & \text{hvis } i \text{ er i gruppe } j \\ 0 & \text{ellers} \end{cases}$$

Vi starter med å renummerere Y_{ij} som Y_i fortløpende, i=1,..., JK (K obs i hver gruppe).

Da kan vi skrive dette som en lineær modell uten konstantledd

$$\mu_j = \mathrm{E}[Y_i] = \mu_1 x_{i1} + \mu_2 x_{i2} + \cdots + \mu_J x_{iJ}$$

Hjørnepunkt-parametrisering = "treatment-kontrast"

Vi kan velge en av gruppene som referansegruppe, f.eks. gruppe 1 og skrive om enveis ANOVA til

$$\begin{aligned} \mu_j = \mathrm{E}[Y_i] &= \mu_1 + (\mu_2 - \mu_1)x_{i2} + \cdots + (\mu_J - \mu_1)x_{iJ} \\ &= \beta_1 + \beta_2 x_{i2} + \cdots + \beta_J x_{iJ} \end{aligned}$$

der altså $\beta_1 = \mu_1$ og $\beta_j = \mu_j - \mu_1$ for $j > 1$.

Denne parametriseringen er naturlig hvis man vil sammenligne $J - 1$ nye behandlinger med en tradisjonell behandling.

Hjørnepunkt-parametrisering / treatment-contrast er default i R.

Sum-parametrisering (kontrast)

Tradisjonelt i ANOVA benyttes imidlertid ofte "sum-parametrisering" med

$$\mu_j = \mathrm{E}[Y_i] = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_J x_{iJ}$$

der

$$\alpha_1 + \alpha_2 + \cdots + \alpha_J = 0$$

Merk at $\sum_{j=1}^{J} x_{ij} = 1$ og med konstantledd α_0 i modellen er det overparametrisert uten en restriksjonen som $\sum_{j=1}^{J} \alpha_j = 0$

Med sum-parametrisering blir $\alpha_J = -(\alpha_1 + \cdots + \alpha_{J-1})$ og

$$\begin{aligned}
\mu_j &= \alpha_0 + \alpha_1 x_{i1} + \cdots + \alpha_{J-1} x_{i,J-1} - (\alpha_1 + \cdots + \alpha_{J-1}) x_{iJ} \\
&= \alpha_0 + \alpha_1 (x_{i1} - x_{iJ}) + \cdots + \alpha_{J-1} (x_{i,J-1} - x_{iJ}) \\
&= \alpha_0 + \alpha_1 x'_{i1} + \cdots + \alpha_{J-1} x'_{i,J-1}
\end{aligned}$$

Sum-parametrisering (kontrast), forts.

Sum-parametriseringen gir altså J parametre i - på samme måte som hjørnepunkt-paramterisering - men med forklaringsvariable

$$x'_{ij} = x_{ij} - x_{iJ}$$

Sum-kontrast spesifiseres i R ved

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

Se bare bort fra `"contr.poly"` som benyttes for en spesiell type kategorisk forklaringsvariabel.

For å komme tilbake til hjørnepunk/treatment-parametrisering:

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

Eks: Kategoriske ¨kovariater

Y = Inntekt etter kjønn og sosioøkonomisk gruppe:

	Sted 1	Sted 2	Sted 3
Mann	300 350 370 360	400 370 420 390	400 430 420 410
Kvinne	300 320 310 305	350 370 340 355	370 380 360 365

Altså to kategoriske forklaringsvariable = faktorer i R-terminologi:

1. Kjønn med 2 nivåer
2. Sted med 3 nivåer

Eks: Inntekt over sted

Ser bort fra kjønn. Kun en faktor og altså enveis ANOVA.

```
> inntekt<-c(300,350,370,360,400,370,420,390,400,430,420,410,300,320,31
> kjonn<-c(rep(1,12),rep(2,12))
> sted<-rep(c(1,1,1,1,2,2,2,2,3,3,3,3),2)
```

Vi har tidligere brukt

> sted = as.factor(sted)
> fit=aov(inntekt~sted)

som egentlig kaller funksjonen lm. Vi kan like gjerne bruke

> modell=lm(inntekt~sted)
> anova(modell)

som også gir oss modellparametre og designmatrisen. Men lm bruker default hjørneparametrisering, så vi må huske å spesifisere sum-kontrast først!

Eks: Inntekt over sted med sum-kontrast

```
> options(contrasts=c("contr.sum","contr.poly"))
> summary(lm(inntekt~factor(sted)))

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    364.375      5.619  64.841  < 2e-16 ***
factor(sted)1  -37.500      7.947  -4.719 0.000117 ***
factor(sted)2   10.000      7.947   1.258 0.222090
---
Residual standard error: 27.53 on 21 degrees of freedom
Multiple R-Squared: 0.5321,     Adjusted R-squared: 0.4875
F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344

> anova(lm(inntekt~factor(sted)))
Analysis of Variance Table

Response: inntekt
             Df  Sum Sq Mean Sq F value   Pr(>F)
factor(sted)  2 18100.0  9050.0  11.941 0.000344 ***
Residuals    21 15915.6   757.9
---
```

Eksempel inntekt med hjørnepunkt-parametrisering

```
> summary(lm(inntekt~factor(sted)))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   326.875      9.733  33.583  < 2e-16 ***
factor(sted)2  47.500     13.765   3.451 0.002394 **
factor(sted)3  65.000     13.765   4.722 0.000116 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27.53 on 21 degrees of freedom
Multiple R-Squared: 0.5321,     Adjusted R-squared: 0.4875
F-statistic: 11.94 on 2 and 21 DF,  p-value: 0.000344

> anova(lm(inntekt~factor(sted)))
Analysis of Variance Table

Response: inntekt
             Df  Sum Sq Mean Sq F value   Pr(>F)
factor(sted)  2 18100.0  9050.0  11.941 0.000344 ***
Residuals    21 15915.6   757.9
```

Få ut designmatrise med sum-kontrast (obs deler av matrisen syns ikke, den skal ha 24 linjer)

```
> options(contrasts=c("contr.sum","contr.poly"))
> enveisfit<-lm(inntekt~factor(sted),x=T)
> enveisfit$x
   (Intercept) factor(sted)1 factor(sted)2
1            1             1             0
2            1             1             0
3            1             1             0
4            1             1             0
5            1             0             1
6            1             0             1
7            1             0             1
8            1             0             1
9            1            -1            -1
10           1            -1            -1
11           1            -1            -1
12           1            -1            -1
13           1             1             0
14           1             1             0
15           1             1             0
16           1             1             0
17           1             0             1
```

Toveis variansanalyse uten interaksjon

$Y_i \sim \mathrm{N}(\mathrm{E}[Y_i], \sigma^2)$ uavhengige med

- nivå j på faktor 1 med ialt J nivåer
- nivå k på faktor 2 med ialt K nivåer

Med hjørnepunkt-parametrisering kodes 1. faktor ved $x_{ij} = 1$ mens $x_{ij'} = 0$ og 2. faktor ved $z_{ik} = 1$ mens $z_{ik'} = 0$ slik at forventningen blir (med $\beta_1 = \alpha_1 = 0$)

$$\mathrm{E}[Y_i] = \beta_0 + \sum_{j=2}^{J} \beta_j x_{ij} + \sum_{k=2}^{K} \alpha_k z_{ik} = \beta_0 + \beta_j + \alpha_k$$

Toveis ANOVA uten interaksjon med sum-kontrast

$Y_i \sim \mathrm{N}(\mathrm{E}[Y_i], \sigma^2)$ uavhengige med

- nivå j på faktor 1 med ialt J nivåer
- nivå k på faktor 2 med ialt K nivåer

Med x_{ij} ogz_{ik} som ved hjørnepunkt-parametrisering kodes nå 1. faktor ved $x'_{ij} = x_{ij} - x_{iJ}$ og 2. faktor ved $z'_{ik} = z_{ik} - z_{iK}$

ANOVA og regresjon generelt

K-veis ANOVA = Multippel regresjon med dummyvariable som angir nivå for hver faktor.

Kan analyseres selv om designet ikke er "balansert", dvs. ulikt antall observasjoner for hver kombinasjon av nivåer.

Kan gjerne blande skalakovariater og kategoriske kovariater i en modell.