10. Recap: Algemeen Lineair Model - Additief Model

Lieven Clement

statOmics, Ghent University (https://statomics.github.io)

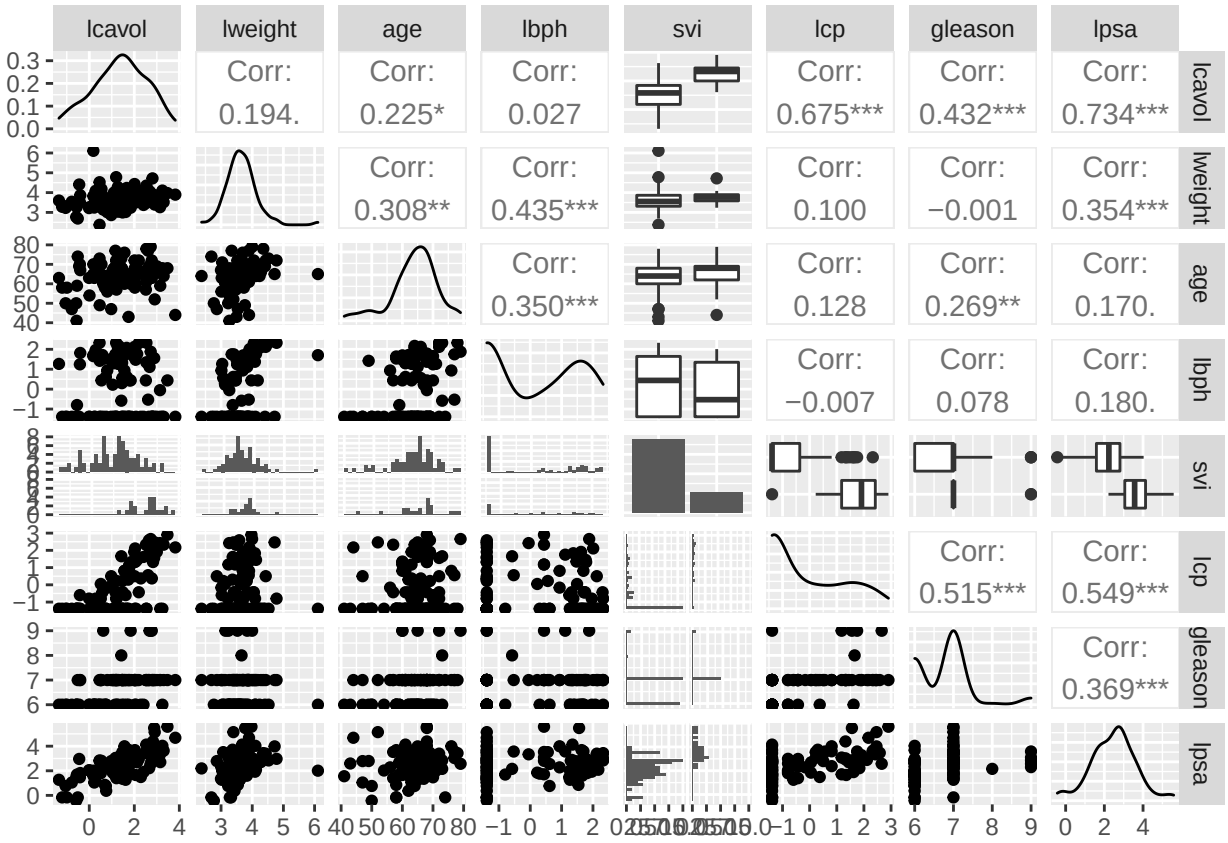
Contents

Dataset - Onderzoeksvraag - Design?

- Prostaatkanker case studie
- Associatie tussen prostaat specifiek antigen concentratie en andere predictoren
- Type response?
- Type predictoren?

Data-exploratie

```
prostate <- read_csv("https://raw.githubusercontent.com/statomics/sbc20/master/data/prostate.csv")

prostate <- prostate %>%
  mutate(svi = as.factor(svi))

library(GGally)
prostate %>%
  dplyr::select(-pgg45)  %>%
  ggpairs()
```



- Schatting voor parameter β_v mogelijks geen zuiver effect van tumor volume.
- Zelfs als lcavol niet is geassocieerd met het lpsa, dan nog kunnen patiënten met een groter tumor volume een hoger lpsa hebben omdat ze bijvoorbeeld een aantasting van de zaadblaasjes hebben (svi status 1). → Confounding.
- Door de svi status in het model op te nemen corrigeren we voor de mogelijkse confounding.

Vertalen van onderzoeksvraag naar populatie parameters: effectgrootte

$$\mathrm{E}\left(Y|X_v, X_w, X_s\right) = \beta_0 + \beta_v X_v + \beta_w X_w + \beta_s X_s$$

- Associatie van predictoren met log PSA: hellingen van het model

- Meer accurate predicties door meerdere predictoren simultaan in rekening te brengen
- Interpretatie?
  - verschil in gemiddelde uitkomst tussen subjecten die in één eenheid van log tumor volume (X_v) verschillen, maar dezelfde waarde hebben voor de overige verklarende variabelen (X_w en X_s) in het model.
  - Associatie tussen log PSA en de predictor log tumor volume waarbij gecorrigeerd wordt voor de overige predictoren, hier dus associatie van log PSA en het log tumor volume na correctie voor log prostaatgewicht en svi-status.

Schatten van effectgrootte a.d.h.v. steekproef

- Kleinste kwadratentechniek

```
lmV <- lm(lpsa~lcavol, prostate)
summary(lmV)
```

```
Call:
lm(formula = lpsa ~ lcavol, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.67624 -0.41648  0.09859  0.50709  1.89672

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.50730    0.12194   12.36   <2e-16 ***
lcavol       0.71932    0.06819   10.55   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7875 on 95 degrees of freedom
Multiple R-squared:  0.5394,    Adjusted R-squared:  0.5346
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

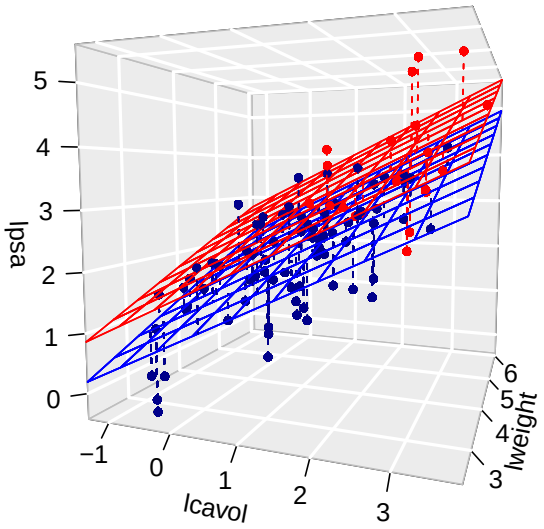
```
lmVWS <- lm(lpsa~lcavol + lweight + svi, prostate)
summary(lmVWS)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
```

```
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

De parameter bij lcavol geeft nu aan dat patiënten met een tumorvolume dat 1% hoger ligt, maar eenzelfde prostaat gewicht en svi status hebben, een prostaat antigeen concentratie zullen hebben dat gemiddeld slechts 0.55% hoger ligt.

De reden dat we eerder een verschil van meer dan 0.72% vonden, kan worden verklaard doordat patiënten met een verschil in tumorvolume vaak ook verschillen in prostaat gewicht en svi status en omdat prostaat gewicht en svi mogelijks ook een associatie vertonen met log PSA



Inferentie

- Kunnen we hetgeen we zien in de steekproef vertalen naar de populatie toe?
- Hiervoor moeten we rekening houden dat we maar een heel klein deel van de populatie hebben kunnen bemonsteren.
- Gevens, statistieken en conclusies zijn stochastisch. Ze variëren van steekproef tot steekproef.
- We moeten die variabiliteit in kunnen schatten o.b.v. één enkele steekproef!

Aannames?

Representatieve steekproef:

$\hat{\beta}_j$ is een onvertekende schatter van β als steekproef representatief is

$$E[\hat{\beta}_j] = \beta_j$$

Normaliteit

$$Y_i \sim N(\mu_i, \sigma^2)$$
$$Y_i \sim N(\beta_0 + \beta_v x_{iv} + \beta_w x_{iw} + \beta_s x_{is}, \sigma^2) \longrightarrow \hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j})$$

en lineaire combinaties van de model parameterschatters zijn ook normaal verdeeld.

$$\longrightarrow L^T\hat{\beta} \sim N(L^T\beta, \sigma^2_{L^T\hat{\boldsymbol{\beta}}})$$

Onafhankelijkheid en gelijkheid van variantie

$$\sigma^2_{L^T\hat{\boldsymbol{\beta}}} = c_L\sigma^2$$

- σ^2?

$$\hat{\sigma}^2 = MSE = \sum_{i=1}^{n} \frac{(Y_i - \hat{Y}_i)^2}{n-p}$$

$$SE_{L^T\hat{\boldsymbol{\beta}}} = c_L\hat{\sigma}$$

- t statistiek

$$T = \frac{L\hat{\boldsymbol{\beta}} - L\boldsymbol{\beta}}{\mathrm{SE}_{L\hat{\boldsymbol{\beta}}}} \sim t_{n-p}$$

- BI en T-test $H_0 : L\boldsymbol{\beta} = 0$ vs $H_1 : L\boldsymbol{\beta} \neq 0$
- F statistiek volgt F-verdeling onder de H_0

$$F = \frac{MSR_2 - MSR_1}{MSE} \sim F_{p_2-p_1, n-p_2}$$

R - output

```
library(car)
summary(lmVWS)
```

```
Call:
lm(formula = lpsa ~ lcavol + lweight + svi, data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.72966 -0.45767  0.02814  0.46404  1.57012

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.26807    0.54350  -0.493  0.62301
lcavol       0.55164    0.07467   7.388  6.3e-11 ***
lweight      0.50854    0.15017   3.386  0.00104 **
sviinvasion  0.66616    0.20978   3.176  0.00203 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7168 on 93 degrees of freedom
Multiple R-squared:  0.6264,    Adjusted R-squared:  0.6144
F-statistic: 51.99 on 3 and 93 DF,  p-value: < 2.2e-16
```

```
Anova(lmVWS, type = "III")
```

```
Anova Table (Type III tests)

Response: lpsa
            Sum Sq Df F value    Pr(>F)
(Intercept)  0.125  1  0.2433  0.623009
lcavol      28.045  1 54.5809 6.304e-11 ***
lweight      5.892  1 11.4678  0.001039 **
svi          5.181  1 10.0841  0.002029 **
Residuals   47.785 93
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusie

De associaties tussen lpsa $\leftrightarrow$ log kanker volume, lpsa $\leftrightarrow$ log prostaat gewicht en lpsa $\leftrightarrow$ status van de zaadblaasjes zijn respectievelijk extreem significant ($p << 0.001$) en sterk significant ($p = 0.001$ en $p = 0.002$).

- interpretaties van de hellingen en BI!

Wat als aannames niet zijn voldaan?

- Normaliteit en heteroscedasticiteit niet voldaan: transformatie van Y
- Lineariteit niet voldaan: transformatie van X of hogere orde termen (interacties en machten $X^2, X^3, ...$).
- Normaliteit niet voldaan: bij grote steekproeven CLT