# Overzichtsdocument voor examen Biostatistiek: Partim Univariate Statistiek

Lieven Clement

## 1 Algemeen Lineair model

### 1.1 Data

- Prostaatkanker case studie
- Associatie tussen log prostaat specifiek antigen (lpsa) concentratie en log cancer volume (v), log gewicht (w) en status van de zaadblaasjes (s).
- Schatting voor parameter $\beta_v$ 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). $\rightarrow$ Confounding.
- Door de svi status in het model op te nemen corrigeren we voor de mogelijkse confounding.

### 1.2 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.

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

- Kleinste kwadratentechniek

## 1.4 Inferentie

### 1.4.1 Aannames?

- Representatieve steekproef:

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

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

- Normaliteit

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

- 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

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

- $\sigma^2$?

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

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

- 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 $H_0$

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

## 1.5 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

## 1.6 Model met interacties

- Effect modificatie!
- **Associatie van een predictor en de respons variabele** hangt van de waarde van een **andere predictor**.

### 1.6.1 Interactie tussen continue predictor variabele en een categorische predictor?

Om deze **interactie** of **effectmodificatie** tussen variabelen $X_v$ en $X_s$, en $X_w$ en $X_s$ statistisch te modelleren, kan men de producten van beide variabelen in kwestie aan het model toevoegen

$$Y_i = \beta_0 + \beta_v x_{iv} + \beta_w x_{iw} + \beta_s x_{is} + \beta_{vs} x_{iv} x_{is} + \beta_{ws} x_{iw} x_{is} + \epsilon_i$$

Deze termen kwantificeren de *interactie-effecten* van respectievelijk de predictoren $x_v$ en $x_s$, en, $x_v$ en $x_s$ op de gemiddelde uitkomst.

In dit model worden de termen $\beta_v x_{iv}$, $\beta_w x_{iw}$ en $\beta_s x_{is}$ de *hoofdeffecten* van de predictoren $x_v$, $x_w$ en $x_s$ genoemd.

Omdat $X_S$ een dummy variabele is, verkrijgen we verschillende regressievlakken:

1. Model voor $X_s = 0$:
$$Y = \beta_0 + \beta_v X_v + \beta_w X_w + \epsilon$$
waar de hoofdeffecten de hellingen voor lcavol en lweight zijn
2. en het model voor $X_s = 1$:
$$\begin{aligned} Y &= \beta_0 + \beta_v X_v + \beta_s + \beta_w X_w + \beta_{vs} X_v + \beta_{ws} X_w + \epsilon \\ &= (\beta_0 + \beta_s) + (\beta_v + \beta_{vs}) X_v + (\beta_w + \beta_{ws}) X_w + \epsilon \end{aligned}$$
met intercept $\beta_0 + \beta_s$ en hellingen $\beta_v + \beta_{vs}$ en $\beta_w + \beta_{ws}$

- De helling voor lcavol en lweight hangt af van de status van de zaadblaasjes!

### 1.6.2 Interactie tussen continue predictoren?

$$Y_i = \beta_0 + \beta_v x_{iv} + \beta_w x_{iw} + \beta_s x_{is} + \beta_{vw} x_{iv} x_{iw} + \epsilon_i$$

Deze term kwantificeert het *interactie-effect* van de predictoren $x_v$ en $x_w$ op de gemiddelde uitkomst.

Het effect van een verschil in 1 eenheid in $X_v$ op de gemiddelde uitkomst bedraagt nu:

$$\begin{aligned} &E(Y|X_v = x_v + 1, X_w = x_w, X_s = x_s) - E(Y|X_v = x_v, X_w = x_w, X_s = x_s) \\ &= [\beta_0 + \beta_v(x_v + 1) + \beta_w x_w + \beta_s x_s + \beta_{vw}(x_v + 1)x_w] - [\beta_0 + \beta_v x_v + \beta_w x_w + \beta_s x_s + \beta_{vw}(x_v)x_w] \\ &= \beta_v + \beta_{vw} x_w \end{aligned}$$

- De helling voor lcavol hangt m.a.w. af van het log gewicht van de prostaat!
- We kunnen hetzelfde doen voor lweight. Helling voor lweight hangt af van het log volume van de tumor!

## 1.7 Anova tabel en aanvullende kwadratensommen?

Beschouw de volgende twee **geneste** regressiemodellen voor predictoren $x_1, \ldots, x_{p-k}$ en $x_1, \ldots, x_{p-1}$:

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-k} x_{ip-k} + \epsilon_i,$$

met $\epsilon_i$ iid $N(0, \sigma_1^2)$, en

$$Y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{p-1} x_{ip-1} + \epsilon_i,$$

met $\epsilon_i$ iid $N(0, \sigma_2^2)$.

Voor het eerste (gereduceerde) model geldt de decompositie

$$\text{SSTot} = \text{SSR}_1 + \text{SSE}_1$$

en voor het tweede (niet-gereduceerde) model

$$\text{SSTot} = \text{SSR}_2 + \text{SSE}_2$$

(SSTot is uiteraard dezelfde in beide modellen omdat dit niet afhangt van het regressiemodel).

$$\text{SSTot} = \text{SSR}_{2|1} + \text{SSR}_1 + \text{SSE}_2$$

$$F = \frac{\text{SSR}_{2|1}}{\text{SSE}_2}$$

- Testen voor alle niveaus van een factor simultaan: omnibus test.
- Testen voor totale effect van een predictor: hoofdeffecten + interacties
- ...

## 1.8 Diagnostiek

### 1.8.1 Multicollineariteit

$$\text{VIF}_j = \left(1 - R_j^2\right)^{-1}$$

### 1.8.2 Invloedrijke Observaties

1. Cooks distance

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\text{MSE}}$$

- Extreme Cook's distance als het het 50% percentiel van de $F_{p,n-p}$-verdeling overschrijdt.

2. DFBETAS

$$\text{DFBETAS}_{j(i)} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\text{SD}(\hat{\beta}_j)}$$

- DFBETAS extreem is wanneer ze 1 overschrijdt in kleine tot middelgrote datasets en $2/\sqrt{n}$ in grote datasets

## 1.9 Contrasten

### 1.9.1 Bloeddruk voorbeeld

**1.9.1.1 Remediering heteroscedasticiteit** Illustratie: Bij grote steekproeven kunnen we corrigeren voor heteroscedasticiteit.

```
mSd <- lm(mBp1$res %>% abs ~ mBp1$fitted)
```

We schatten het model nu opnieuw:

```
mBp3 <- lm(BPSysAve ~ Age*Gender, bpData, w = 1/mSd$fitted^2)
```

**1.9.1.2 Besluitvorming** De onderzoeksvragen vertalen zich in de volgende nullhypotheses:

1. Associatie tussen bloeddruk en leeftijd bij de vrouwen?

$$H_0 : \beta_{\text{Age}} = 0 \text{ vs } H_1 : \beta_{\text{Age}} \neq 0$$

2. Associatie tussen bloeddruk en leeftijd bij de mannen?

$$H_0 : \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} = 0 \text{ vs } H_1 : \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} \neq 0$$

3. Is de Associatie tussen bloeddruk en leeftijd verschillend bij mannen en vrouwen?

$$H_0 : \beta_{\text{Age:Gendermale}} = 0 \text{ vs } H_1 : \beta_{\text{Age:Gendermale}} \neq 0$$

- We kunnen onderzoeksvraag 1 en 3 onmiddelijk toetsen o.b.v. de model output.
- Onderzoeksvraag 2 is echter een lineaire combinatie van twee parameters.

$$\begin{aligned}
L\boldsymbol{\beta} &= 0 \\
\begin{bmatrix} 0 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_g \\ \beta_a \\ \beta_{a:g} \end{bmatrix} &= 0 \\
0\beta_0 + 0\beta_g + 1\beta_a + 1\beta_{a:g} &= 0 \\
\beta_a + \beta_{a:g} &= 0
\end{aligned}$$

- Bovendien is er ook het probleem dat we meerdere toetsen nodig hebben om de associaties te bestuderen.

1. We toetsen eerste de omnibus hypothese dat er geen associatie is tussen leeftijd en de bloeddruk.

$$H_0 : \beta_{\text{Age}} = \beta_{\text{Age}} + \beta_{\text{Age:Gendermale}} = \beta_{\text{Age:Gendermale}} = 0$$

- Dat vereenvoudigt zich tot het toetsen dat

$$H_0 : \beta_{\text{Age}} = \beta_{\text{Age:Gendermale}} = 0$$

- Wat we kunnen evalueren door twee modellen te vergelijken. Een model met enkel het gender effect en volledige model met Gender, Age en Gender x Age interactie.

2. Als we deze hypothese kunnen verwerpen voeren we posthoc analyses uit voor elk van de 3 contrasten.

De posthoc testen kunnen we opnieuw uitvoeren a.d.h.v. het multcomp pakket.

```
bpPosthoc <- glht(mBp3, linfct=c(
  "Age = 0",
  "Age + Age:Gendermale = 0",
  "Age:Gendermale = 0")
  )
bpPosthoc %>% summary

bpPosthocBI <- bpPosthoc %>% confint
bpPosthocBI
```

### 1.9.2 Testosteron concentratie bij volwassen mannen in de NHANES studie

Veronderstel dat we wensen te bestuderen of er een associatie is leeftijd en gewicht van volwassen mannen in de Amerikaanse populatie op de de testosteron concentratie.

```
modTes <- lm(formula = Testosterone ~ Age * Weight, data = NHANES %>%
    filter(Age > 18 & Gender == "male" & !is.na(Testosterone) &!is.na(Weight) & !is.na(Age)))
```

- Er is geen significante interactie. Het aanvaarden van de nulhypothese is een zwakke conclusie. Daarom kunnen we opteren om de interactie in het model te laten.
- Merk op dat de hoofdeffecten geen zinvolle interpretatie hebben!
- Omdat er geen significante interactie is, is het zinvol om een uitspraak te doen over de associatie van gewicht en testosteron, en leeftijd en testosteron.

We kunnen hierover een uitspraak doen door te marginaliseren over alle leeftijden (gewichten) van de mannen in het experiment.

$$\frac{\sum_{i=1}^{n}(\beta_a + \beta_{w:a}X_w)}{n} = \beta_a + \beta_{w:a}\bar{X}_w$$

$$\frac{\sum_{i=1}^{n}(\beta_w + \beta_{w:a}X_a)}{n} = \beta_w + \beta_{w:a}\bar{X}_a$$

## 1.10 Factoriële proeven

### 1.10.1 Data

48 ratten werden at random toegewezen aan

- 3 giffen (I,II,III) and
- 4 behandelingen (A,B,C,D),

en,

- de overlevingstijd werd opgemeten (eenheid: 10 h)

We modelleren de "snelheid van sterven" met een hoofdeffect voor gif en behandeling en een gif $\times$ behandeling interactie.

$$\begin{aligned} y_i \quad = \quad & \beta_0 + \\ & \beta_{II} x_{iII} + \beta_{III} x_{iIII} + \\ & \beta_B x_{iB} + \beta_C x_{iC} + \beta_D x_{iD} + \\ & \beta_{II:B} x_{iII} x_{iB} + \beta_{II:C} x_{iII} x_{iC} + \beta_{II:D} x_{iII} x_{iD} + \\ & \beta_{III:B} x_{iIII} x_{iB} + \beta_{III:C} x_{iIII} x_{iC} + \beta_{III:D} x_{iIII} x_{iD} + \epsilon_i \end{aligned}$$

met $i = 1, \ldots, n$, $n = 48$, en, $x_{iII}$, $x_{iIII}$, $x_{iB}$, $x_{iC}$ en $x_{iD}$ dummy variabelen voor respectievelijk gif II, III, behandeling B, C, en D.

##$ Inferentie

Een interactie tussen gif en behandeling impliceert dat we het effect van het type gif afzonderlijk moeten bestuderen voor elke behandeling:

1. Voor behandeling A moeten we dan volgende nulhypotheses toetsen:

- II-I: $H_0 : \beta_{II} = 0$
- III-I: $H_0 : \beta_{III} = 0$
- III-II: $H_0 : \beta_{III} - \beta_{II} = 0$

2. Voor behandeling B:

- II-I: $H_0 : \beta_{II} + \beta_{II:B} = 0$
- III-I: $H_0 : \beta_{III} + \beta_{III:B} = 0$
- III-II: $H_0 : \beta_{III} + \beta_{III:B} - \beta_{II} - \beta_{II:B} = 0$

3. Voor behandeling C:

- II-I: $H_0 : \beta_{II} + \beta_{II:C} = 0$
- III-I: $H_0 : \beta_{III} + \beta_{III:C} = 0$
- III-II: $H_0 : \beta_{III} + \beta_{III:C} - \beta_{II} - \beta_{II:C} = 0$

4. Voor behandeling D:

- II-I: $H_0 : \beta_{II} + \beta_{II:D} = 0$
- III-I: $H_0 : \beta_{III} + \beta_{III:D} = 0$

- III-II: $H_0 : \beta_{III} + \beta_{III:D} - \beta_{II} - \beta_{II:D} = 0$

Hetzelfde geldt wanneer we het effect van de behandeling bestuderen:

1. Voor gif I toetsen we dan nulhypothese

- B-A: $H_0 : \beta_B = 0$
- C-A: $H_0 : \beta_C = 0$
- D-A: $H_0 : \beta_D = 0$
- C-B: $H_0 : \beta_C - \beta_B = 0$
- D-B: $H_0 : \beta_D - \beta_B = 0$
- D-C: $H_0 : \beta_D - \beta_C = 0$

2. Gif II

- B-A: $H_0 : \beta_B + \beta_{II:B} = 0$
- C-A: $H_0 : \beta_C + \beta_{II:C} = 0$
- D-A: $H_0 : \beta_D + \beta_{II:D} = 0$
- C-B: $H_0 : \beta_C + \beta_{II:C} - \beta_B - \beta_{II:B} = 0$
- D-B: $H_0 : \beta_D + \beta_{II:D} - \beta_B - \beta_{II:B} = 0$
- D-C: $H_0 : \beta_D + \beta_{II:D} - \beta_C - \beta_{II:C} = 0$

3. Gif III

- B-A: $H_0 : \beta_B + \beta_{III:B} = 0$
- C-A: $H_0 : \beta_C + \beta_{III:C} = 0$
- D-A: $H_0 : \beta_D + \beta_{II:D} = 0$
- C-B: $H_0 : \beta_C + \beta_{III:C} - \beta_B - \beta_{III:B} = 0$
- D-B: $H_0 : \beta_D + \beta_{III:D} - \beta_B - \beta_{III:B} = 0$
- D-C: $H_0 : \beta_D + \beta_{III:D} - \beta_C - \beta_{III:C} = 0$

In onze studie was de interactie echter niet significant.

- Coventioneel Analyse met additief model zonder interactie
- Alternatief: Analyse met model met interactie waarbij we

1. de effectgrootte voor de pairsgewijze vergelijkingen tussen de verschillende giffen (II-I, III-I en III- II) te schatten door ze uit te middelen over alle behandelingen (A, B, C, en D), en,
2. de effectgrootte voor de pairsgewijze vergelijkingen tussen de verschillende behandelingen (B-A, C-A, D-A, C-B, D-B en D-C) te schatten door ze uit te middelen over alle giffen (I, II, III).

Dat zou ons gelijkaardige schattingen van de effectgroottes moeten geven als deze voor het additieve model waarbij we de interactie term uit het model hadden geweerd.

B.v. voor gif III vs gif II zou dat in volgende contrast resulteren:

- III-II:

$$H_0 : \begin{array}{r} \frac{\beta_{III} - \beta_{II}}{4} + \\ \frac{\beta_{III} + \beta_{III:B} - \beta_{II} - \beta_{II:B}}{4} + \\ \frac{\beta_{III} + \beta_{III:C} - \beta_{II} - \beta_{II:C}}{4} + \\ \frac{\beta_{III} + \beta_{III:D} - \beta_{II} - \beta_{II:D}}{4} \end{array} = 0$$

$$H_0 : \beta_{III} + \frac{1}{4} \times \beta_{III:B} + \frac{1}{4} \times \beta_{III:C} + \frac{1}{4} \times \beta_{III:D} - \beta_{II} - \frac{1}{4} \times \beta_{II:B} - \frac{1}{4} \times \beta_{II:C} - \frac{1}{4} \times \beta_{II:D} = 0$$

# 2 Power, steekproefgrootte en andere design aspecten.

## 2.1 Variantie schatter?

$$\hat{\mathbf{\Sigma}}_{\hat{\boldsymbol{\beta}}} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

De onzekerheid op model parameters hangt dus af van de residuele variabiliteit en het de proefopzet!

- Hoe groter $\mathbf{X}^T\mathbf{X}$ hoe meer informatie het experiment zal hebben over de model parameters en hoe kleiner hun variantie en standaard errors!
- Factoriële designs?
- Designs with continue predictoren?

---

De effectgrootte van interesse is typisch een lineaire combinatie van de model parameters,

$$l_0 \times \beta_0 + l_1 \times \beta_1 + ... + l_{p-1} \times \beta_{p-1} = \mathbf{L}^T\boldsymbol{\beta}$$

De nulhypothese van onze test kan dan worden geschreven als

$$H_0 : \mathbf{L}^T\boldsymbol{\beta} = 0$$

vs de alternatieve hypothese

$$H_0 : \mathbf{L}^T\boldsymbol{\beta} \neq 0$$

En het bewijs in het experiment tegen $H_0$ kan worden gekwantificeerd met de t-test statistiek:

$$t = \frac{\mathbf{L}^T\hat{\boldsymbol{\beta}} - 0}{\text{se}_{\mathbf{L}^T\hat{\boldsymbol{\beta}}}}$$

die een t-distributie volgt met n-p vrijheidsgraden onder de nul hypothese als alle aannames van het model geldig zijn.

De kracht van de toets is dan

$$P(p < 0.05|H_1)$$

en hangt af van

- de werkelijke effectgrootte in de populatie $\mathbf{L}^T\boldsymbol{\beta}$.
- Het aantal observaties: SE en df van t-test.
- Keuze van de designpunten
- Keuze van significantie-niveau $\alpha$.

We kunnen de power schatten met simulaties als

1. voldaan is aan de aannamens van het model ==> piloot experiment
2. standaard deviatie is gekend ==> piloot experiment
3. de echte effectgrootte in de populatie ==> die is ongekend daarom moet een minimale effectgrootte worden gespecifieerd die we wensen op te pikken in het nieuwe experiment.
4. steekproef is gespecifieerd ==> Keuze van grootte van de steekproef en designpunten

# 3 Randomized complete block design (RCB)

- Observaties ingedeeld in blokken
- In elk blok wordt elke behandeling geëvalueerd.

- RBC beperkt de randomisatie: de behandelingen worden binnen blokken gerandomiseerd.
- In de analyse moet voor blokken worden gecorrigeerd
- Een gepaard design is het meest eenvoudige RCB: met blokgrootte 2.
- Effecten kunnen binnen block worden geschat:
- In muisvoorbeeld waarin proteine expressie (variable `intensity`) in verschillende celtypes (variabele `celltype`) wordt gemeten voor elke muis (variable `mouse` is block)

```
lm(intensity~celltype + mouse)
```

De winst in power van een randomized complete block design is dus een afweging tussen de variabiliteit die kan worden verklaart met het blokeffect en het verlies in vrijheidsgraden.

Als je je de formule voor de variantie covariantie matrix van de parameter schatters bijhaalt zien we

$$\hat{\mathbf{\Sigma}}^2_{\hat{\boldsymbol{\beta}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\hat{\sigma}^2$$

dat het RCB een impact heeft op

- $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ alsook op
- $\hat{\sigma}^2$ van de residuen!

→ We kunnen de variantie in de expressie tussen de proefdieren/blokken isoleren uit onze analyse!

→ Dat reduceert de variantie van de residuen en leidt tot een toename in power als de variabiliteit tussen muizen/blokken groot is.

Merk op dat,

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p} = \frac{SSTot - SSM - SSCT}{n-p}$$

- Blokken is dus nuttig als de reductie in SSE groot is in vergelijking met het verlies in vrijheidsgraden.
- Dus als SSM een groot deel in van de totale variabiliteit kan verklaren.

Verder heeft het verlies in vrijheidsgraden ook een impact op de t-verdeling die zal worden gebruikt voor inferentie, die hierdoor bredere staarten zal hebben.