BIAS

The beast

Klinkenberg

24 Sep 2014

Wat is BIAS

Things that lead us to the wrong conclusions (Field)

$$outcome_i = model_i + error_i$$

$$model_i = b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni}$$

- X = predictor variables
- b = parameters

BIAS

Verkeerde conclussies over:

- Parameters b_i
- Standaard error en betrouwbaarheidsintervallen
- Toetsingsgrootheden en *p-waarden*

means $\rightarrow$ SE $\rightarrow$ CI

SE $\rightarrow$ toetsingsgrootheid $\rightarrow$ *p-waarden*

The beasts:

- Uitbijters (Outliers)
- Assumpties

Voorbeeld

Eigen IQ schatting van mannen en vrouwen. Wat we willen is een uitspraak doen over het verschil in de populatie. Niet enkel deze sample. We willen een inferentie maken (Vandaar de term inferentiële statistiek).

```
data = read.csv("IQ.csv")
data[12:17,]
```

```
##          Timestamp IQ.van.je.buur Eigen.IQ sekse
## 12 20/09/13 11:06             145      120     0
## 13 20/09/13 11:06             125      125     0
## 14 20/09/13 11:06             120      110     0
## 15 20/09/13 11:06             123      125     1
## 16 20/09/13 11:06             145      125     1
## 17 20/09/13 11:06             120      120     0
```

Voorbeeld

We zien dat de vrouwen als 0 gecodeerd zijn en mannen als 1. We kunnen dan het regressie model invullen voor dit onderzoek.

$$\text{Schatting eigen IQ}_i = b_0 + b_1 Sekse_i + error_i$$

```
aggregate(Eigen.IQ ~ factor(sekse), data, mean)
```

```
##   factor(sekse) Eigen.IQ
## 1             0    120.7
## 2             1    121.8
```

We kunnen nu de b's berekenen: $b_0 = 120.713$ en $b_1 = 1.0918$

Voorbeeld

$$\text{Schatting eigen IQ}_i = b_0 + b_1 Sekse_i + error_i$$

Als we dan het regressie model invullen, krijgen we:

```
##    Eigen.IQ   b.0   b.1 sekse   error
## 12      120 120.7 1.092     0  -0.713
## 13      125 120.7 1.092     0   4.287
## 14      110 120.7 1.092     0 -10.713
## 15      125 120.7 1.092     1   3.195
## 16      125 120.7 1.092     1   3.195
## 17      120 120.7 1.092     0  -0.713
```

De gemiddelden vormen dus indirect de parameters b's in dit regressie model. Deze b's zijn de schatters van de populatie β's.

Voorbeeld

En wat nou als deze gemiddelden niet zo goed zijn?

Bijvoorbeeld omdat er extreme uitbijters tussen zitten.

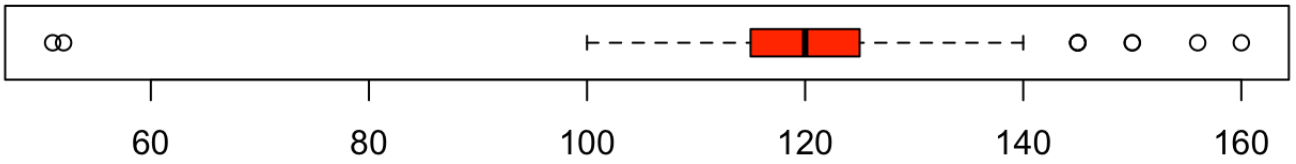
Voorbeeld

Zonder deze outliers ziet het er net wat anders uit.

```
##   factor(sekse) Eigen.IQ
## 1             0    120.3
## 2             1    121.4
```

```
##    Eigen.IQ   b.0   b.1 sekse   error
## 16      125 120.3 1.149     1  3.5789
## 17      120 120.3 1.149     0 -0.2718
## 18      115 120.3 1.149     0 -5.2718
## 19      125 120.3 1.149     1  3.5789
## 20      120 120.3 1.149     0 -0.2718
## 21      115 120.3 1.149     0 -5.2718
```

Uitbijters



Uitbijters kunnen grote invloed hebben op de gemiddelden.

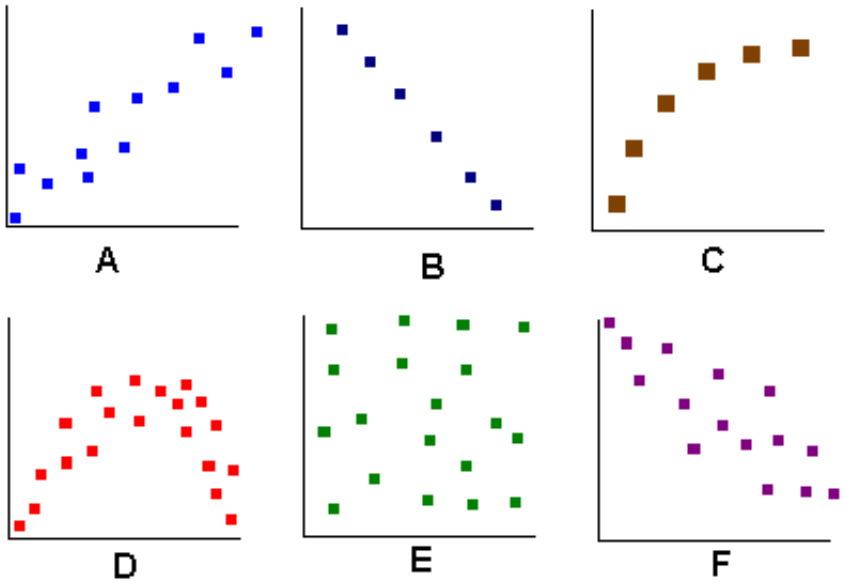
- Verwijderen op basis van boxplot.
- Verwijderen op basis van 3 standaard deviaties.
- Trimmed mean
- Winsorizing

Assumpties

- Additiviteit en lineairiteit
- Normaliteit
- Homoscedasticiteit/homogeniteit van variantie
- Onafhankelijkheid

Additiviteit en lineairiteit

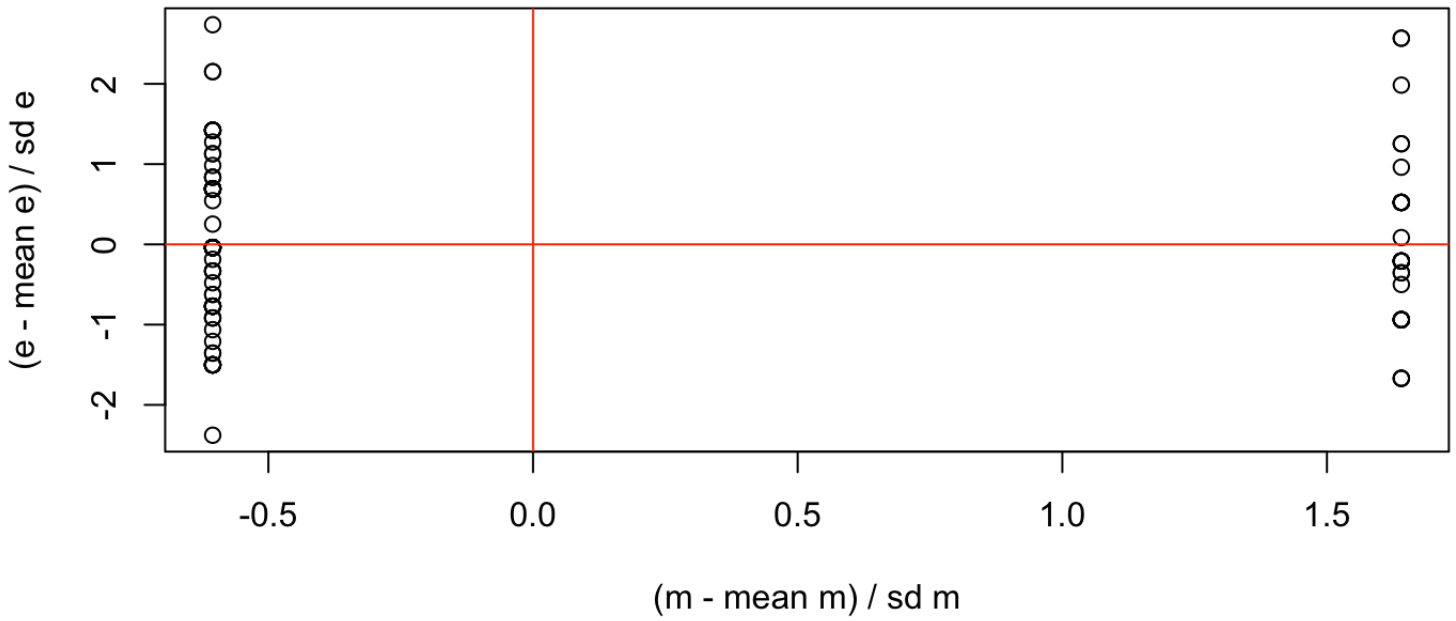
De afhankelijke variabele is in werkelijkheid lineair gerelateerd aan de predictoren.



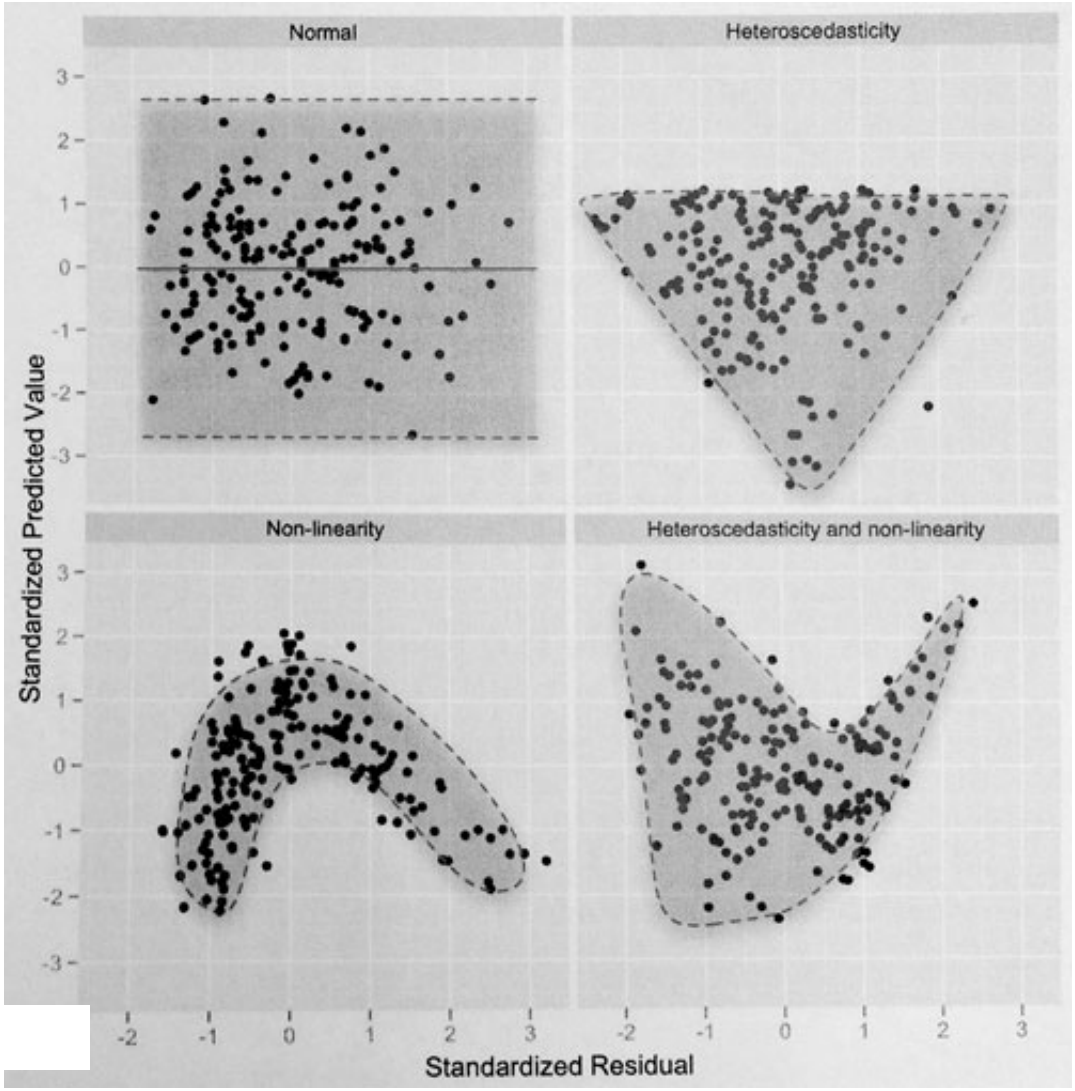
$$\text{MODEL}_i = b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_{ni}$$

Additiviteit en lineairiteit

Dit is te controleren door een plot te maken van de gestandaardiseerde error/residu en de gestandaardiseerde verwachte uitkomst/model.



Additiviteit en lineairiteit



Normaliteit

- Parameter schattingen b's
- Betrouwbaarheidsintervallen (SE * **1.196**)
- Nul hypothese toetsing
- Error

Het gaat niet om de normaliteit van de data maar van de populatie verdeling. Deze willen we testen aan de hand van de data.

Geen zorgen bij grote samples (Centrale limietstelling).

Normaliteit

Te bekijken met:

- Skewness en Kurtosis

Te toetsen met:

- Kolmogorov-Smirnof test
- Shapiro-Wilk test

Maar hoe groter de sample hoe kleiner de *p-waarde* bij gelijke toetsingsgrootheden. Dus dat bijt elkaar een beetje.

- transformatie van de outcome variable.

Homoscedasticiteit/homogeniteit van variantie

Van invloed op:

- Parameters b's
- NHT

De assumptie van de nul hypothese is dat de nul verdeling waar is. Dus bij verschillende samples uit die verdeling, laten we zeggen mannen en vrouwen op IQ, verwachten we dat de variantie van beide groepen identiek is. Anders zou onze assumptie niet gelden.

In algemene termen kunnen we dus zeggen dat op elk niveau van de predictorvariabele de varianties gelijk moeten zijn.

Homoscedasticiteit/homogeniteit

Onafhankelijkheid

De observaties die gedaan zijn, lees: de rijen in SPSS of de proefpersonen in je onderzoek moeten onafhankelijk van alkaar een reactie gegeven hebben op de outcome variable. Het antwoord van persoon B moet niet afhangen van die van pesoon A.

Vragen?