The beast of BIAS

Klinkenberg

28 sep 2017

Inhoud

- what is bias
- outliers
- assumptions
- additivity and linearity
- normality
- homoscedasticityhomogenity of variance
- independence

What 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

Wrong conclusions about:

- Parameters b_i
- Standaard error and confidence intervals
- Test statistics and p-values

means $\rightarrow$ SE $\rightarrow$ CI

SE $\rightarrow$ test statistics $\rightarrow$ p-values

The beasts

- Outliers
- Violations of assumptions

Example

IQ estimations of males and febecouse. We want to know the differences in the population not the sample. We therefore want to make an inference about the population, hence the name inferential statistics.

```
data = read.table("../../topics/t-test_independent/IQ.csv", sep = ' ', header
names(data)[3] <- "male"
data$male <- ifelse(data$male == "male" , 1, 0)
data[12:15,]
```

We can see that fameles are coded as 0 and males as 1. Such coding can be used in a linear regression equation.

$$\text{IQ you}_i = b_0 + b_1 male_i + error_i$$

```
means <- aggregate(IQ.you ~ factor(male), data, mean); means
```

```
##   factor(male)   IQ.you
## 1            0 123.6735
## 2            1 118.9821
```

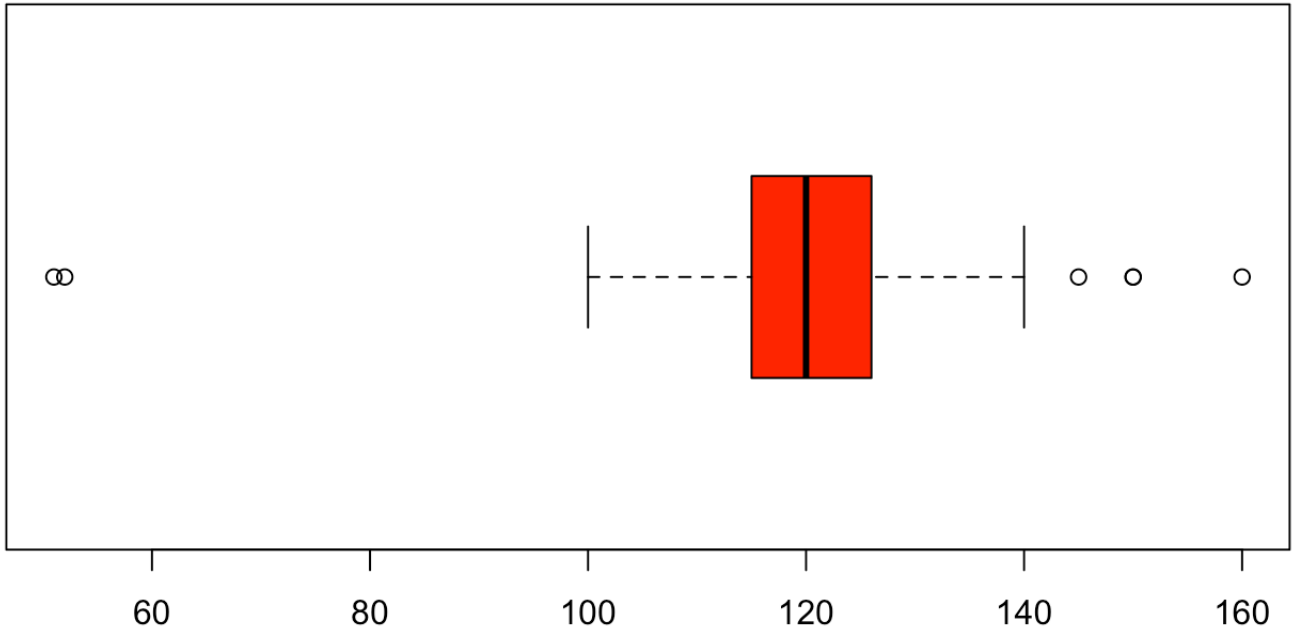
We can now calculate the b's: $b_0 = 123.67$ and $b_1 = -4.69$

$$\text{IQ you}_i = b_0 + b_1 male_i + error_i$$

If we apply this to the regression model we get:

```
##         b.0   b.1 male  model IQ.you  error
## [1,] 123.67 -4.69    0 123.67    120  -3.67
## [2,] 123.67 -4.69    1 118.98    120   1.02
## [3,] 123.67 -4.69    0 123.67    120  -3.67
## [4,] 123.67 -4.69    1 118.98    110  -8.98
## [5,] 123.67 -4.69    0 123.67    110 -13.67
## [6,] 123.67 -4.69    1 118.98    119   0.02
## [7,] 123.67 -4.69    1 118.98    128   9.02
## [8,] 123.67 -4.69    0 123.67    104 -19.67
```

The means indirectly represent the parameters b's in this regression model. These b's are the estimates of the population parameters β's.



t what if these means are not correct, because of an extreme outlier

Outliers

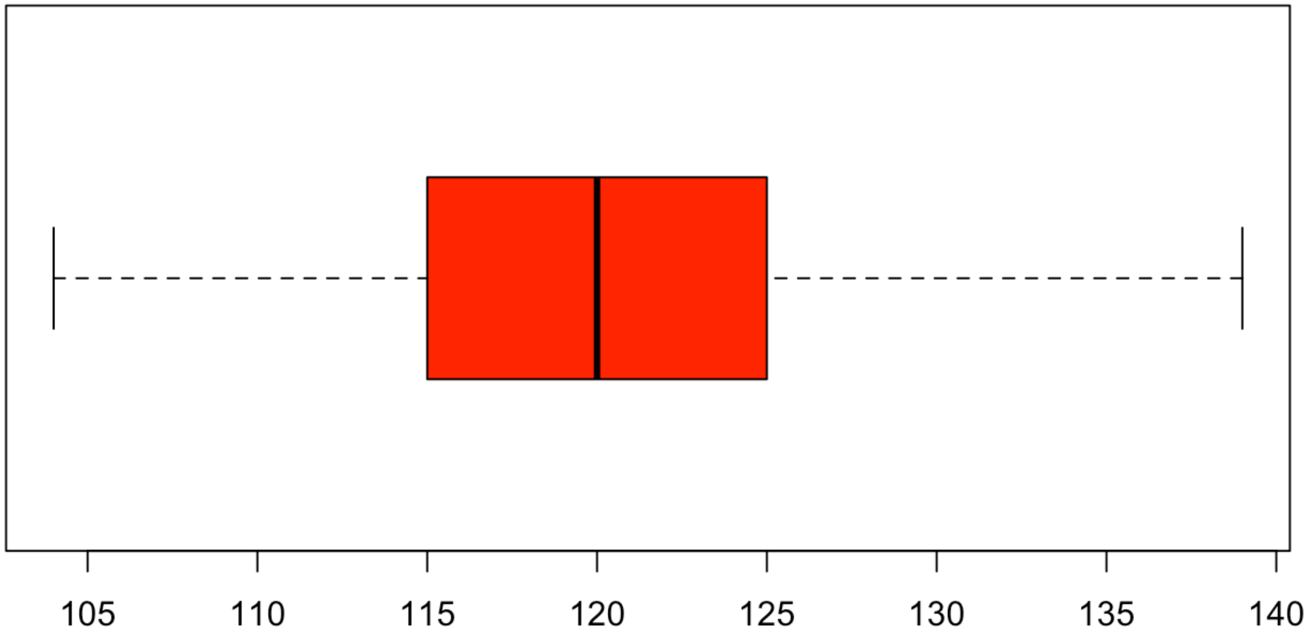
Outliers can have a huge impact on the estimations

- **Trim** Delete based on boxplot.
- **Trim** Delete based on 3 standard deviations.
- **Trim** Trimmed mean: Delete upper and lower percentages.
- **Winsorizing** Replace outliers with highest non outlier.

Without these outliers the results look a bit different.

```
##   factor(male)   IQ.you
## 1            0 121.3333
## 2            1 121.1224
```

```
##    IQ.you      b.0        b.1 male      error
## 12    125 121.3333 -0.2108844    1   3.877551
## 13    111 121.3333 -0.2108844    1 -10.122449
## 15    115 121.3333 -0.2108844    1  -6.122449
## 16    110 121.3333 -0.2108844    0 -11.333333
## 17    125 121.3333 -0.2108844    0   3.666667
## 18    139 121.3333 -0.2108844    0  17.666667
```

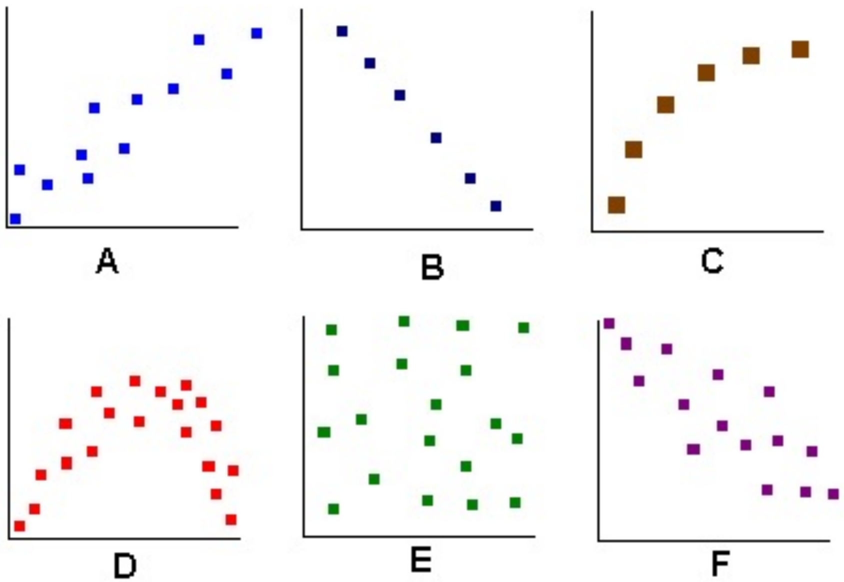

105
110
115
120
125
130
135
140

Assumptions

- Additivity and linearity
- Normality
- Homoscedasticity/homogenity of variance
- Independence

Additivity and linearity

The outcome variable is linearly related to the predictors.



relations

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

Additivity and linearity

We can check this by looking at the scatterplot of the predictors with the outcome variable.

Normality

- Parameter estimates b's
- Confidence intervals (SE * **1.196**)
- “Null hypothesis significance testing”
- Errors

Not the normality of the sample but the normality of the parameter β in the population. We will test this assumption based on the data, though with large samples the centrel limit theorem ensures that the parameters are bell shaped.

Centrel limit theorem

Normality

You can look at:

- Skewness and Kurtosis

We can test with:

- Kolmogorov-Smirnof test
- Shapiro-Wilk test

But, the bigger the sample the smaller the p-value at equal test statistic. So we are losing power at large samples.

- We can also transform the variable

Homoscedasticity/homogenity of variance

Influences:

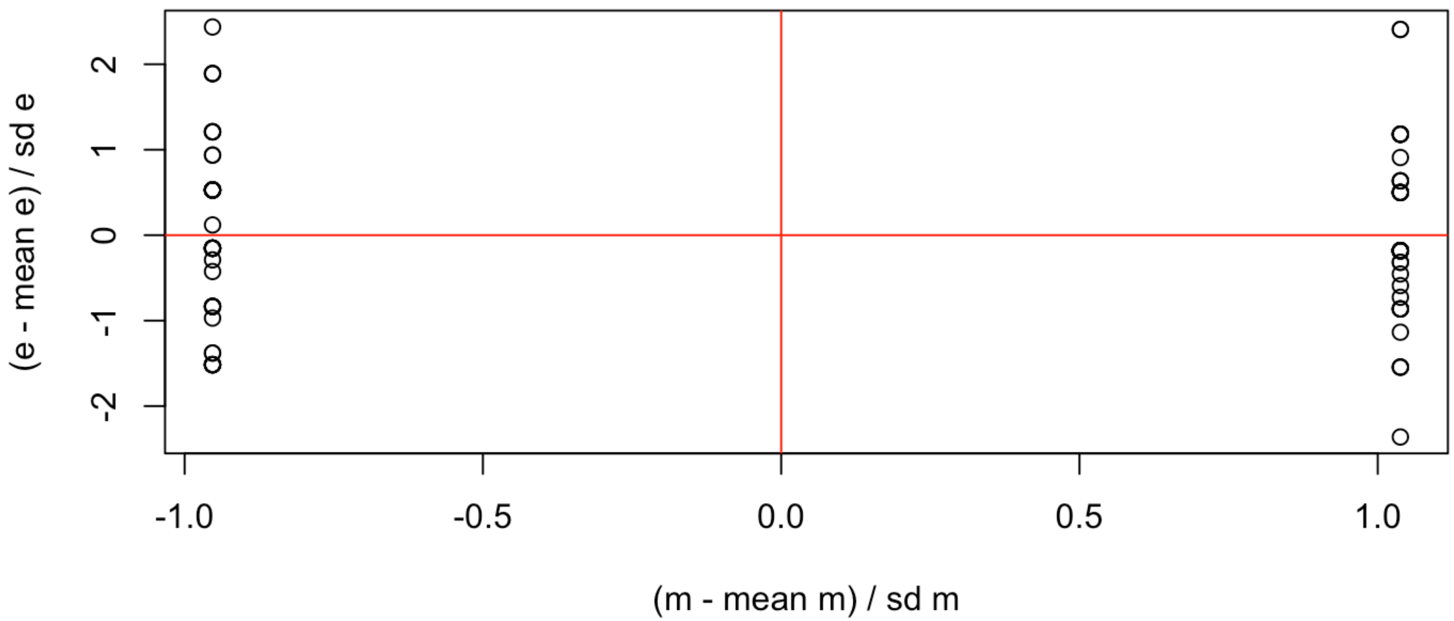
- Parameters b's
- NHT

The null hypothesis assumes the null distribution to be true. Therefore, different sampples from that districution should have equal variances. Otherwise the assumption could not hold.

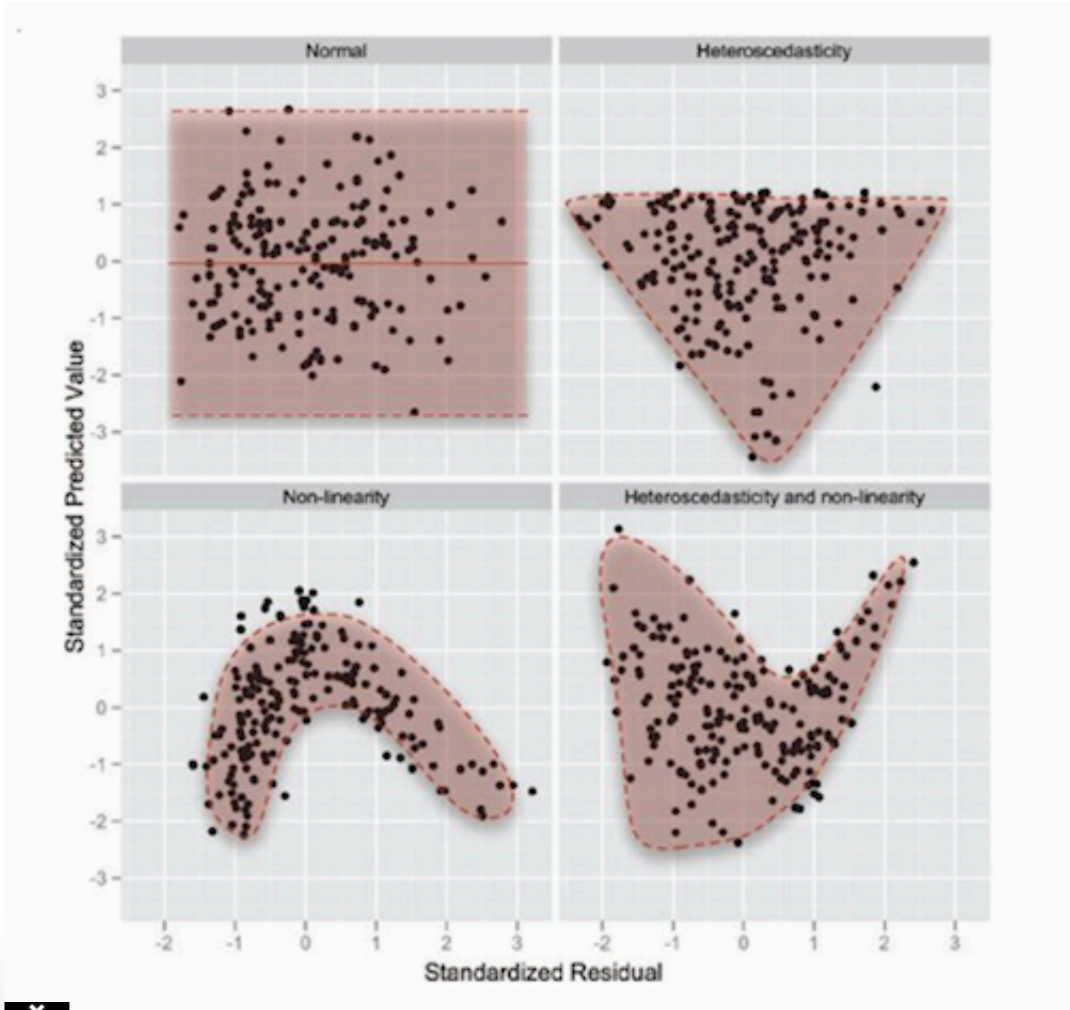
In general, we can say that on every value of the predictor variable the variances in the outcome variable should be equal.

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

We can check this by plotting the standardised error/resiual and the standardised expected outcome/model.



(e - mean e) / sd e
-2
-1
0
1
2
-1.0
-0.5
0.0
0.5
1.0
(m - mean m) / sd m



Lineairiteit

Independence

The observed outcome (rows in SPSS or participants in your research) should be independent from each other. The answer of person B should not depend on the answer of person A.

Whisper

- @shklinkenberg
- Klinkenberg
- S.Klinkenberg@UvA.nl
- ShKlinkenberg

END