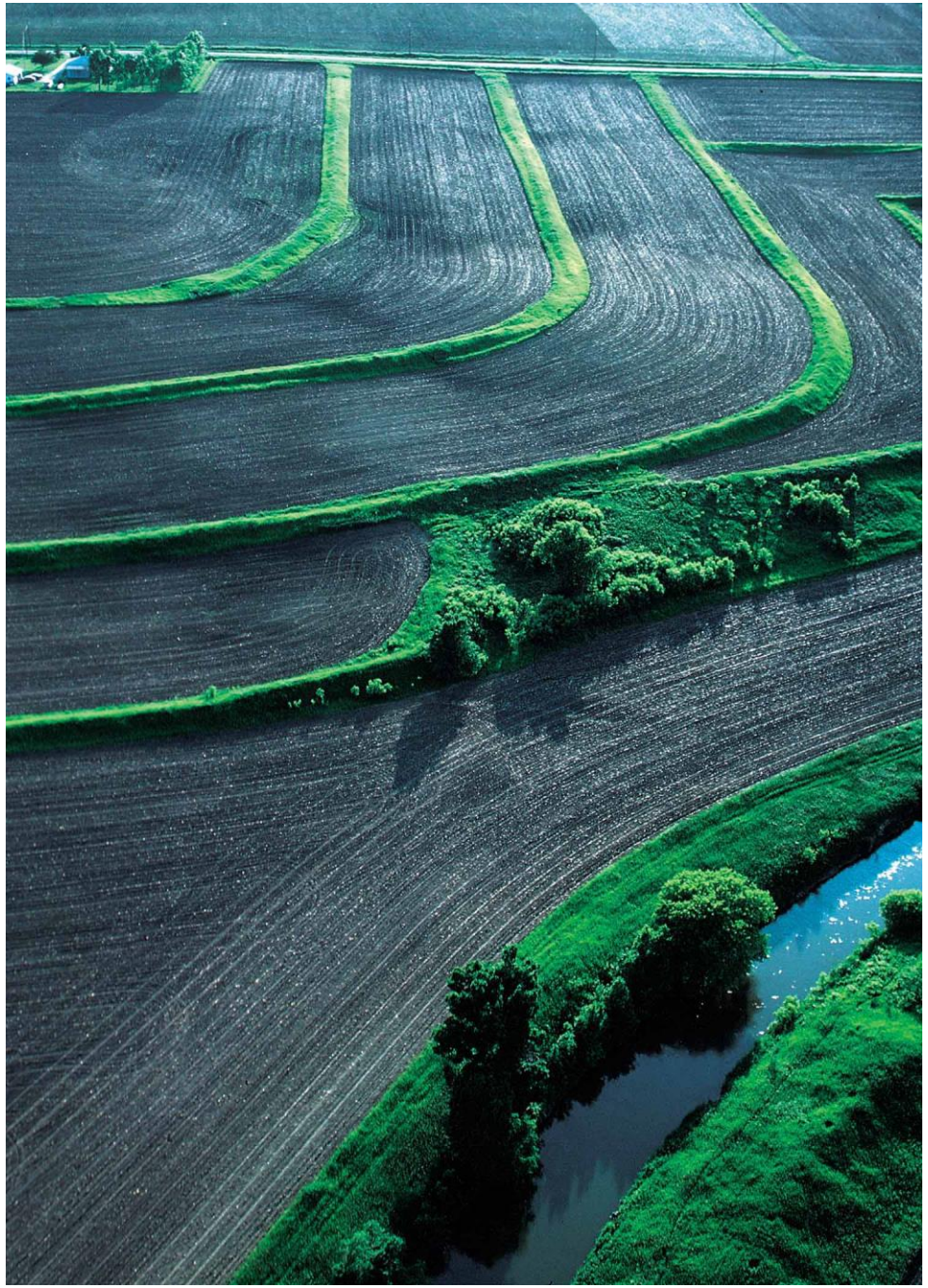


Photo by: Lynn Betts, USDA Natural Resources Conservation Service
https://commons.wikimedia.org/wiki/Agriculture#/media/File:TerracesBuffers.JPG

Examining Residuals for Validation and Added Confidence

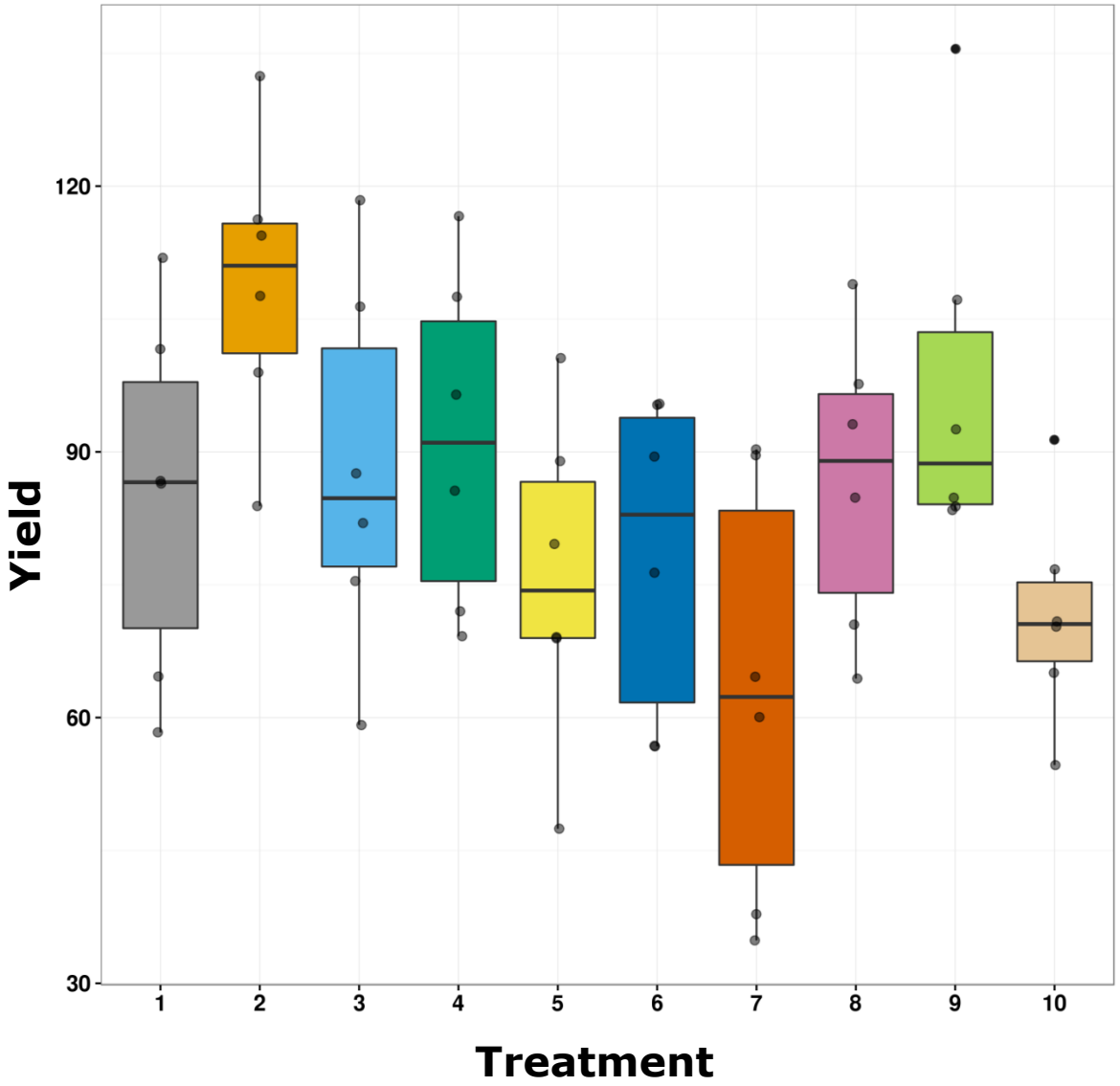
Rachel A. Hillmer, Ph.D.

Biostatistician, Koch Biological Solutions

11.4.2016

Data in this presentation

- Were generated using the statistical programming language, R.
- The statistical properties of the data (mean, standard deviation, distribution) resemble those of real field trial data. However, the effects shown are not those of actual biological products.



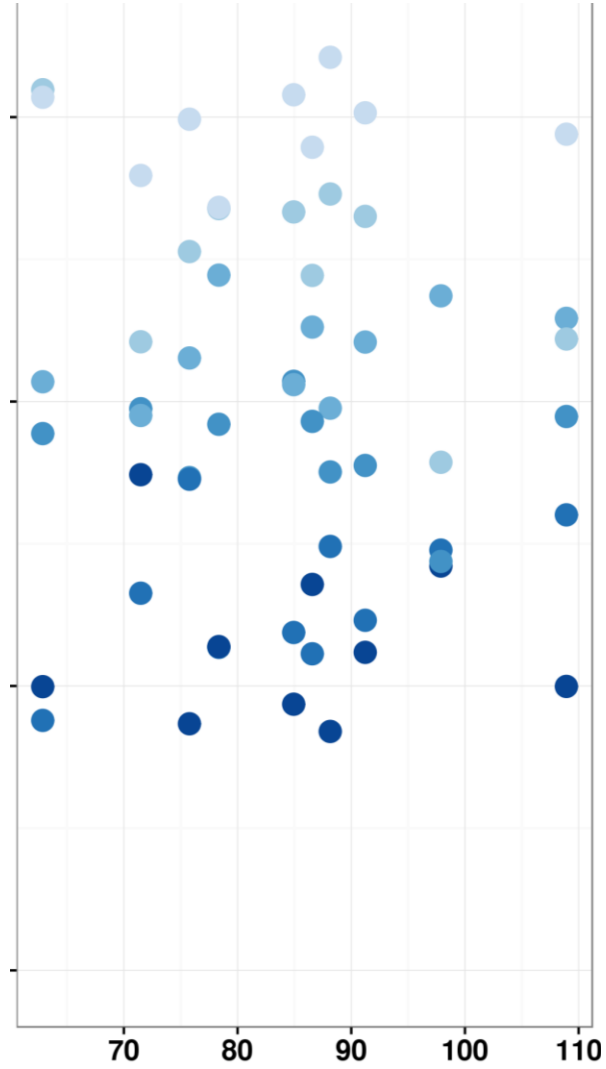
Statistics fails when statistical assumptions fail

Photo by: Daniel Leininger
https://commons.wikimedia.org/wiki/File:Old_Vault_Door.jpg

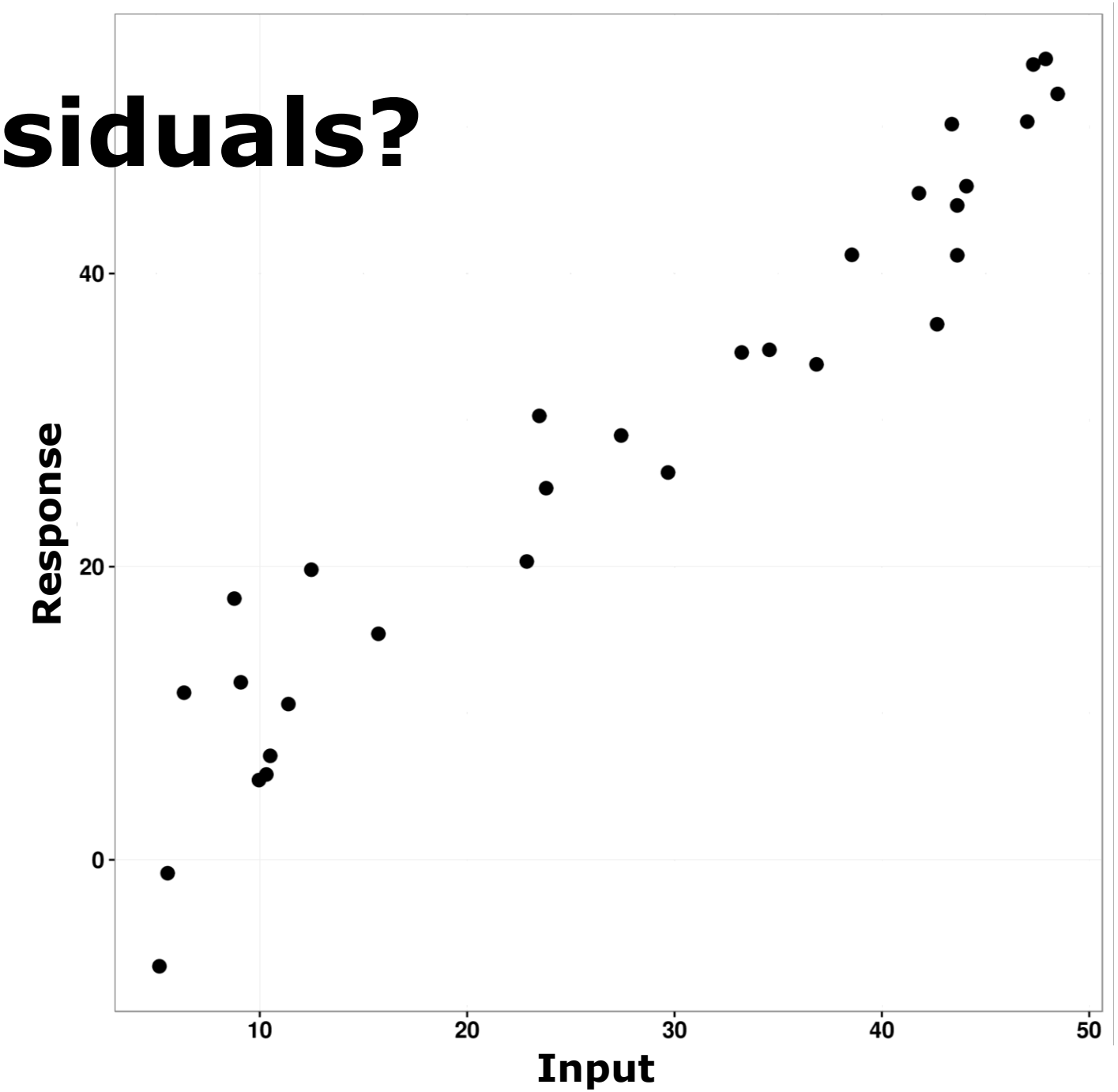
Residuals allow us to assess the validity of statistical assumptions

- Is the data normally distributed?
- Is the data heteroskedastic?
- Are there strong outliers?
- Are there additional experimental factors influencing the response variable?

- …and the residuals help us plan for next time.

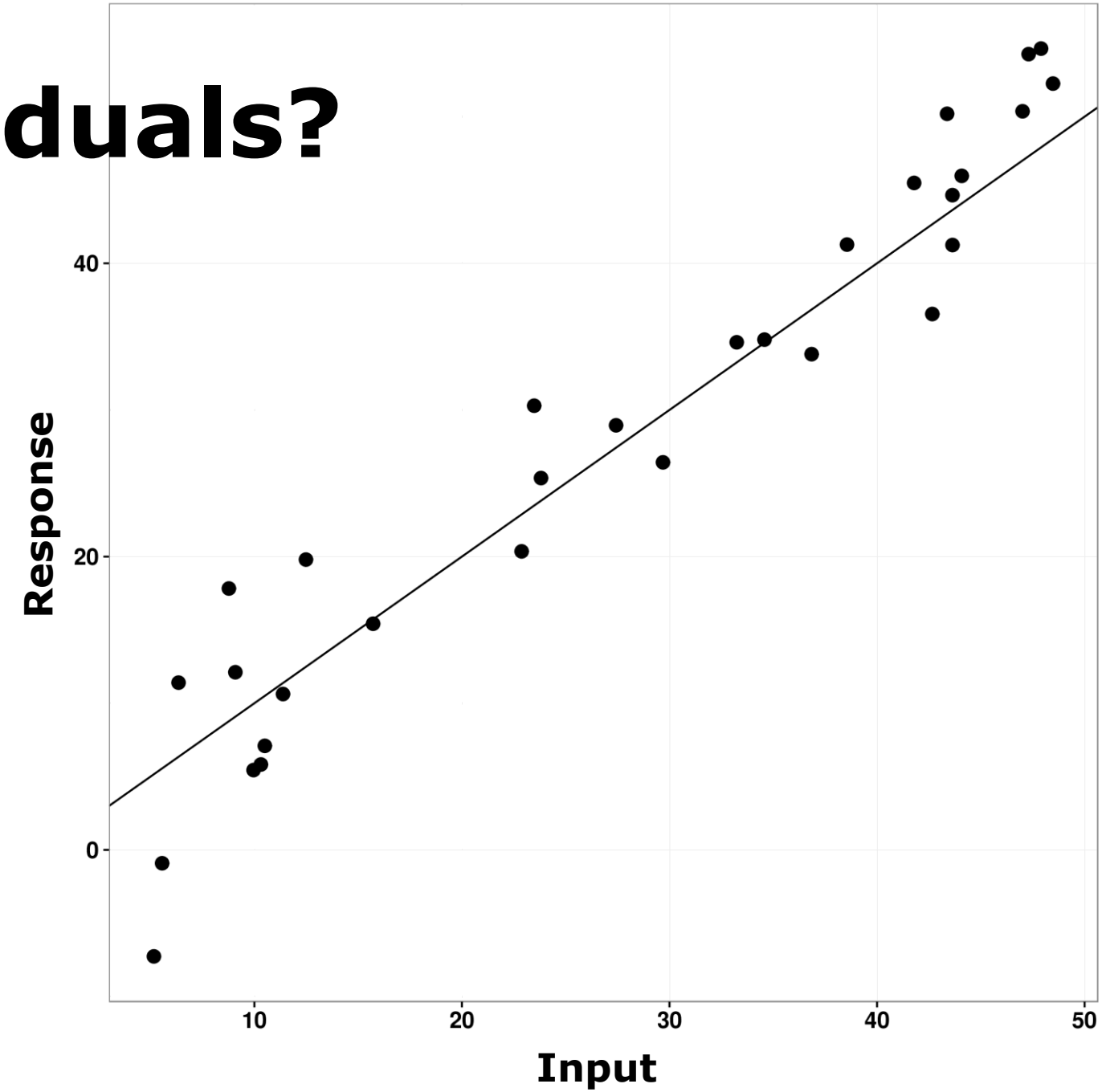


What are residuals?

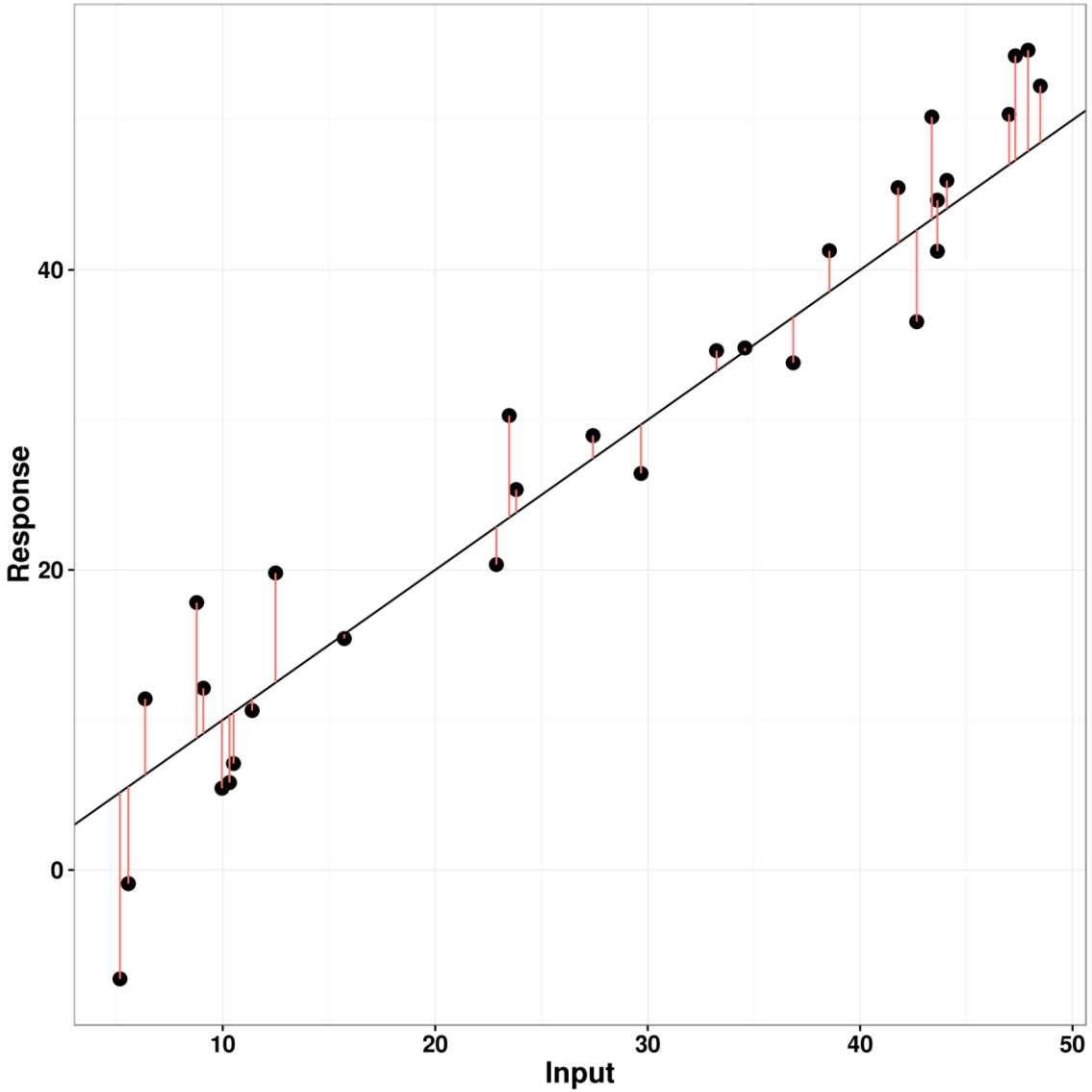


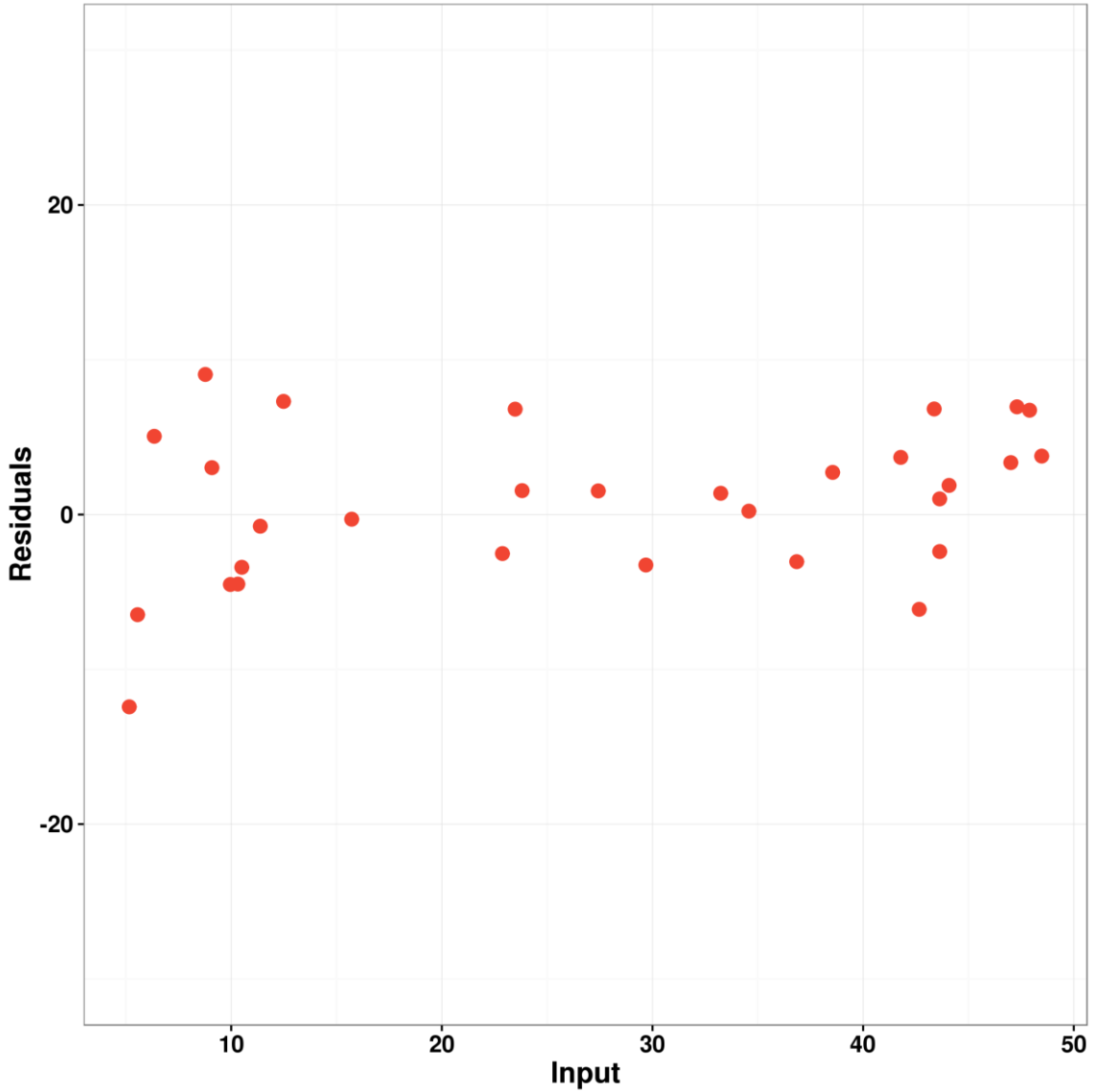
What are residuals?

- Residuals are the left-over variation when all known effects have been accounted for.



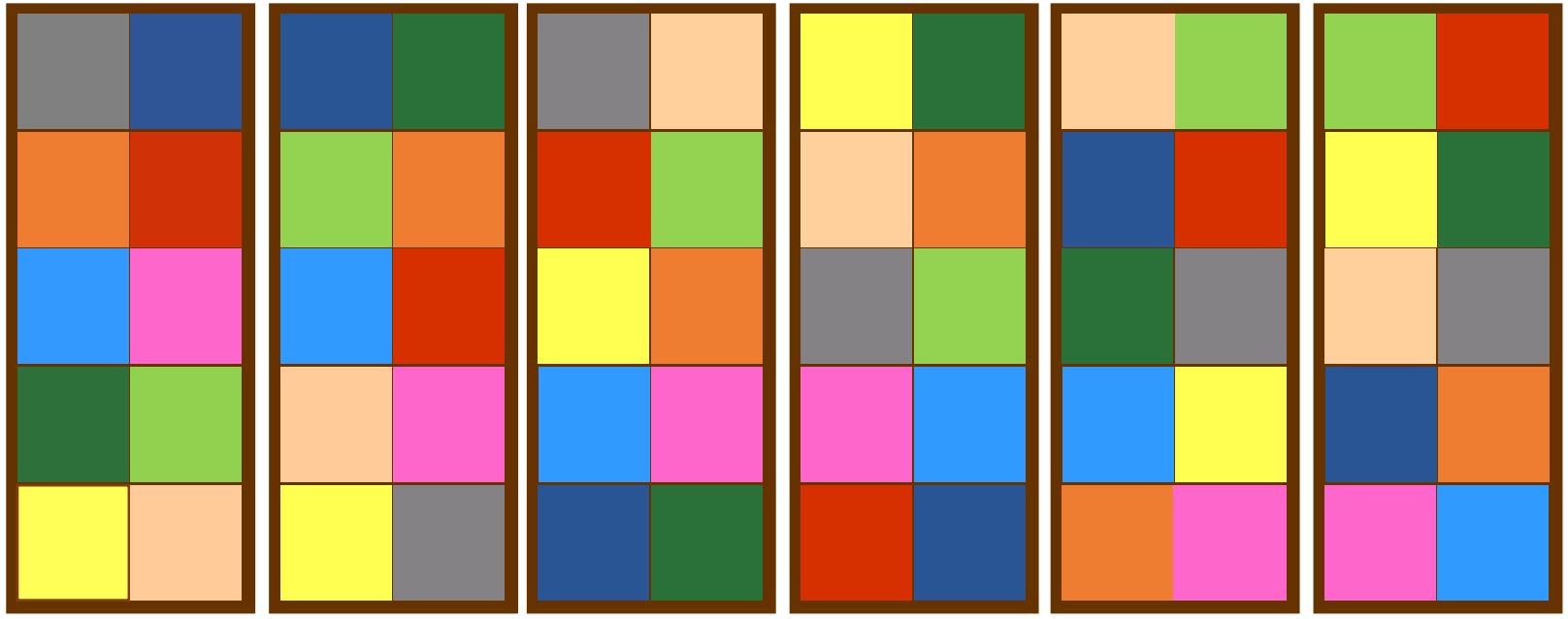
Residuals are left-over variation



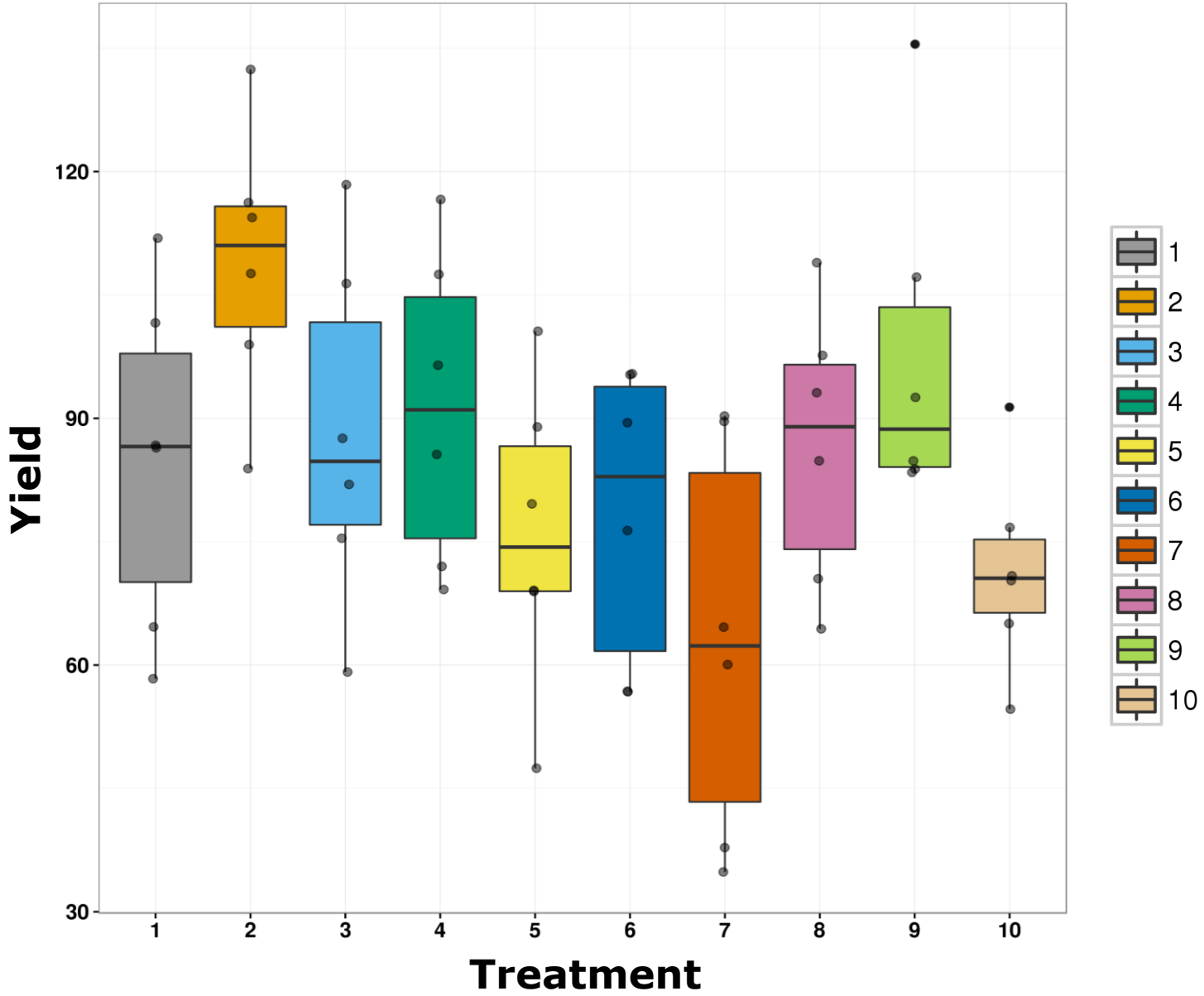


A more relevant example:

- Randomized complete block design, to deal with the major source of variability in field trials: soil.

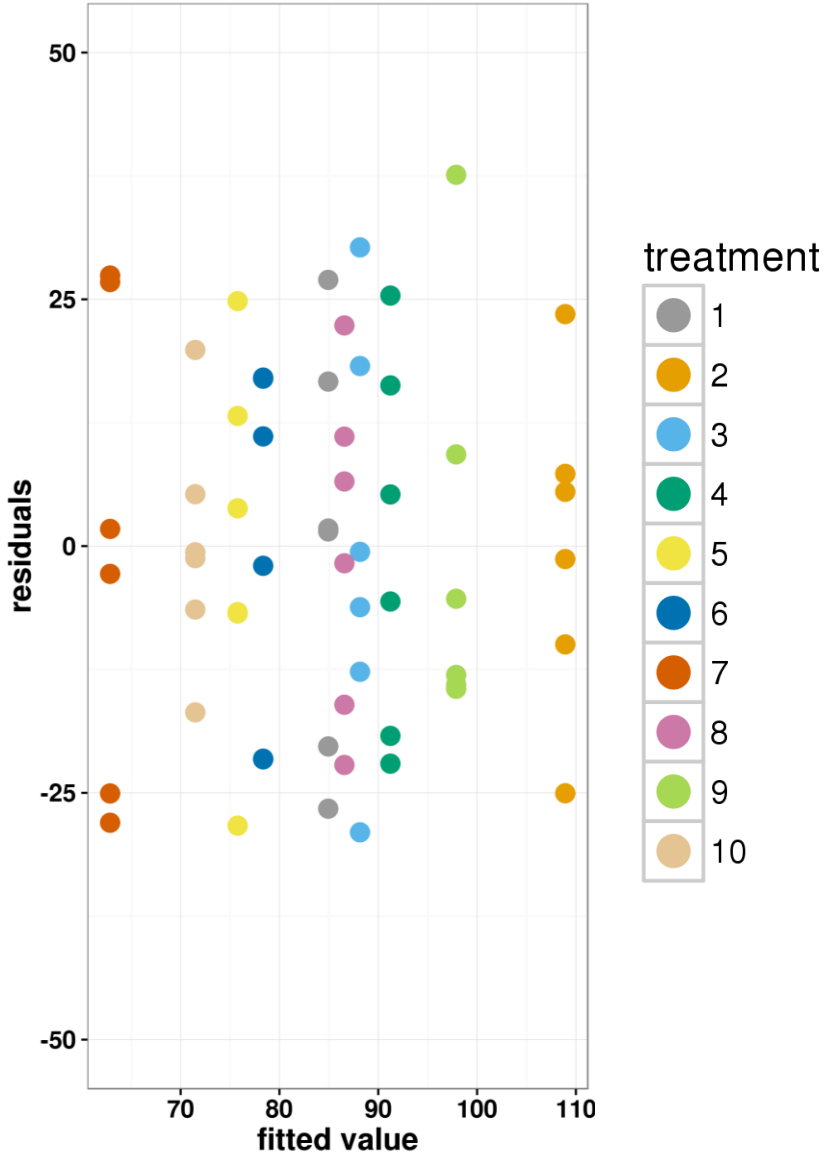


Yield by treatment

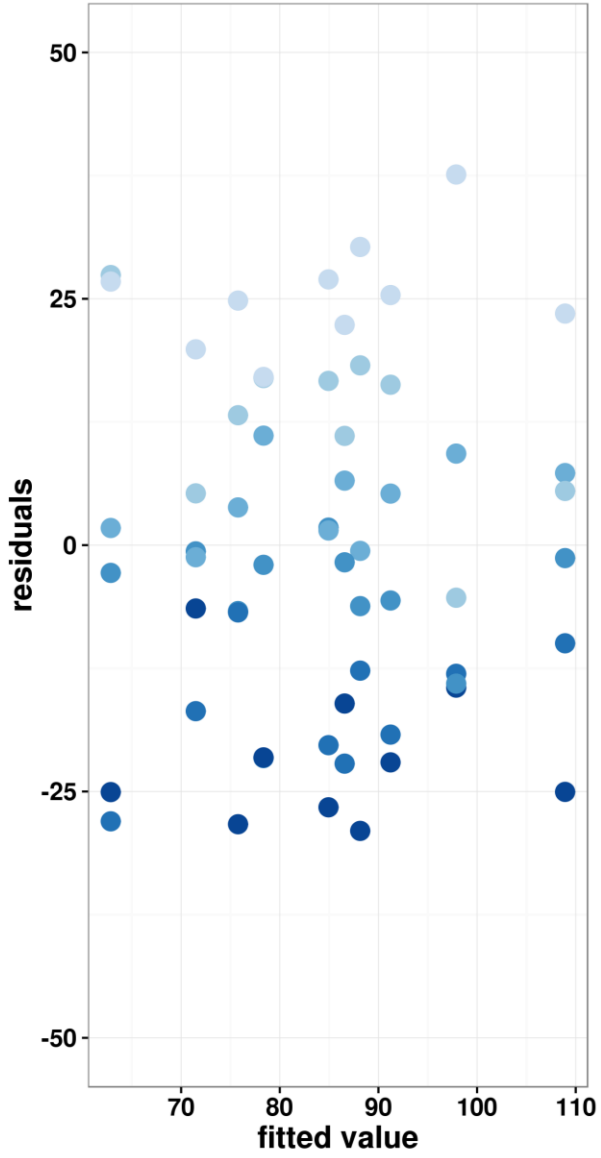


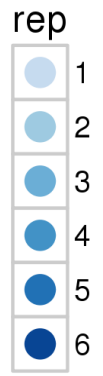
- Data were generated in R and do not represent plant responses to actual biologicals.

Residuals

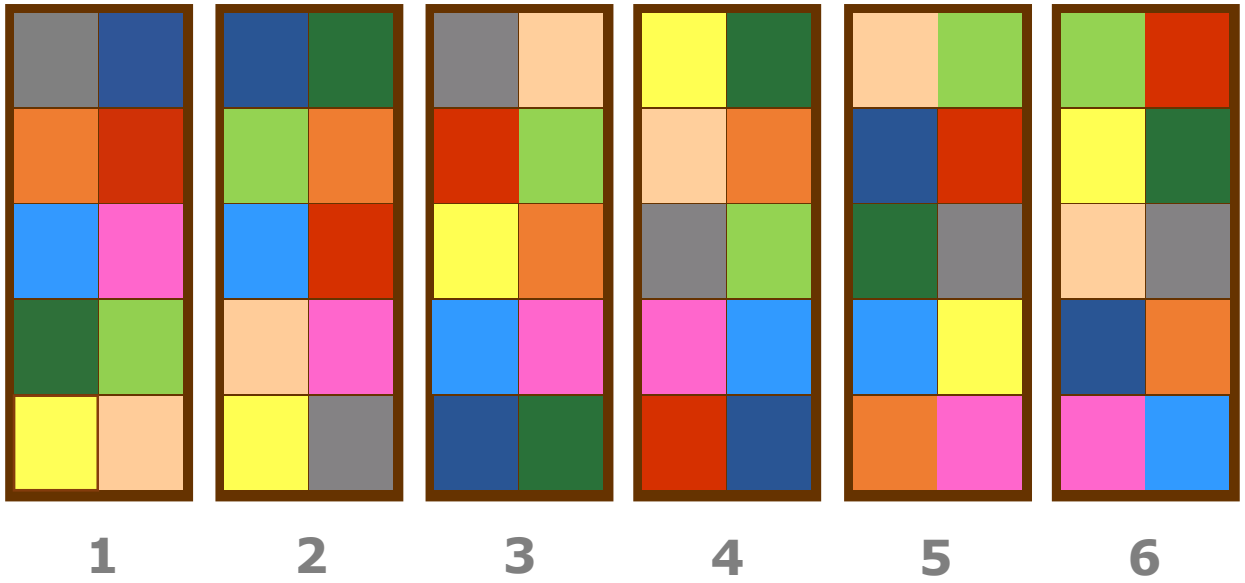

50
25
0
-25
-50
residuals
70
80
90
100
110
fitted value

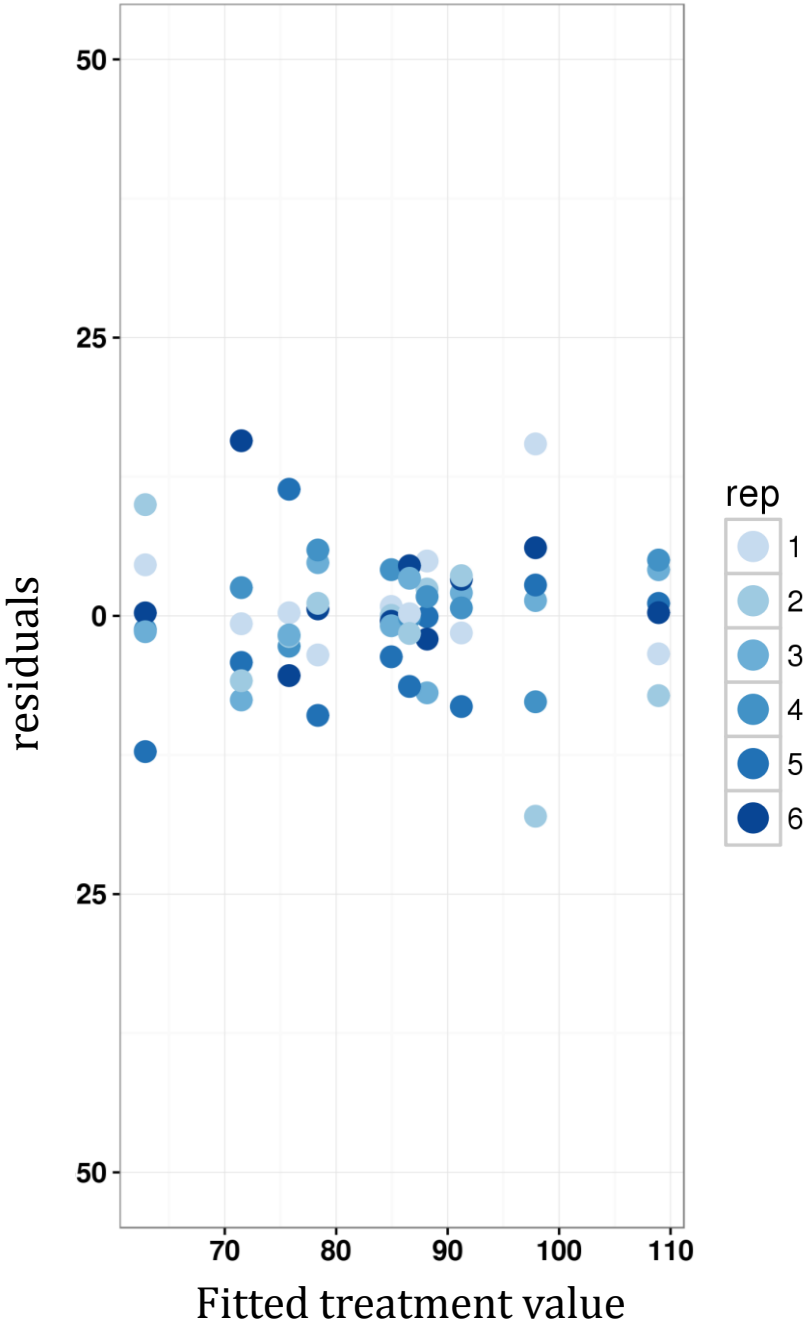
treatment
1
2
3
4
5
6
7
8
9
10


50
25
0
-25
-50
residuals
70
80
90
100
110
fitted value


rep
1
2
3
4
5
6

Residuals corrected for replicate effects



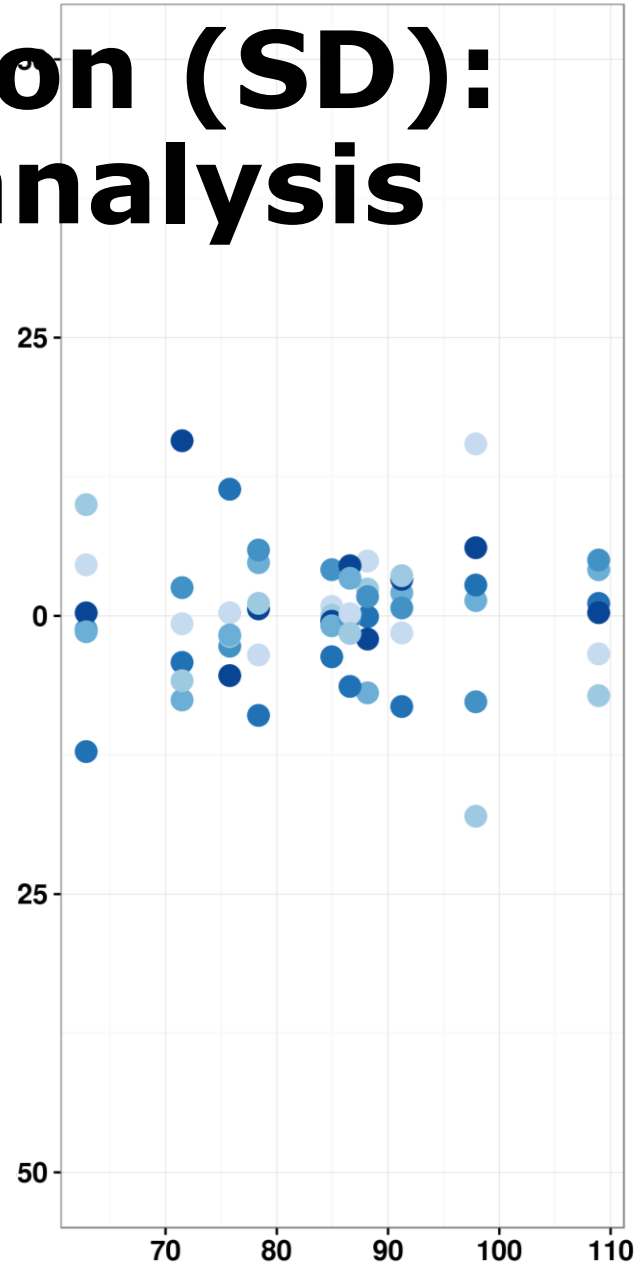


Residual standard deviation (SD): starting point for power analysis

Given:

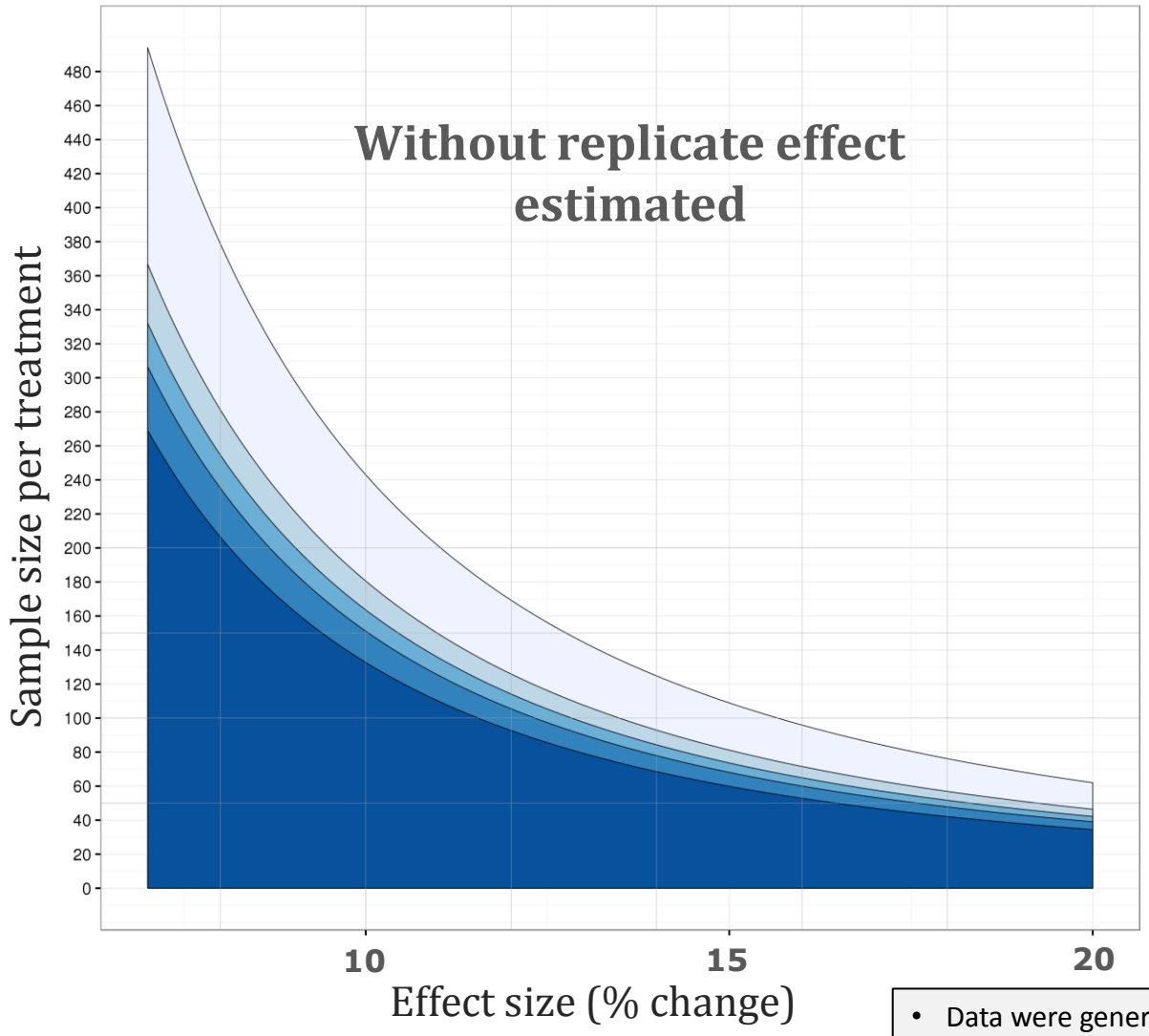
- Effect size
- Significance level (Type I error)
- Power (Type II error)
- Residual standard deviation

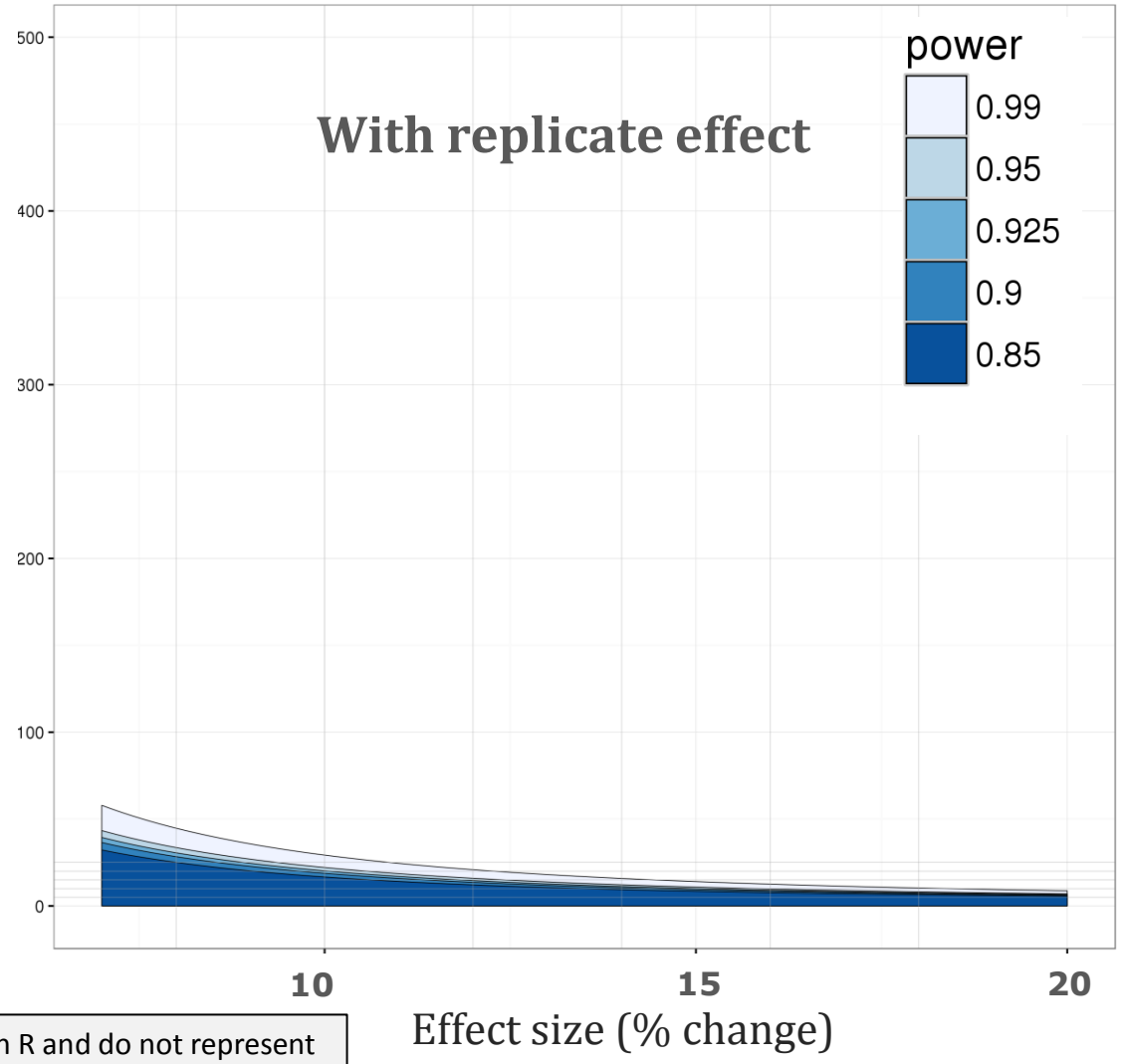
...Sample size can be calculated.



Power analysis visualization

P-value: 0.01
Adj. P-value ≈ 0.1

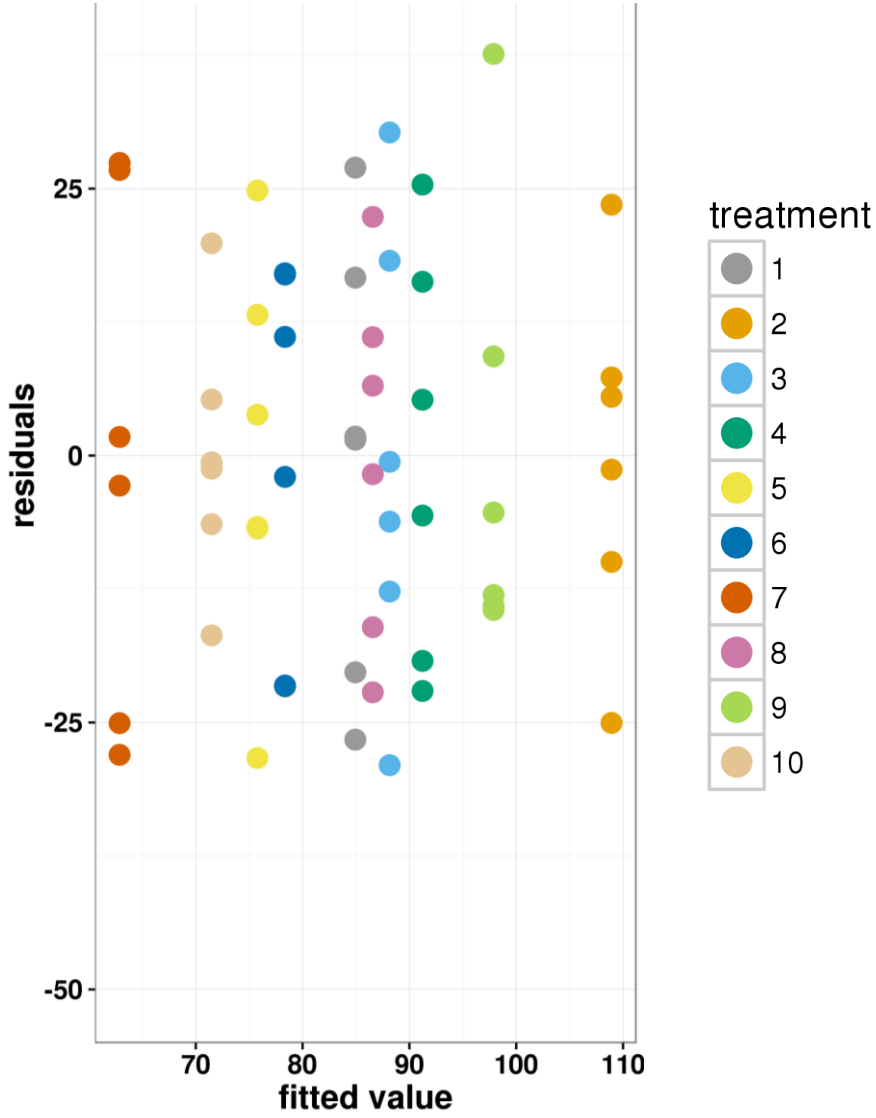




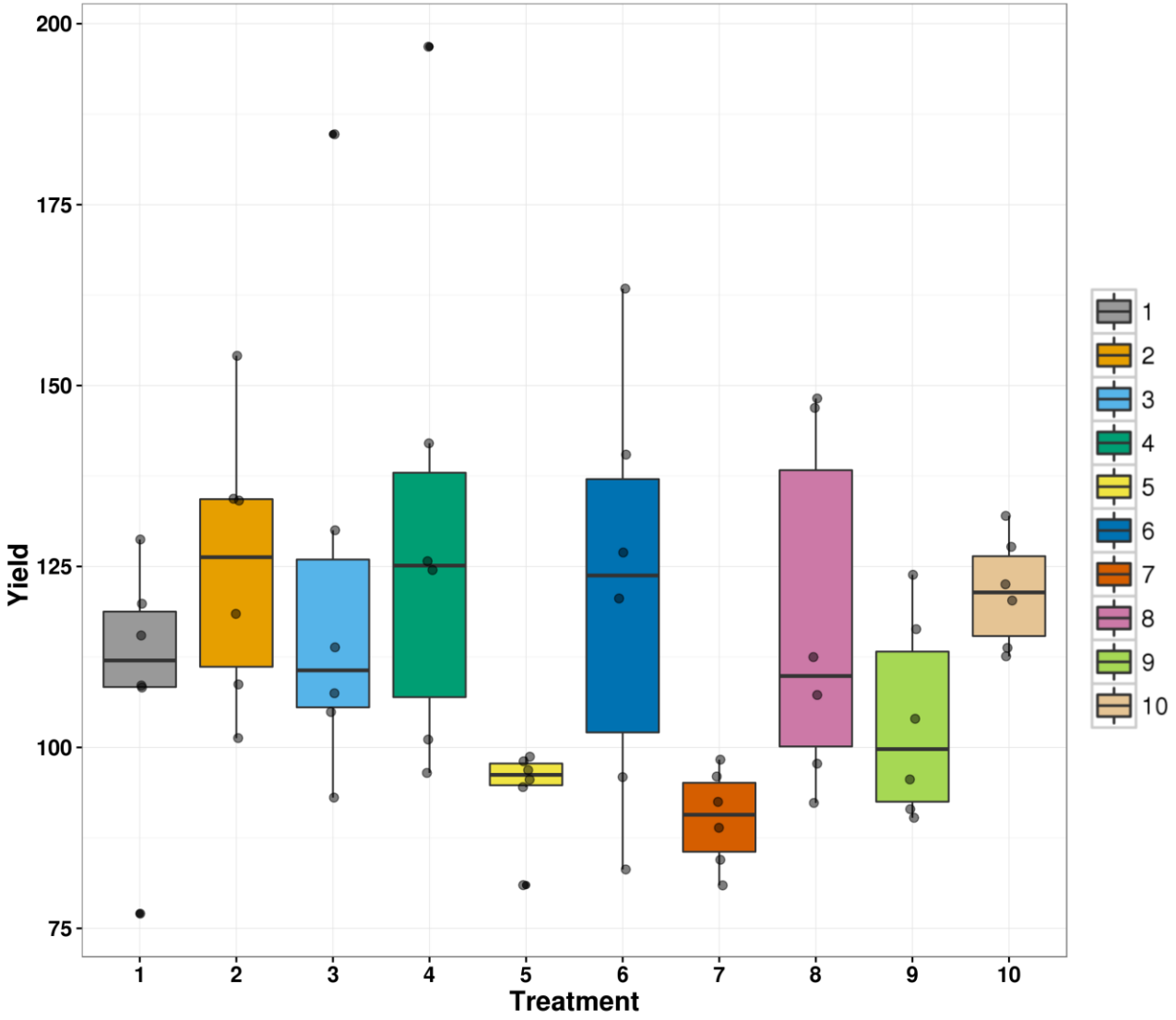
- Data were generated in R and do not represent plant responses to actual biologicals.

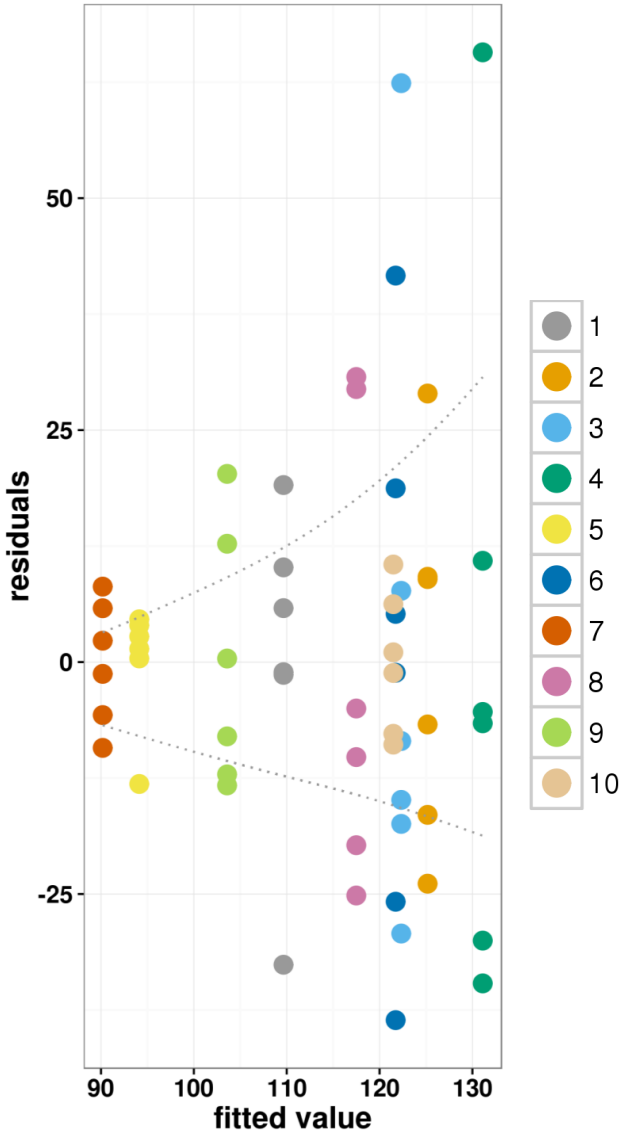
Statistics Modeling Assumption:

- variance independent of the mean (homoskedastic)



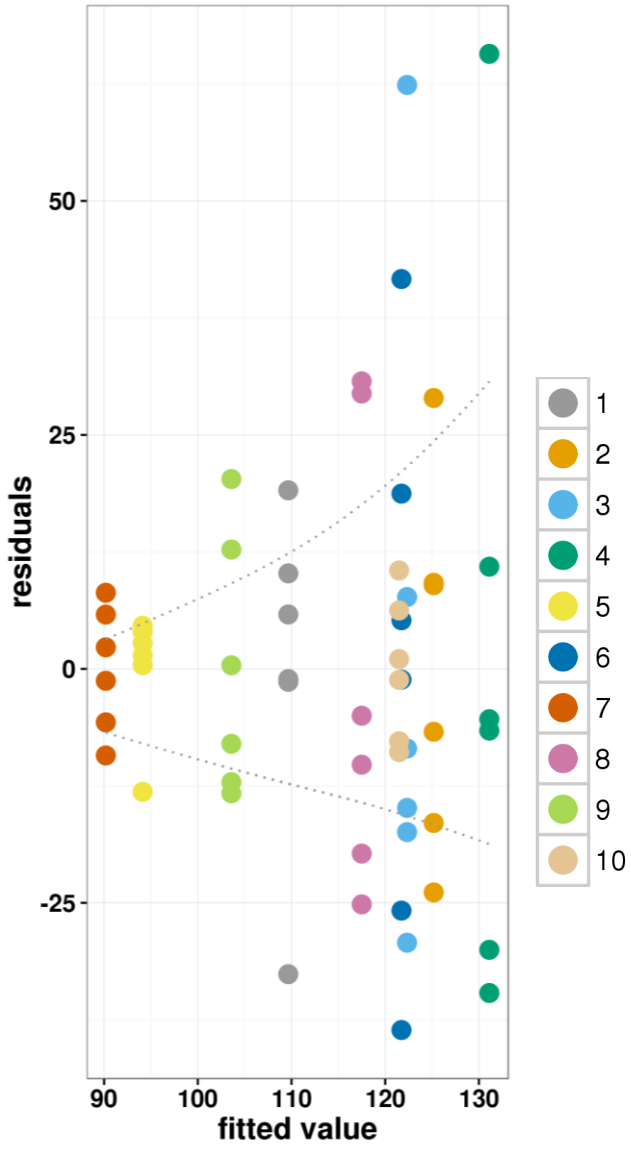
Example: data needing transformation

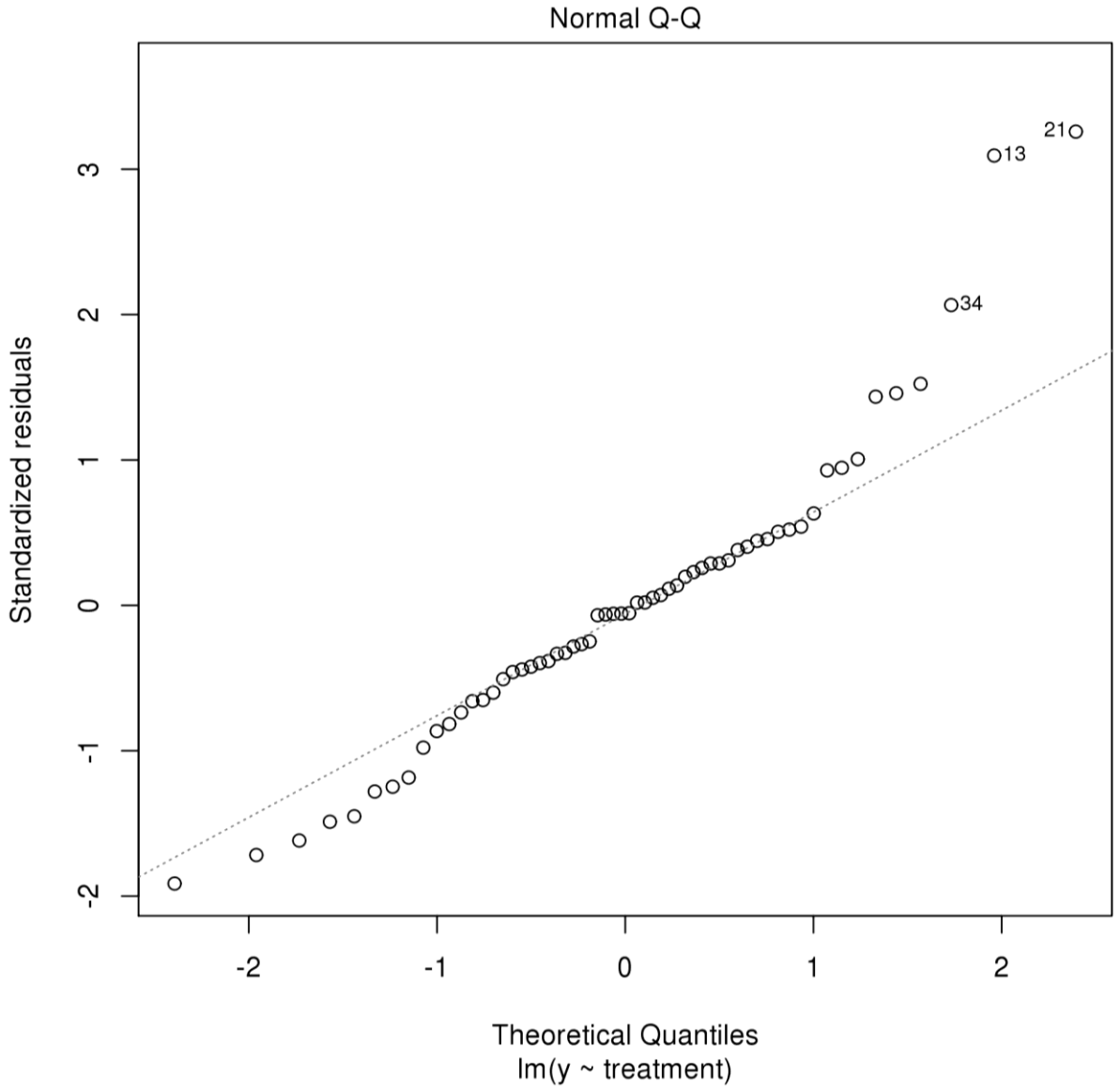




- Data were generated in R and do not represent plant responses to actual biologicals.

Residuals: Quantile-quantile norm plot





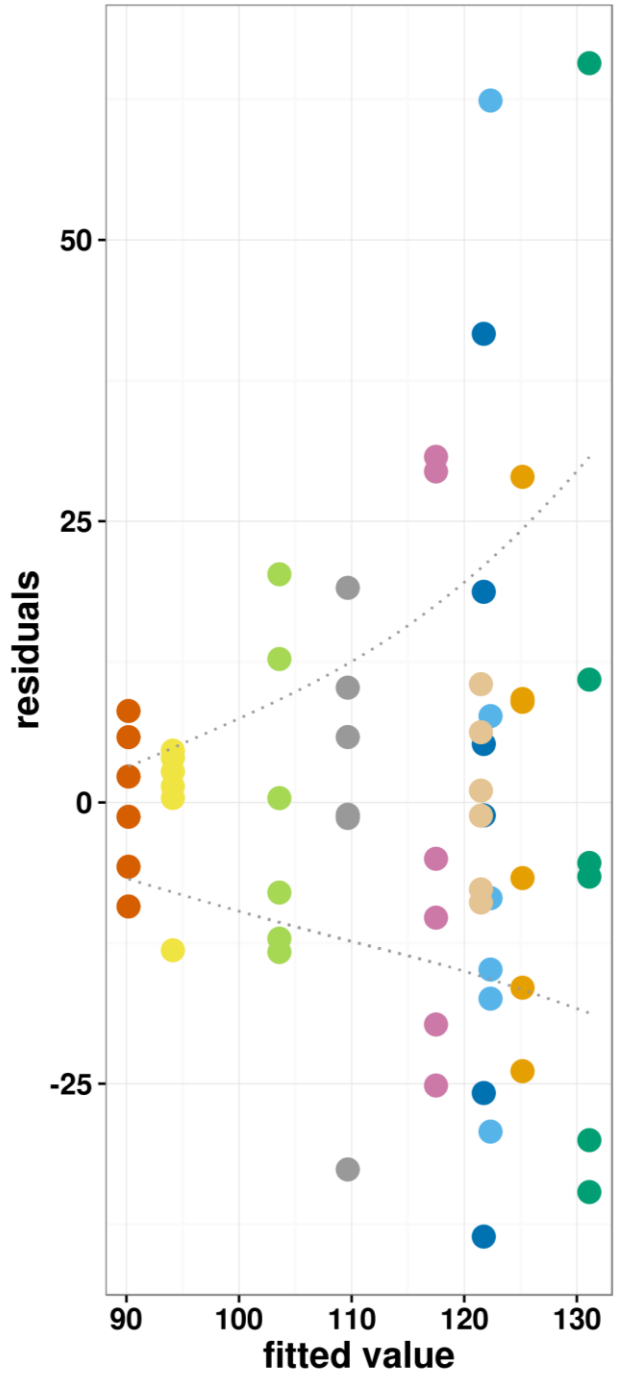
Dealing with heteroskedastic data

- Cases where I've seen log transformation (or other remediation) needed:
  - biomass data (sometimes)
  - disease studies where infection severity is being scored
  - water use efficiency

- No rule; if warranted, investigate data transformation.

Take-home messages

- Know the large-effect variables that influence your measurement of interest
- Know the assumptions your statistical tests are making and validate those assumptions



KOCH™

BIOLOGICAL SOLUTIONS

VISION: Be a global leader in delivering targeted, science-driven biological products that improve crop productivity

www.kochbiologicalsolutions.com
www.kochagronomicservices.com
https://www.youtube.com/watch?v=mqZr4Ed39vU