# Categorizing variables measured with error

Hendriek Boshuizen, Stratos-initiative Topic group 4

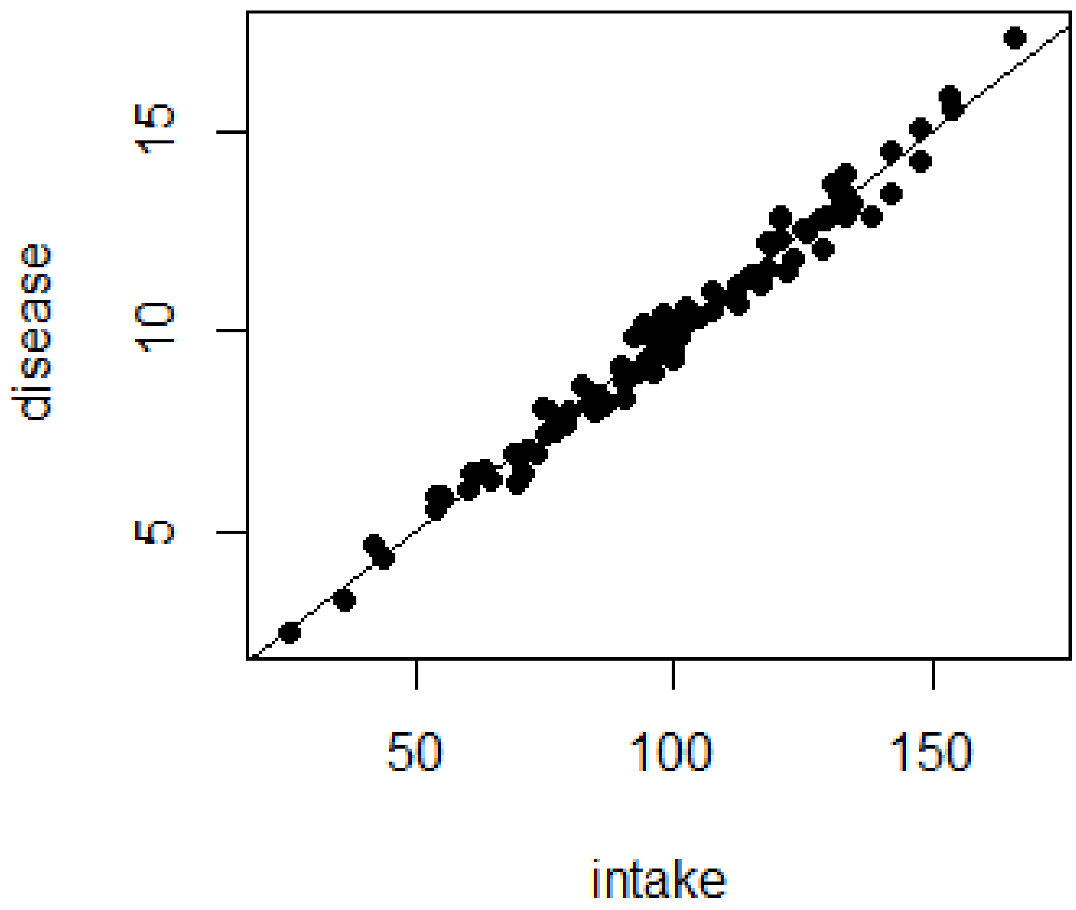
ICDAM 2021, February 8th 2021

WAGENINGEN
UNIVERSITY & RESEARCH

# Background

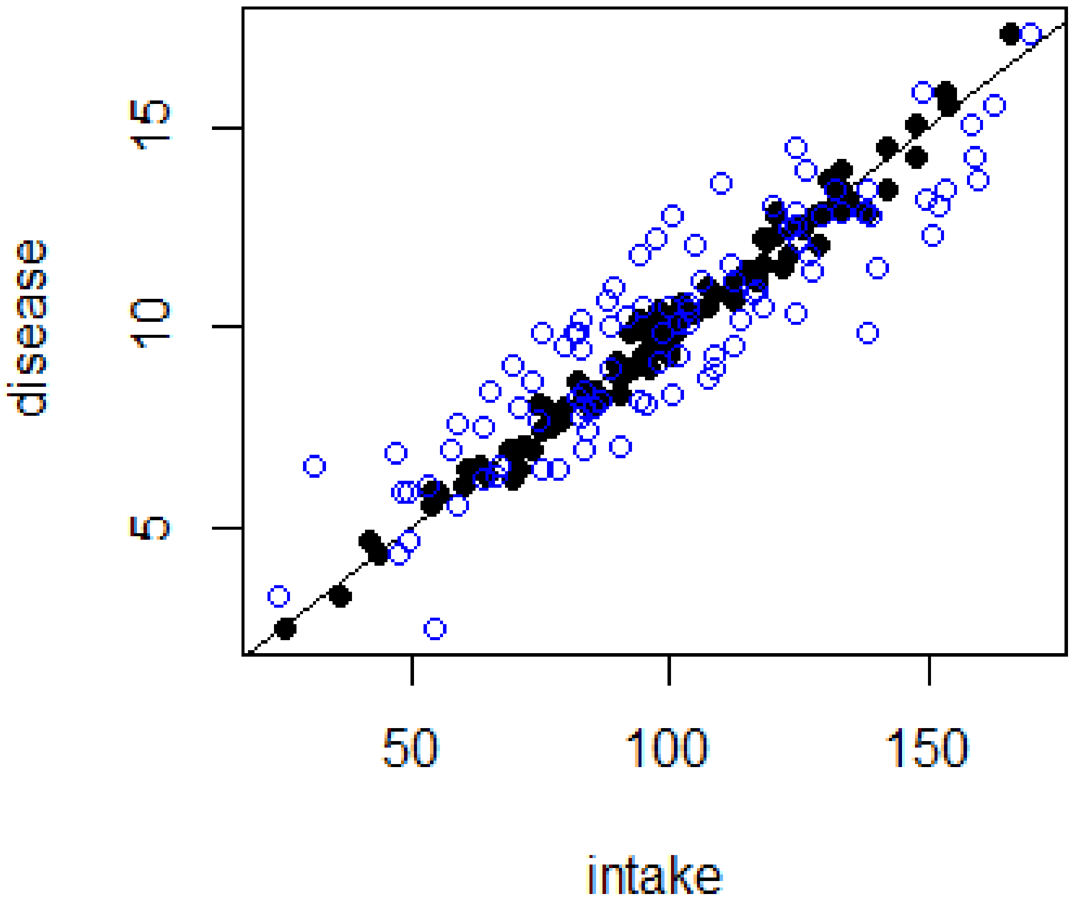
- Categorization of food consumption often done in nutritional studies
- Reason: to study shape / no assumption on shape
- Categorization in general has disadvantages
  - Loss of information / power
  - Can be manipulated
  - Inferior to splines

# Measurement error in continuous variable




WAGENING
UNIVERSITY & RESEA

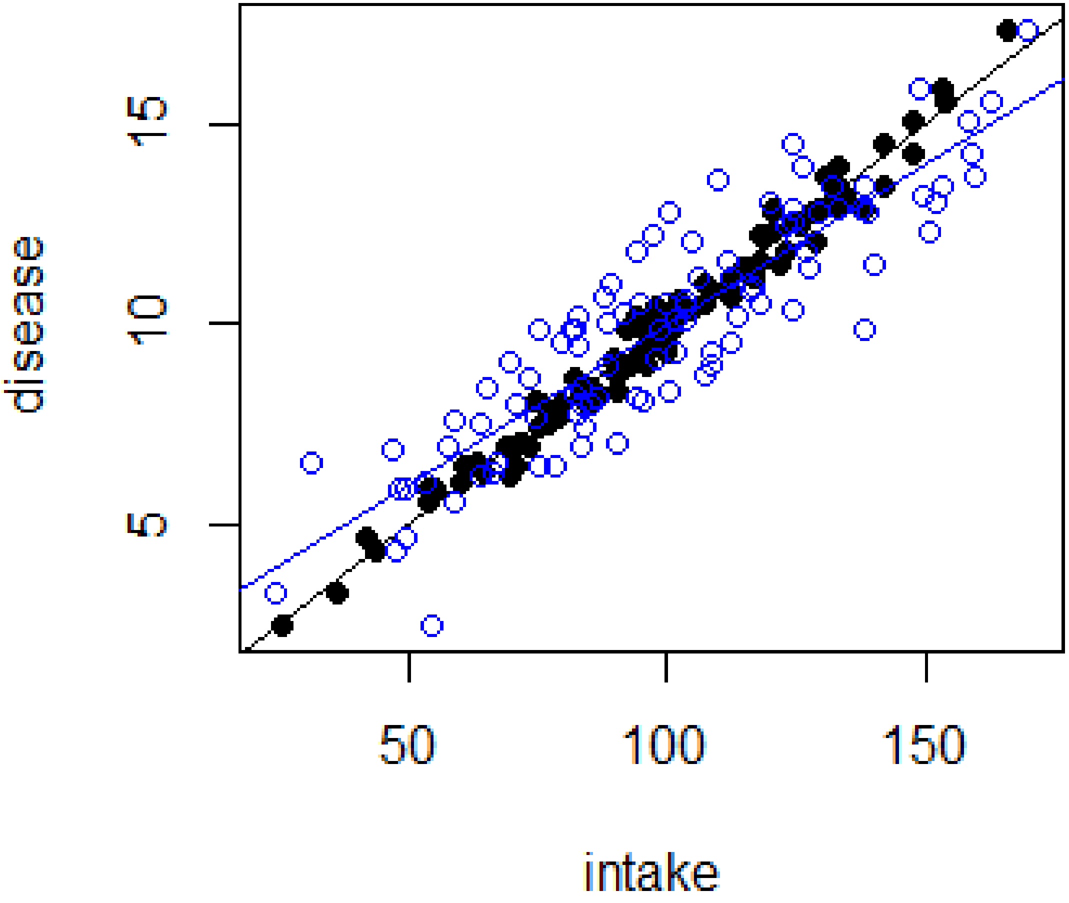
# Measurement error in continuous variable



# Measurement error in continuous variable

➔ Regression Dilution

" Attenuation"

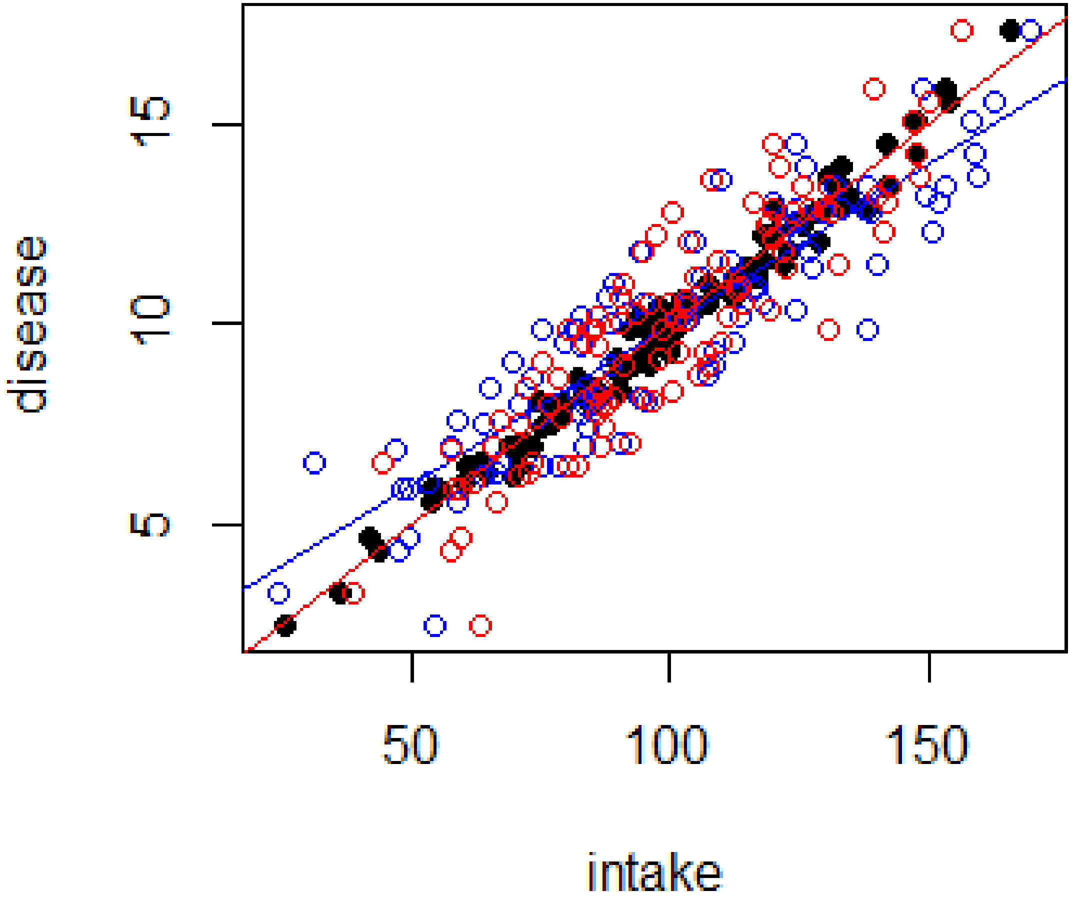



WAGENINGEN UNIVERSITY & RESEARCH

# Measurement error in continuous variable

Regression calibration: use predicted intake

(predicted from mis-measured intake; only alloyed standard needed)



# What if intake is categorized?

Two ways to categorize:

- Quantiles (tertiles, quartiles, quintiles)

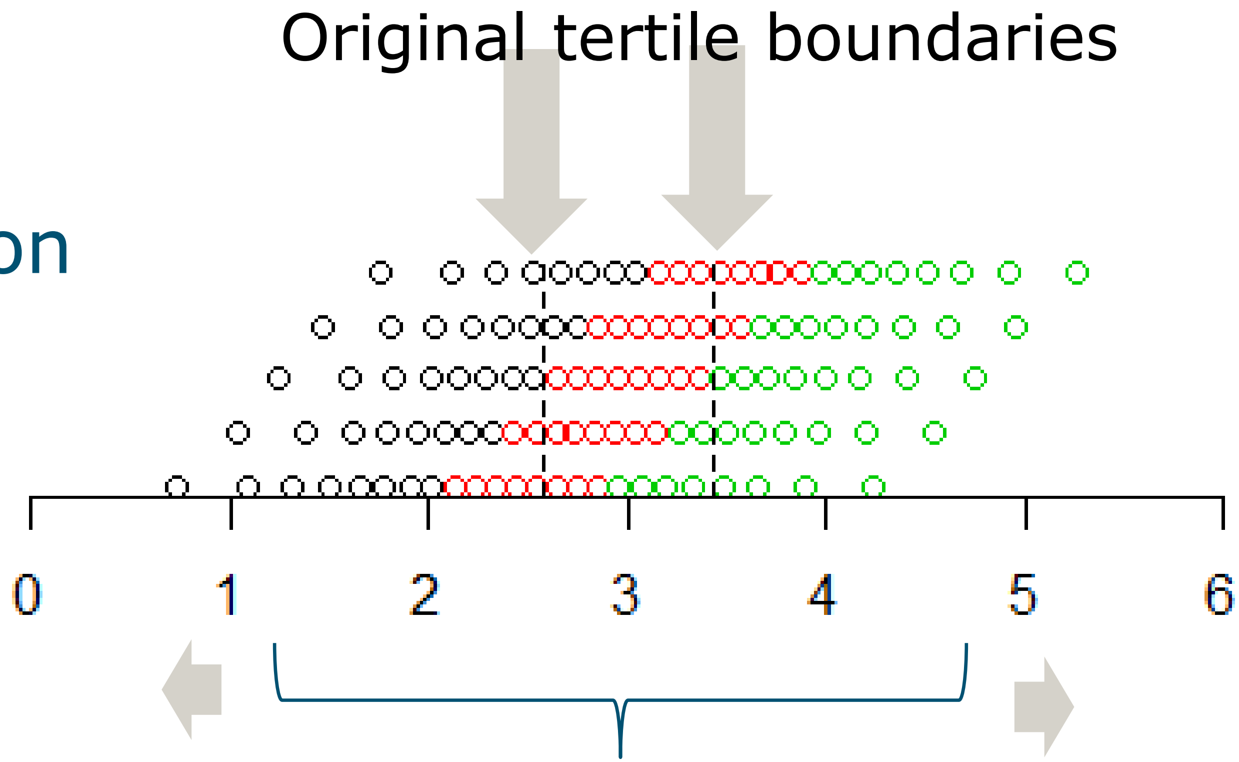
- A priori (<100 g/d, 100-200 g/d, > 200 g/d)

# Tertile without measurement error

## Without error

# Original tertile boundaries

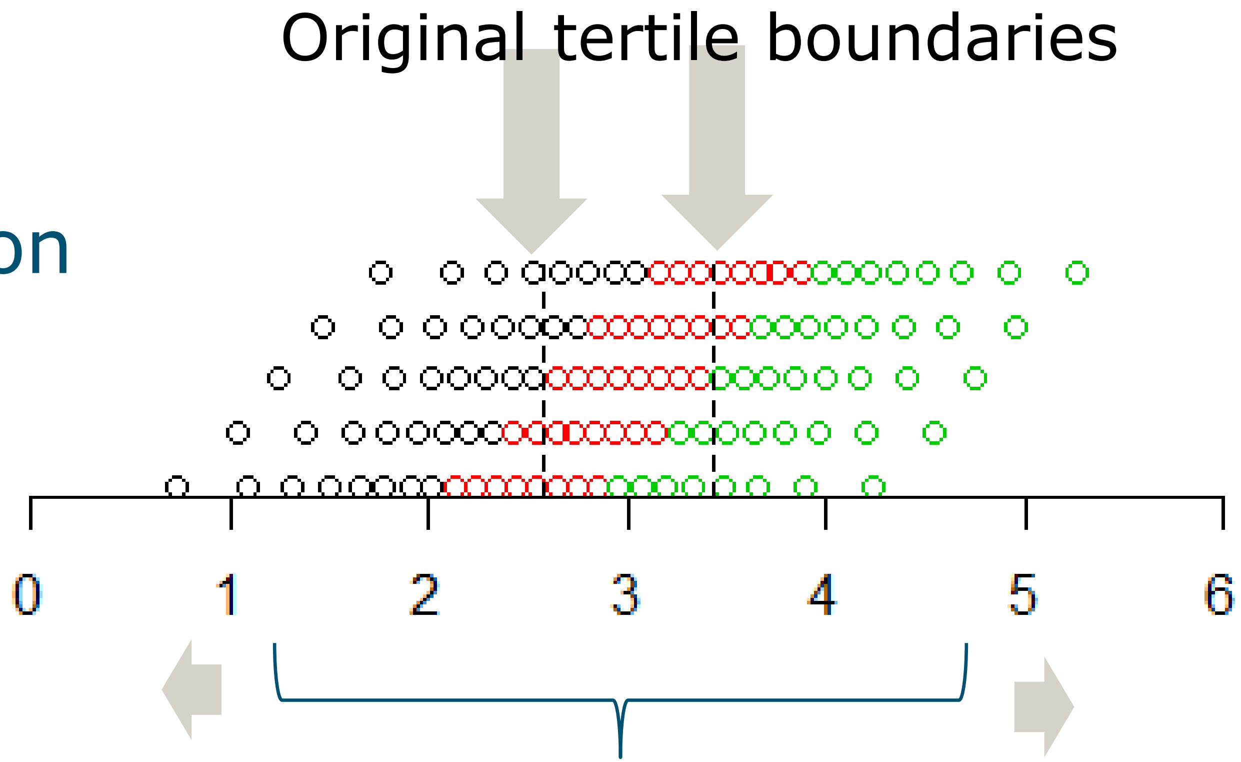
With error:
Misclassification




Less cases in middle category ➔ no longer tertile

# Original tertile boundaries

## With error: Misclassification




Less cases in middle category ➔ no longer tertile

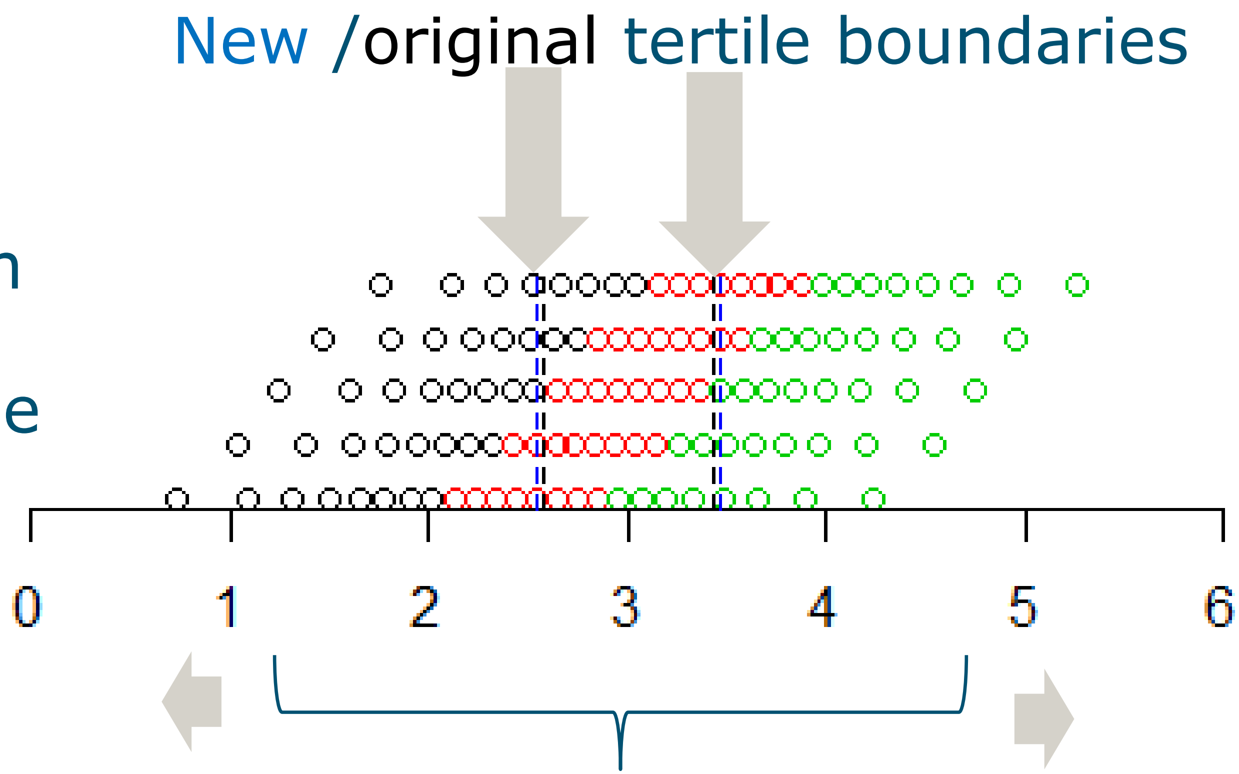
More chance on misclassification closer to cut-off ➔ Differential misclassification

# Use tertile boundary

With error:

Misclassification

More in extreme categories




If normally distributed additive error:
attenuation equal to: correlation true/mismeasured = √(attenuation $factor_{cont}$)
Other cases (including a priori categories): no easy formula

WAGENINGEN
UNIVERSITY & RESEARCH

# What to do?

- Adjust using misclassification rates: need repeated alloyed measurement, complex with confounding

- Quantiles: Divide by correlation true/mismeasured

- A priori: Use calibrated variables = expected value of true intake

  $C_i = E(T_i)$

Assumption:

$$E_{Ci_{\in\ cat}}(C_i) = \quad E_{Ci_{\in\ cat}}E(T_i) \ ?\!\sim E_{Ti_{\in\ cat}}E(T_i\ )$$

# Setup Simulations

Effect of Protein on Systolic Blood Pressure

Measurement error: Trijsburg et al. 2015

SE FFQ ~ population SD

Simulated: linear relation (best case)

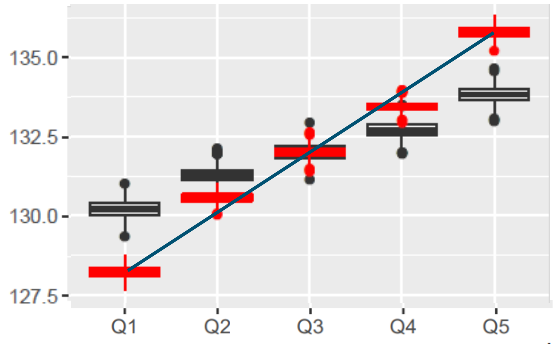
WAGENINGEN
UNIVERSITY & RESEARCH

# Two types of categorization

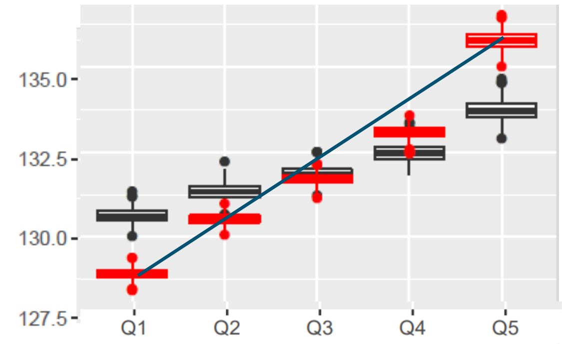
- Quintiles
  - Adjustment: correlation true/mismeasured intake
- A priori
  - Plot against average calibrated values

# Without adjustments

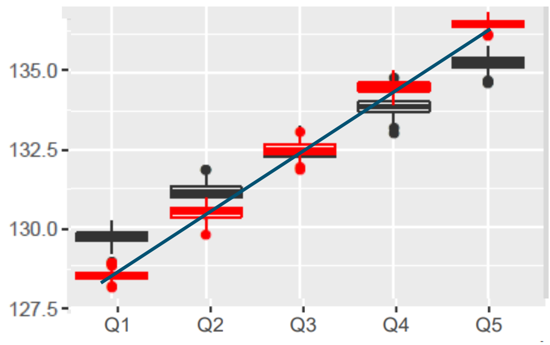
## Normal intake




## Lognormal intake




## Uniform intake




Without error

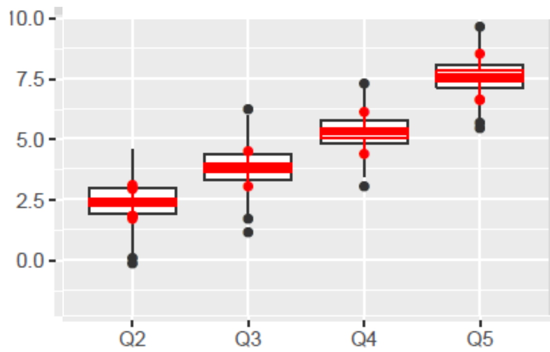
With error

Line between
Q1 and Q5

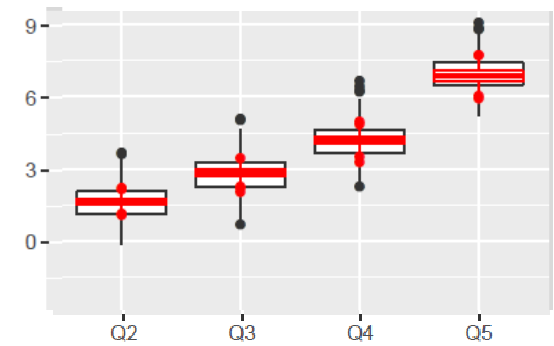
WAGENINGEN
UNIVERSITY & RESEARCH

# With adjustment ($1/corr_{TRUE, MISMEASURED}$)

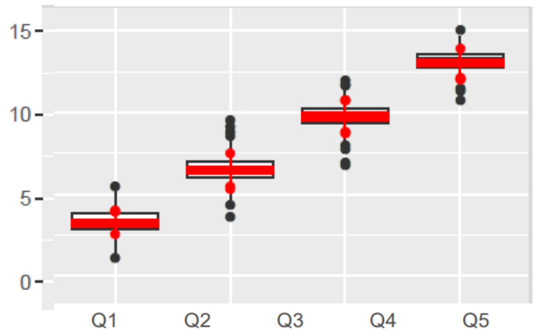
## Normal intake




## Lognormal intake




## Uniform intake



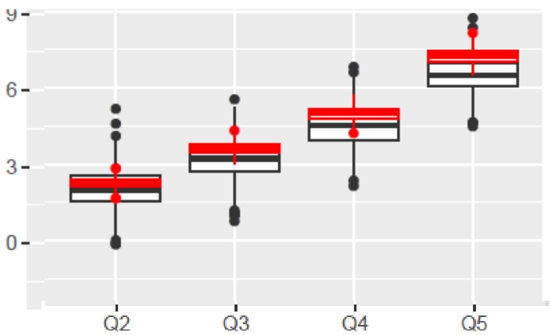

Without error

With error

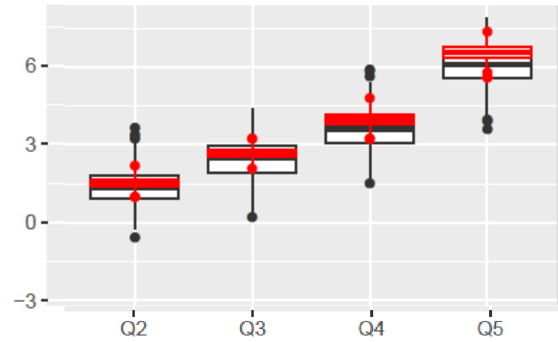
# What about data with confounding?

# Confounding, adjustment (1/corr$_{\text{TRUE, MISMEASURED}}$)

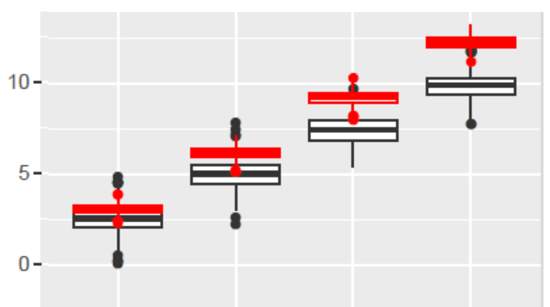
## Normal intake




## Lognormal intake




## Uniform intake




Without error

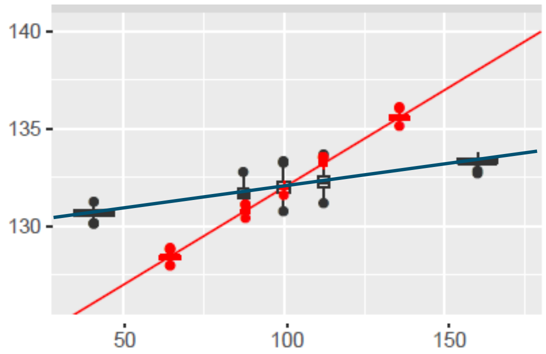
With error

WAGENINGEN
UNIVERSITY & RESEARCH

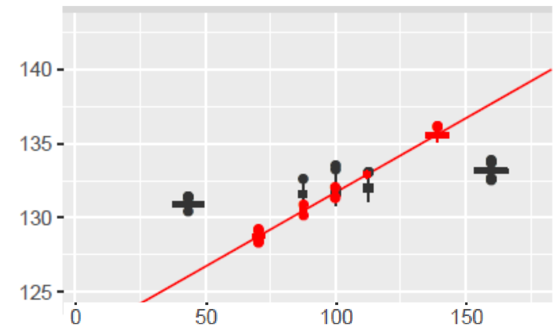
# A priori categories

# Plot against category mean intake

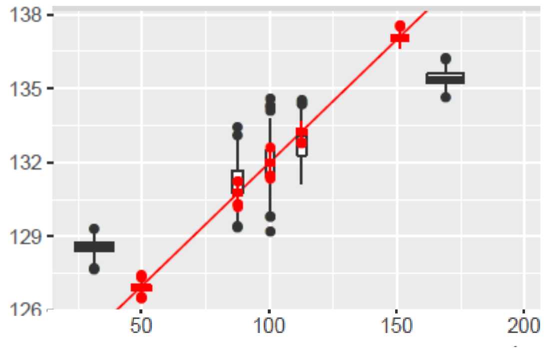
## Normal intake




## Lognormal intake




## Uniform intake

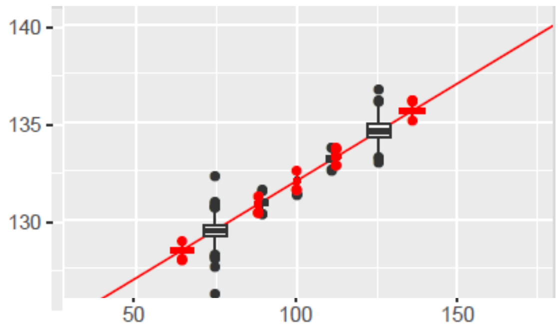



Without error

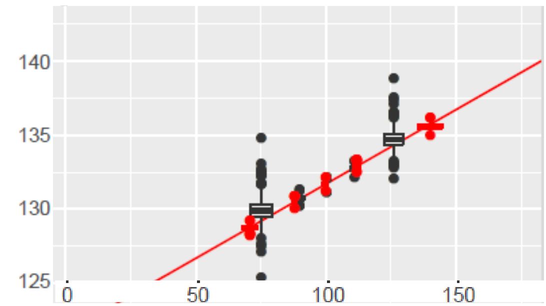
With error

# Calibrated data

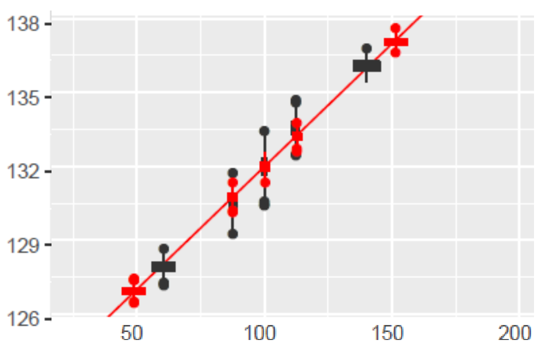
## Normal intake




## Lognormal intake




## Uniform intake



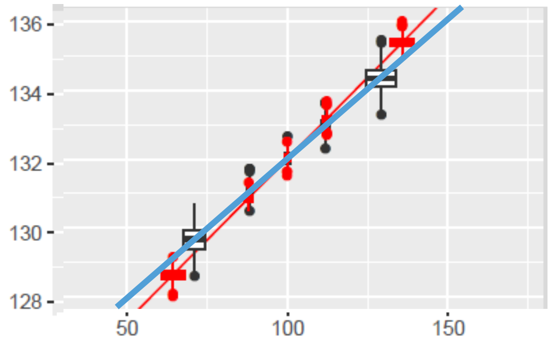

Without error

Calibrated

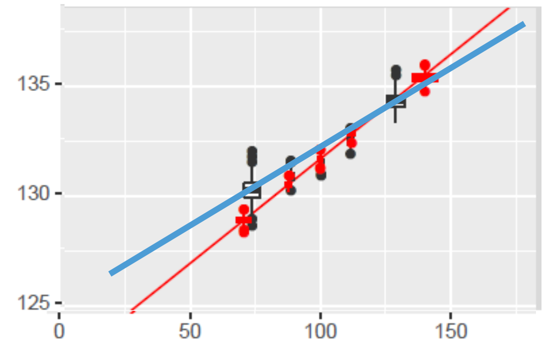
➔ Almost OK

# Calibrated data with confounding

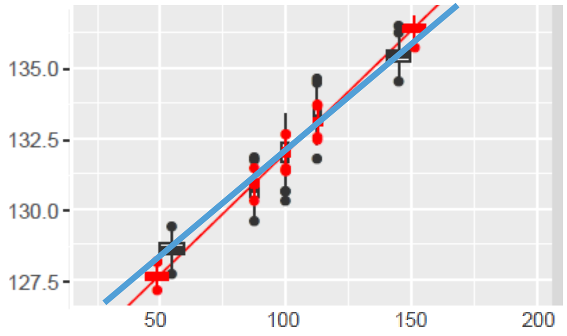
## Normal intake




## Lognormal intake




## Uniform intake




Without error

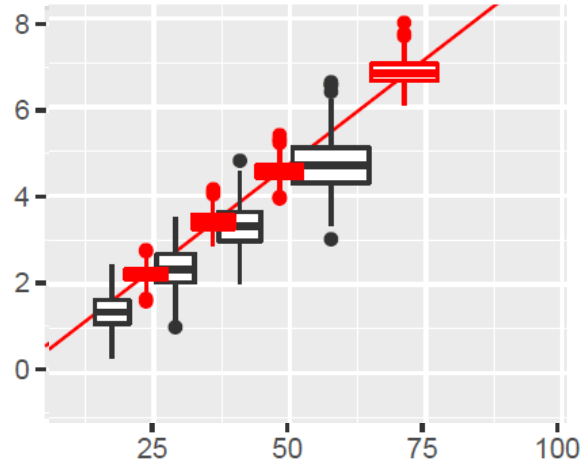
Calibrated

True line

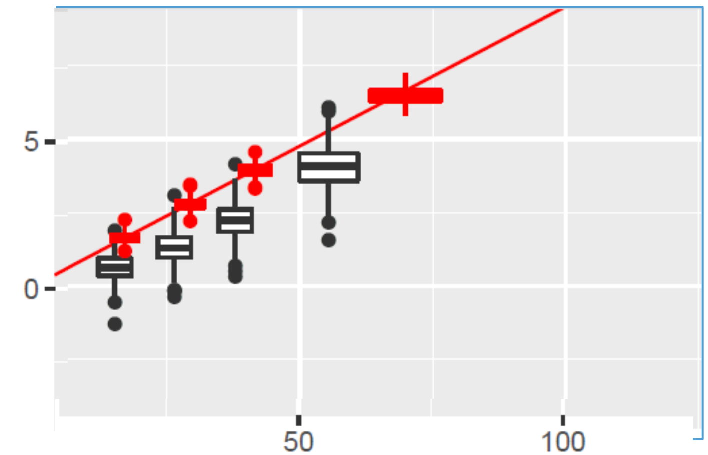
Line on adjusted values

# Difference with lowest category

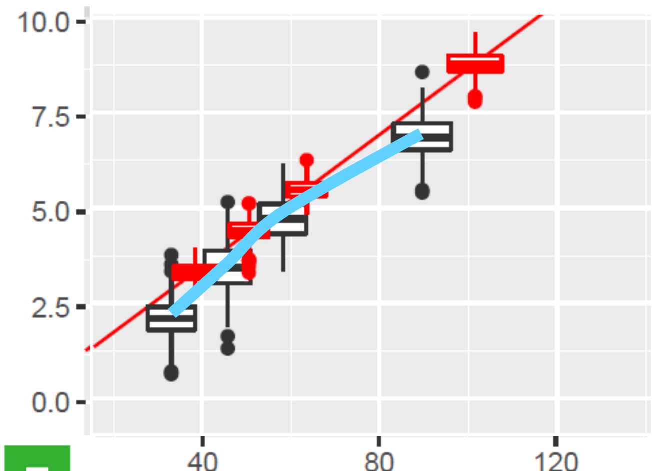
## Normal intake




## Lognormal intake




## Uniform intake




Without error

Calibrated

True line

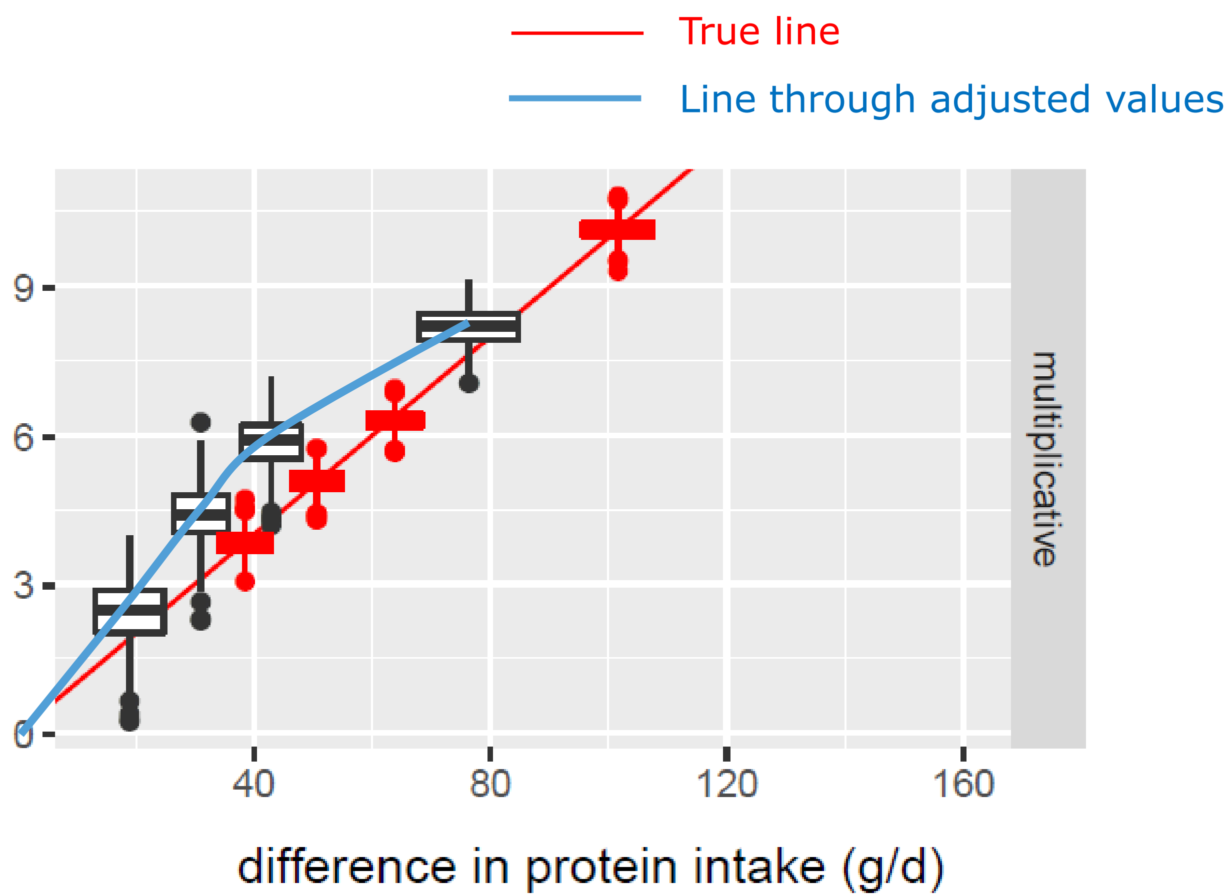
Line through adjusted values

To low

Different shape!

UNIVERSITY & RESEARCH

# Multiplicative error, no confounding




WAGENINGEN
UNIVERSITY & RESEARCH

# Summary

## Quantiles

- $\text{Correlation}_{\text{TRUE,MISMEASURED}}$ [$\sqrt{(\text{attenuation factor}_{\text{cont}})}$] exact for additive error and normal distribution, no confounding

- Approximate for other distributions

- Not very well in presence of confounders

# Summary

## A priori categories

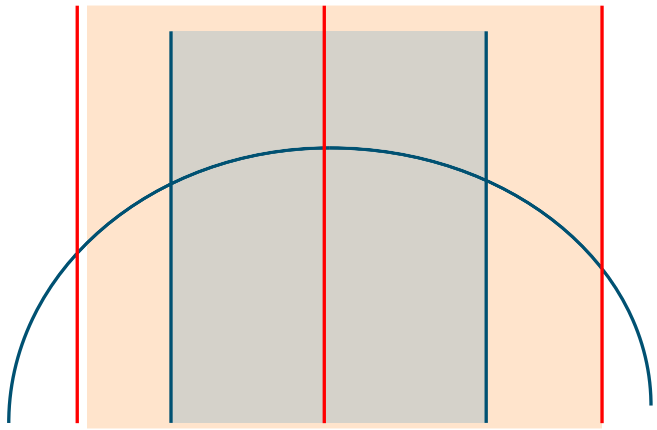
- Calibration works approximately for additive error
- Other error types ➔ possible influence on shape
- Confounding makes this worse
- Part of problem is that effect of confounder is not estimated correctly when data are categorized

# Shape of relation (non linear models)

- Important reason: study shape of relation
- Shape is influenced by measurement error, especially by multiplicative error
- Keogh, Strawbridge & White, Epidemiologic Methods, 1(1):2012 : p-splines higher power

# Shape of relation (non linear models)

Choice of cut-offs important, categories blunt the shape



Blunting by categorization transfers effect to confounder

# Message

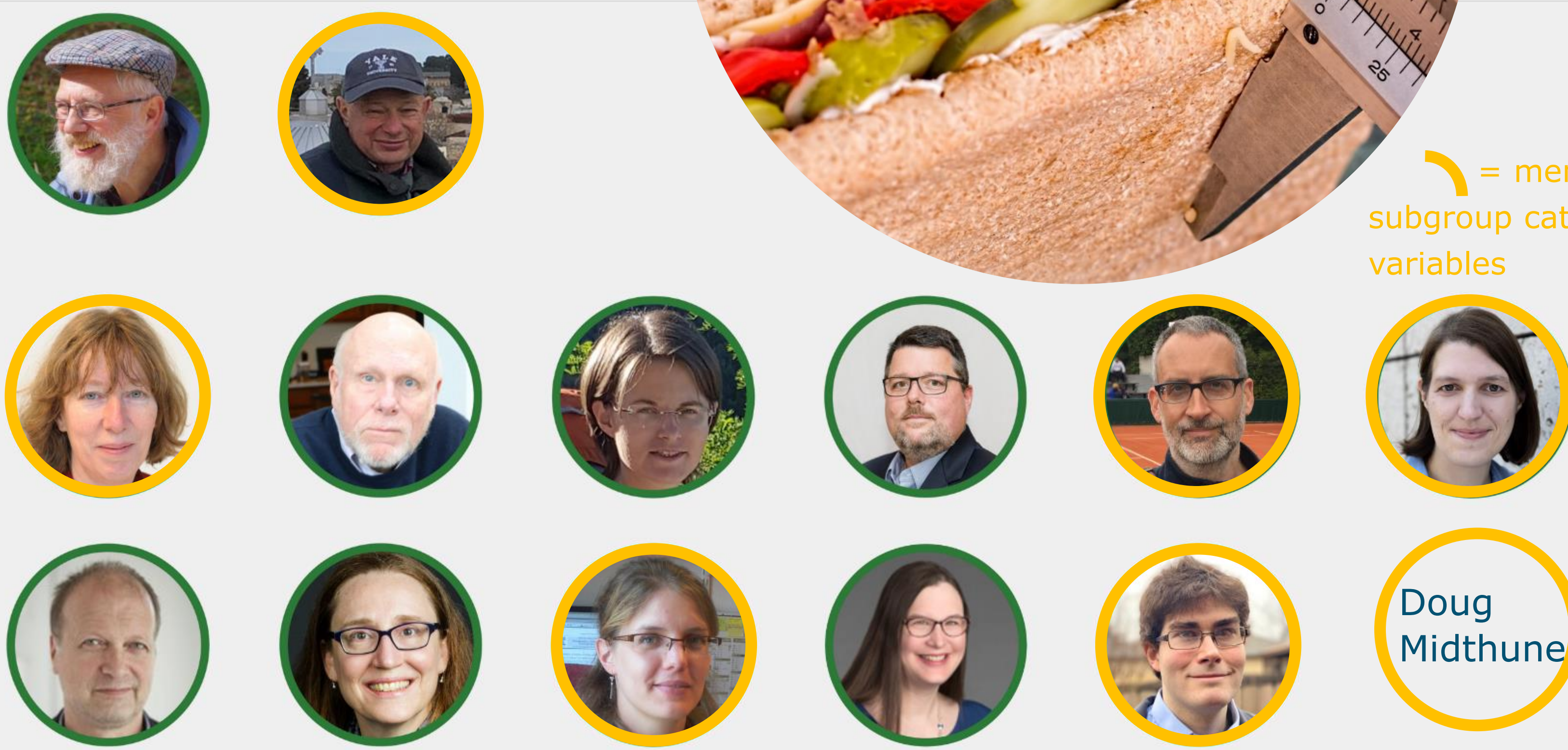
- No easy fix for measurement error in categorized variables, unless no confounding
- Sticking with continuous variables and using smoothing methods best solution
- Also for other reasons

# Thank you for your attention!

Measurement error and misclassification

STRATOS INITIATIVE

Topic Group 4



= member subgroup categorized variables

Doug Midthune

WAGENINGEN UNIVERSITY & RESEARCH

www.stratostg4.statistic.uni-muenchen.de

Image by Steve Buissinne from Pixabay