Biogen

Biostatistics

Multicollinearity

Jerome Sepin

April 19, 2023

Why this presentation?

From Schwabe et al. (2022):

"The primary pharmacokinetic parameters were log-transformed and analyzed using an analysis of covariance (ANCOVA) model that included treatment as a fixed effect, baseline weight category ($\geq 60- \leq 80kg / > 80kg- \leq 100kg$), baseline BMI (continuous variable), and study site as covariates."

$$\log(y) \sim trt + weight_{cat.} + BMI_{cont.} + site$$

What is the problem with this?

$$\log(y) \sim trt + weight_{cat.} + BMI_{cont.} + site$$

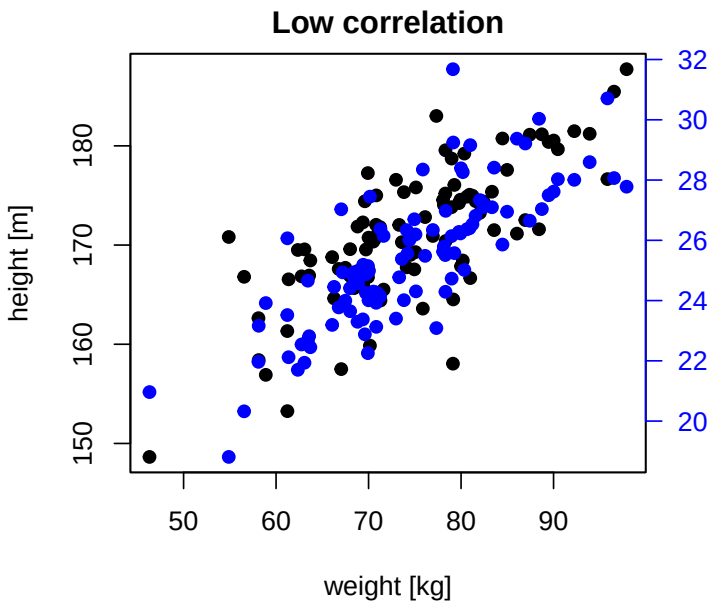
Since $\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2}$

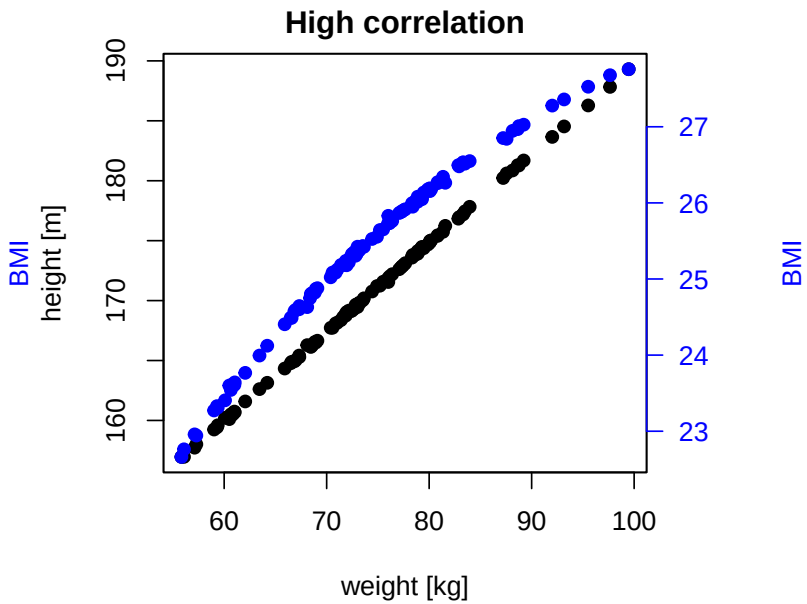
$$\log(y) \sim trt + weight_{cat.} + \frac{weight_{cont.}}{height^2_{cont.}} + site$$

Thus, *weight* and *BMI* may be highly correlated (a special case of collinearity)!

What is the problem with collinear variables in the model?

Let's see what happens when the correlation between *height* and *weight* changes!





Small simulation:

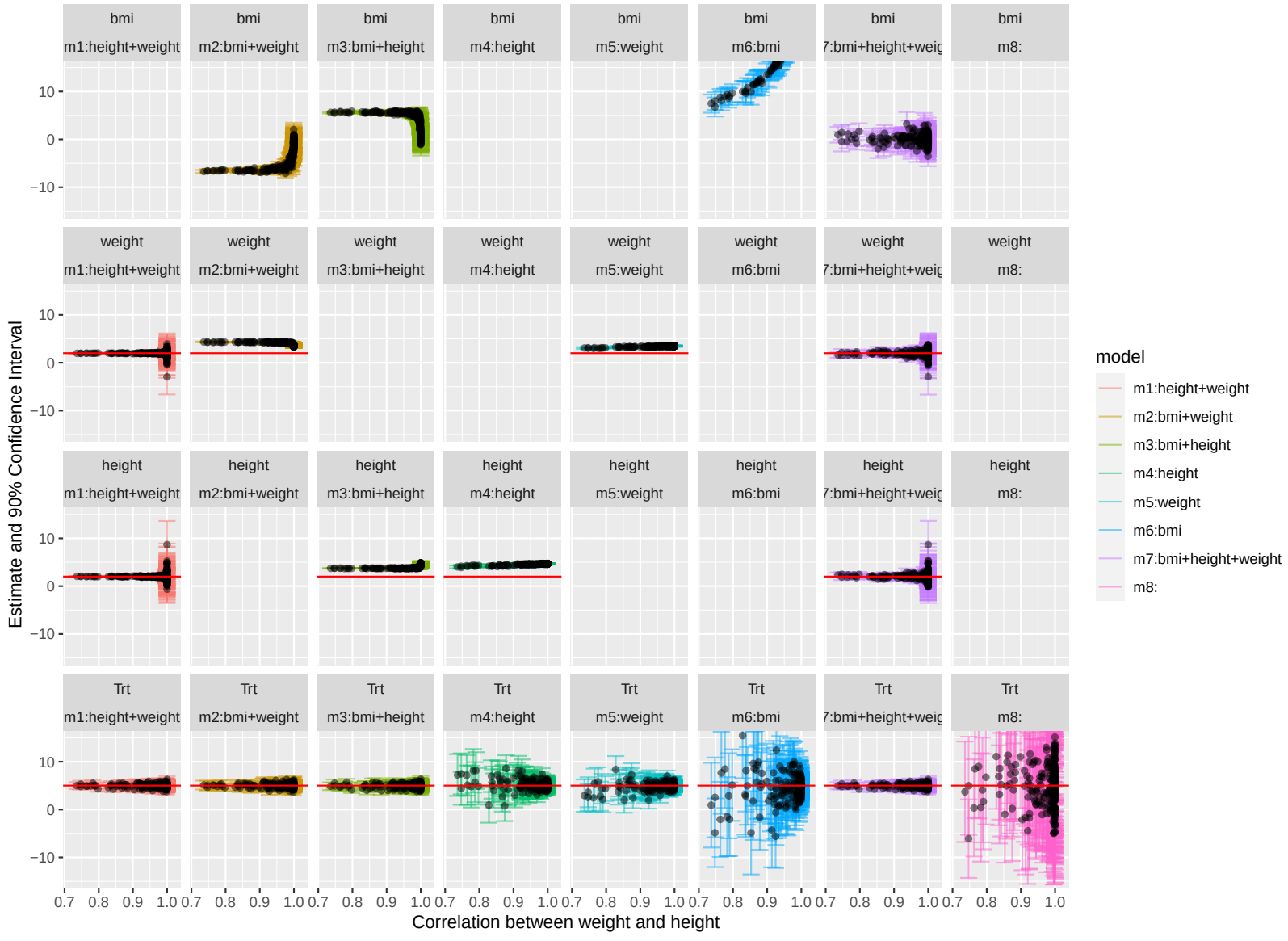
$$\boxed{\log(y) \sim trt + weight_{cat.} + BMI_{cont.} + site}$$

For simplicity without *site* and with continuous variables!

"True" model: $\log(y) \sim 5 + 5 \cdot trt + 2 \cdot weight + 2 \cdot height$

Dichotomia is a problem for a different day...

Small simulation: Results



Relation of variables

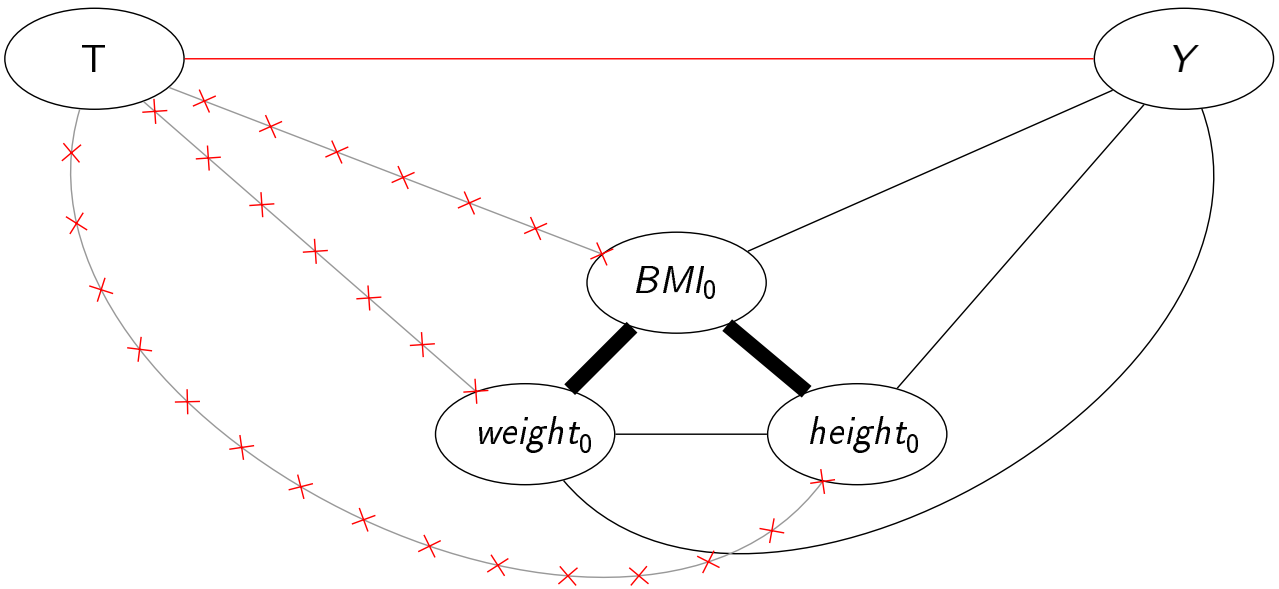


Figure: Randomization removes confounding

$$\boxed{\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2}}$$

But why this model?

Domain knowledge?

(Statistical) Model selection is a problem for a different day...

But why this model?

ICH E9 Expert Working Group (1999): *"In multicentre trials (see Glossary) the randomisation procedures should be organised centrally. It is advisable to have a separate random scheme for each centre, i.e. to stratify by centre or to allocate several whole blocks to each centre. More generally, stratification by important prognostic factors measured at baseline (e.g. severity of disease, age, sex, etc.) may sometimes be valuable in order to promote balanced allocation within strata; this has greater potential benefit in small trials.* *The use of more than two or three stratification factors is rarely necessary, is less successful at achieving balance and is logistically troublesome.* *The use of a dynamic allocation procedure (see below) may help to achieve balance across a number of stratification factors simultaneously provided the rest of the trial procedures can be adjusted to accommodate an approach of this type.* *Factors on which randomisation has been stratified should be accounted for later in the analysis.* *"*

Collinearity != Correlation

- Correlation is a special case of collinearity (includes only two variables)
- Problems can also arise without large (pairwise) correlations
- For example when explanatory variables add up to 100% (recipe)

milk	water	...	salt	total
40%	30%	...	1%	100%
45%	35%	...	2%	100%
...	...	...	...	100%

Recipes remain secret!

Conclusion

- Collinearity increases uncertainty (but only for the variables that contribute)
- "Killing" collinearity may induce bias
- Randomization to the help
- Trade-off between explanatory power and sparsity (bias-variance dilemma)

References

ICH E9 Expert Working Group (1999). Statistical principles for clinical trials: Ich harmonized tripartite guideline. *Statistics in Medicine*, 18:1905–1942.

Schwabe, C., Illes, A., Ullmann, M., Ghori, V., Vincent, E., Petit-Frere, C., Monnet, J., Racault, A. S., and Wynne, C. (2022). Pharmacokinetics and pharmacodynamics of a proposed tocilizumab biosimilar MSB11456 versus both the US-licensed and EU-approved products: a randomized, double-blind trial. *Expert Review of Clinical Immunology*, 18(5):533–543.