Assessment of the detrimental effects of collinearity in classical and transformation models

---

Master Thesis in Biostatistics (STA495)

by

Jerome Sepin
17-932-427

supervised by

PD Dr. Malgorzata Roos

Zurich, March 21, 2023

Assessment of the detrimental effects of collinearity in classical and transformation models

Jerome Sepin

Version March 21, 2023

Acknowledgement

I would like to express my sincere gratitude to everyone who has supported me during the completion of this thesis. I am particularly grateful to my supervisor, PD Dr. Malgorzata Roos, for her guidance, insightful feedbacks, and unwavering support throughout the research process. Without your help and infectious positive mindset, this would have never been possible. Moreover, I would like to thank the whole teaching staff of the Biostatistics Master Program for doing their best in teaching statistics and making my time as student, both instructive and inspiring.

Jerome Sepin
February 2023

Contents

Abstract v

1 Introduction 1

2 Causal inference from complex longitudinal data 3

2.1 TIME-VARYING TREATMENTS (19) . . . 3

2.1.1 The causal effect of time-varying treatments (19.1) . . . 3

2.1.2 Treatment strategies (19.2) . . . 3

2.1.3 Sequentially randomized experiments (19.3) . . . 4

2.1.4 Sequential exchangeability (19.4) . . . 4

2.1.5 Identifiability under some but not all treatment strategies (19.5) . . . 4

2.1.6 Time-varying confounding and time-varying confounders (19.6) . . . 4

2.2 TREATMENT-CONFOUNDER FEEDBACK (20) . . . 5

2.2.1 The elements of treatment-confounder feedback (20.1) . . . 5

2.2.2 The bias of traditional methods (20.2) . . . 6

2.2.3 Why traditional methods fail (20.3) . . . 7

2.3 G-METHODS FOR TIME-VARYING TREATMENTS (21) . . . 7

2.3.1 The g-formula for time-varying treatments (21.1) . . . 7

2.4 Transformation models . . . 8

2.4.1 An example . . . 9

2.5 Differences between `lm` and `tram::Lm` . . . 10

2.5.1 Maximum-Likelihood estimation for the linear regression model . . . 11

2.5.2 Maximum-Likelihood estimation for the transformation model equivalent (`tram::Lm`) . . . 11

A Appendix 13

A.1 Approximate likelihood 13
A.2 Computational reproducibility 14

Bibliography **17**

Abstract

Multiple linear regression techniques are well-established statistical tools that are able to quantify the association between many explanatory variables and one outcome variable in a human-interpretable manner. However, many explanatory variables increase the chance of collinearity, which means that one of them is well explainable by linear combinations of others. It is well known that collinearity has detrimental impacts on multiple linear regression estimands, thus stimulating research on collinearity. For example, Belsley came up with a rule of thumb to detect harmful collinearity, which says that condition indices, and therefore also condition numbers, over 30 indicate consequential collinearity. In the meantime, this rule of thumb has been widely advocated so that it seems to be carved in stone. Therefore, it is important to design a Monte Carlo simulation to clarify the relevance of this cut-off.

Belsley's rule of thumb applies to the omnipresent statistical workhorse, the least-squares model. However, with the rise of computational power, novel transformation models that are able to flexibly transform the outcome have a large impact on the understanding of regression models. It is currently not known whether both, least-squares and the transformation model equivalent, react equally to collinearity. Thus, it is important to clarify whether collinearity diagnostics procedure developed with least-squares can also be used in transformation models.

Furthermore, it can be expected that the sample size can mitigate the detrimental impact of collinearity, but there are currently no exact rules how to do this. Thus, there is a demand for software and well-explained hands-on examples that assist in properly adjusting the study design to account for collinearity.

To address these needs in this master thesis, we designed and conducted a Monte Carlo simulation study where we found no signs of tipping point at Belsley's cut-off value of 30. However, we discovered that the degree of collinearity summarized by one condition number impacts the Wald statistics values of both, the least-squares model and the transformation model equivalent. We also demonstrated that the Wald statistic values differ in general between the two methods. Moreover, we proposed a method for sample size calculation in the least-squares case. The methods developed are implemented in open-source R software, which is integrated in the `Collinearity` package. As additional support, we also demonstrated how to apply these methods in a case study using the `BostonHousing2` data. These examples and functions assess the impact of the detrimental effect of collinearity on multiple linear regression estimands and suggest how to improve the sample size to mitigate this detrimental effect.

Chapter 1

Introduction

Multiple linear regression techniques are well-established and easy-interpretable statistical tools that can incorporate many explanatory variables associated with one outcome variable. Many explanatory variables increase the chance of collinearity, which means that one of them is well explained by a linear combination of others. It is well known that collinearity has detrimental impacts on multiple linear regression estimands.

Chapter 2

Causal inference from complex longitudinal data

This chapter summarizes Hernan and Robins (2023) part III.

2.1 TIME-VARYING TREATMENTS (19)

2.1.1 The causal effect of time-varying treatments (19.1)

time-varying dichotomous treatment A_k at time $k = 0, 1, 2, \ldots, K$

We refer to the treatment history as $\bar{A}_k = (A_0, \ldots, A_K)$ Patients under treatment throughout the whole time K have $\bar{A} = (A_0 = 1, \ldots, A_K = 1) = \bar{1}$ while patients that never receive treatment do have history $\bar{A} = (A_0 = 0, \ldots, A_K = 0) = \bar{0}$. Most likely patients will have some sort of a mixture between those two treatment strategies due to various reasons and therefore a compact (and true) notation is not really possible. The average treatment effect at timepoint k may be defined as:

$$\mathbb{E}\left[Y^{a_k=1}\right] - \mathbb{E}\left[Y^{a_k=0}\right]$$

however, this does not reflect the overall goal to quantify the average causal effect of the time-varying treatment. Due to the fact that in reality the treatment strategies cannot be completely discriminated between $\bar{1}$ or $\bar{0}$, no **unique** average causal effect is defined.

2.1.2 Treatment strategies (19.2)

"Always treat" vs. "Never treat": $\mathbb{E}\left[Y^{\bar{a}=\bar{1}}\right] - \mathbb{E}\left[Y^{\bar{a}=\bar{0}}\right]$

Dynamic treatment strategy: a_k depends on the evolution of the patients time-varying covariate(s) $\bar{L}_k$ (otherwise they are called static). Development of ADA and subsequent discontinuation of the treatment is a form of dynamic treatment strategy.

The average causal effect estimate of a (potentially) time-varying treatment is only well defined if the strategies are specified (and hold). Since at each timepoint k there is the option (random or deterministic) to treat or not to treat, there are a many different treatment strategies available resulting in **many** different causal effects.

2.1.3 Sequentially randomized experiments (19.3)

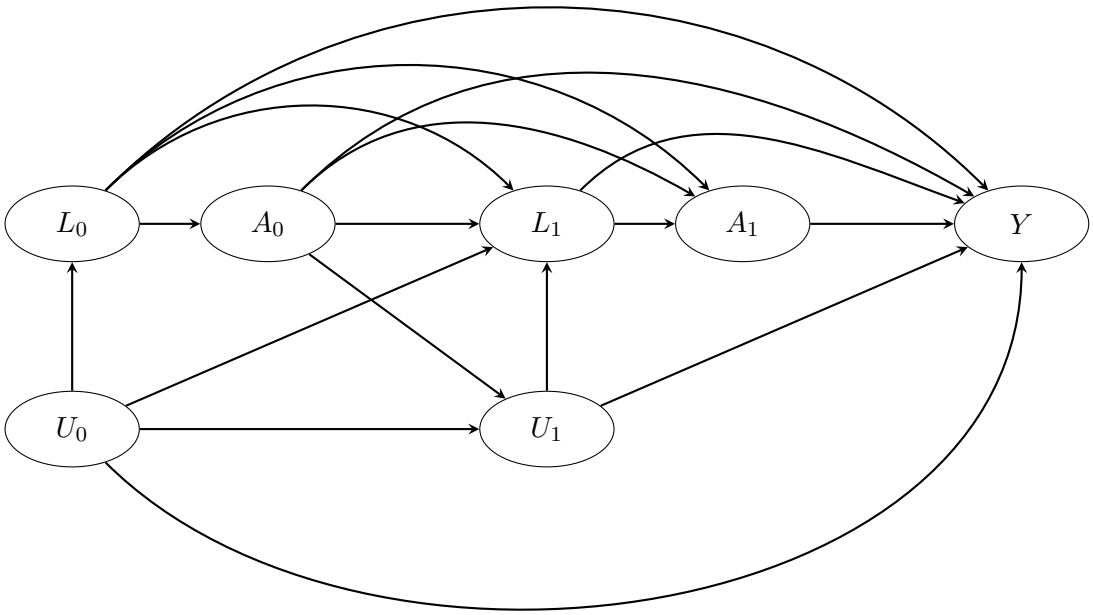

Figure 2.1: Causal diagram of a randomized clinical trial with dynamic treatment strategy, meaning that depending on some of the patients covariates, the treatment may be changed. Corresponds to Figure 19.2 in Hernan and Robins (2023)

2.1.4 Sequential exchangeability (19.4)

Causal inference (unbiased effect estimation) with time-fixed treatment effects require conditional exchangeability $Y^a \perp A \mid L$

(p247: "Conditional exchangeability holds in observational studies if the probability of receiving treatment depends on the measured covariates L and, conditional on L, does not further depend on any unmeasured, common causes of treatment and outcome.")

Sequential conditional exchangeability:

$$Y^g \perp A_k \mid \bar{A}_{k-1} = g\left(\bar{A}_{k-2}, \bar{L}_{k-1}\right), \bar{L}_k \tag{2.1}$$

2.1.5 Identifiability under some but not all treatment strategies (19.5)

Needs knowledge of *d*-separation and SWIG.

2.1.6 Time-varying confounding and time-varying confounders (19.6)

p253: "*Achieving approximate exchangeability requires expert knowledge, which will guide investigators in the design of their studies to measure as many of the relevant variables $\bar{L}_k$ as possible. For example, in an HIV study, experts would agree that time-varying variables like CD4 cell count, viral load, symptoms need to be appropriately measured and adjusted for.*"

p253: "*But the question "Are the measured covariates sufficient to ensure sequential exchangeability?" can never be answered with certainty.*"

2.2 TREATMENT-CONFOUNDER FEEDBACK (20)

"*Suppose that we have a study in which the strongest form of sequential exchangeability holds: the measured time-varying confounders are sufficient to validly estimate the causal effect of any treatment strategy. Then the question is what confounding adjustment method to use. The answer to this question highlights a key problem in causal inference about time-varying treatments: treatment-confounder feedback.*"

If treatment-confounder feedback is apparent, traditional adjustment methods induce bias.

2.2.1 The elements of treatment-confounder feedback (20.1)

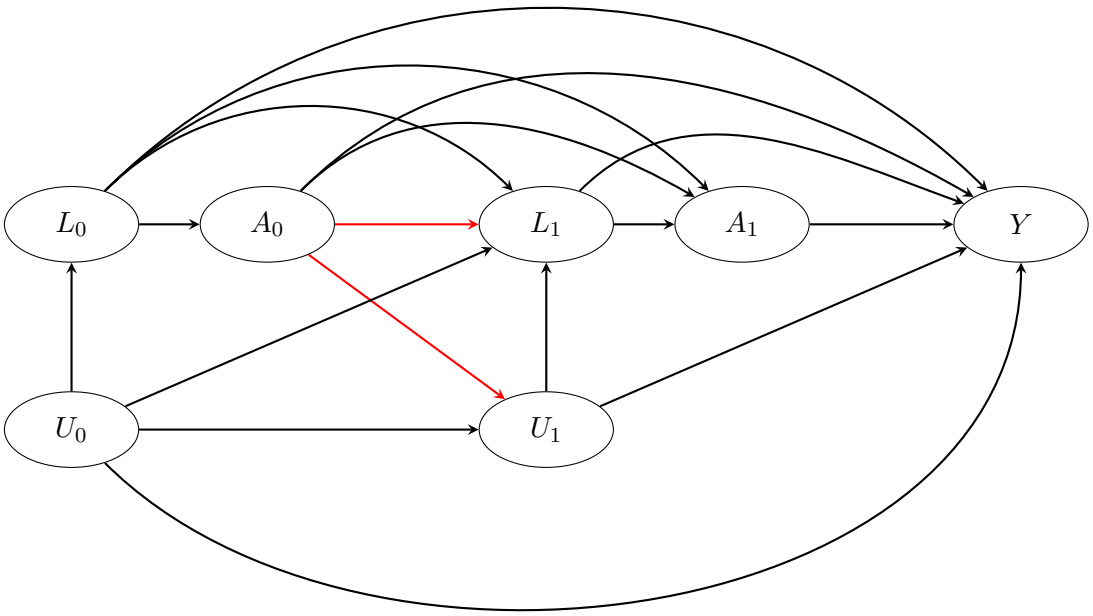

Figure 2.2: Corresponds to Figure 19.2 or 20.1 in Hernan and Robins (2023). The **treatment-confounder feedback** means in this example that the confounder (L_k) affects the treatment (A_k) `and` the treatment (A_k) affects in-turn again the confounder (L_{k+1}). And this **treatment-confounder feedback** can be caused directly or even go over unmeasured variables U_k. Remark: When the two red edges would be removed, we would still have time-varying confounders but not anymore treatment-confounder feedback.

An example why traditional methods fail

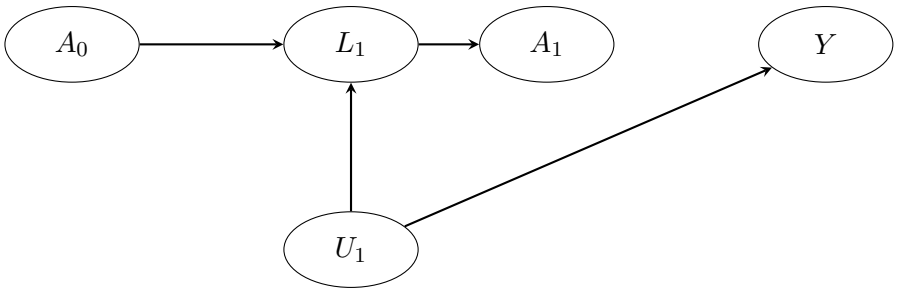

Figure 2.3: Corresponds to Figure 20.3 in Hernan and Robins (2023). Illustrates a sequentially randomized trial with two time-points $k = 0, 1$. Furthermore, treatment A_0 is unconditionally (or marginally) randomized and A_1 is randomized only with respect to confounders L_1. Since no edge from A_1, L_1, A_1 to Y is drawn, there is no effect assumed.

We are interested in estimating the **average causal treatment effect** of the static treatment strategy "always treat" $(a_0 = 1, a_1 = 1)$ versus the static treatment strategy "never treat" $(a_0 = 0, a_1 = 0)$ on the outcome Y. Thus the **average causal treatment effect** is

$$\mathbb{E}\left[Y^{a_0=1,a_1=1}\right] - \mathbb{E}\left[Y^{a_0=0,a_1=0}\right] \tag{2.2}$$

and should be 0 since there are no forward directed paths possible from A_0 or A_1 to Y. In addition, since there are no edges from unmeasured variables U into the treatment A, we can argue Figure 2.3 represents indeed a randomized trial and thus we should be able to use originating data (A_0, L_1, A_1, Y) to conclude (2.2) is equals to 0.

Even though if sequential exchangeability is given (when conditioning on confounder L_1), if we have time-varying confounders **and** treatment-confounder feedback, traditional methods can not correctly adjust for those confounder and thereby induce bias.

2.2.2 The bias of traditional methods (20.2)

Table 2.1: Table 20.1 in Hernan and Robins (2023)

A_0	L_1	A_1	N	Mean Y
0	0	0	2400	84
0	0	1	1600	84
0	1	0	2400	52
0	1	1	9600	52
1	0	0	4800	76
1	0	1	3200	76
1	1	0	1600	44
1	1	1	6400	44

- $\mathbb{P}(A_0 = 1) = 0.5$
- $\mathbb{P}(A_1 = 1 \mid L_1 = 0) = 0.4$ and $\mathbb{P}(A_1 = 1 \mid L_1 = 1) = 0.8$

We can confirm the causal effects of A_0 and A_1 (conditinal on the past) are indeed zero when we treat A_0 and A_1 seperately as time-fixed treatments:

1. The average causal effect in the stratum $(A_0 = 0, L_1 = 0)$:

$$\mathbb{E}\left[Y^{a_1=1} \mid A_0 = 0, L_1 = 0\right] - \mathbb{E}\left[Y^{a_1=0} \mid A_0 = 0, L_1 = 0\right] = 84 - 84 = 0$$

2. The average causal effect in the stratum $(A_0 = 0, L_1 = 1)$:

$$\mathbb{E}\left[Y^{a_1=1} \mid A_0 = 0, L_1 = 1\right] - \mathbb{E}\left[Y^{a_1=0} \mid A_0 = 0, L_1 = 1\right] = 52 - 52 = 0$$

3. The average causal effect in the stratum $(A_0 = 1, L_1 = 0)$:

$$\mathbb{E}\left[Y^{a_1=1} \mid A_0 = 1, L_1 = 0\right] - \mathbb{E}\left[Y^{a_1=0} \mid A_0 = 1, L_1 = 0\right] = 76 - 76 = 0$$

4. The average causal effect in the stratum $(A_0 = 1, L_1 = 1)$:

$$\mathbb{E}\left[Y^{a_1=1} \mid A_0 = 1, L_1 = 1\right] - \mathbb{E}\left[Y^{a_1=0} \mid A_0 = 1, L_1 = 1\right] = 44 - 44 = 0$$

Joining A_0 and A_1 as a joint time-varying treatment is not so easy even though the identifiablity conditions (???) hold and thus the data should be well enough to estimate the effect. Comparing the two treatment strategies "always treat" versus "never treat":

$$\mathbb{E}\left[Y^{a_0=1,a_1=1}\right] - \mathbb{E}\left[Y^{a_0=0,a_1=0}\right]$$

should be equals zero according to the g-null theorem (???).

With the classical methods this is however not the case:

$$\mathbb{E}\left[Y \mid A_0 = 1, A_1 = 1\right] - \mathbb{E}\left[Y \mid A_0 = 0, A_1 = 0\right] = 54.6666667 - 68 = -13.3333333$$
$$\mathbb{E}\left[Y \mid A_0 = 1, L_1 = 0, A_1 = 1\right] - \mathbb{E}\left[Y \mid A_0 = 0, L_1 = 0, A_1 = 0\right] = 76 - 84 = -8$$
$$\mathbb{E}\left[Y \mid A_0 = 1, L_1 = 1, A_1 = 1\right] - \mathbb{E}\left[Y \mid A_0 = 0, L_1 = 0, A_1 = 0\right] = 44 - 52 = -8$$

2.2.3 Why traditional methods fail (20.3)

Stratification can not handle treatment-confounder feedback

In our example in Figure 2.3, stratification with respect to L_1 is problematic. This, because L_1 is a collider for association measure A_0 and thus opens the path $A_0 \rightarrow L_1 \leftarrow U_1 \rightarrow Y$. This means that stratification induces a non-causal relationship between treatment A_0 and unmeasured variable U_1 and therefore also between A_0 and outcome Y, within levels of L_1. Thereby, stratification for L_1 eliminates confounding for A_1 but introduces **selection bias** for A_0.

2.3 G-METHODS FOR TIME-VARYING TREATMENTS (21)

2.3.1 The g-formula for time-varying treatments (21.1)

2.4 Transformation models

Hothorn (2020) nicely proposes a prospective to unify a wide range of statistical models by moving to conditional distributions and thus leaves the models relying on conditional expectation behind. We get there by rearranging the familiar model, noted by Equation, to model the error term $\boldsymbol{\varepsilon}$ as

$$\frac{\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}}{\sigma} = \boldsymbol{\varepsilon}$$

This is done because the error term $\boldsymbol{\varepsilon}$ is the only stochastic component of the model and in this transformed linear model framework we specify the error terms to be standard normally distributed with $\boldsymbol{\varepsilon}[i] \sim \mathcal{N}(0,1)$. Moreover, we can treat the constant term from the least-squares method separately by letting $\boldsymbol{\beta} = \left[\alpha, \tilde{\boldsymbol{\beta}}\right]$ and $\boldsymbol{X} = \left[\mathbf{1}, \tilde{\boldsymbol{X}}\right]$. For one observation, the model takes then the form

$$\frac{\boldsymbol{y}[i] - \alpha - \tilde{\boldsymbol{X}}[i,]\tilde{\boldsymbol{\beta}}}{\sigma} = \boldsymbol{\varepsilon}[i] \sim \mathcal{N}(0,1)$$

Modelling via conditional distribution function, this turns to

$$\mathbf{P}(Y[i] \leq \boldsymbol{y}[i] \mid \tilde{\boldsymbol{X}}[i,]) = \Phi\left(\frac{\boldsymbol{y}[i] - \alpha - \tilde{\boldsymbol{X}}[i,]\tilde{\boldsymbol{\beta}}}{\sigma}\right) \tag{2.3}$$

and to make sense of the name *transformation model* we further reformulate to

$$\mathbf{P}(Y[i] \leq \boldsymbol{y}[i] \mid \tilde{\boldsymbol{X}}[i,]) = \Phi\left(\underbrace{-\frac{\alpha}{\sigma}}_{\theta_0} + \underbrace{\frac{1}{\sigma}}_{\theta_1}\boldsymbol{y}[i] - \tilde{\boldsymbol{X}}[i,]\underbrace{\frac{\tilde{\boldsymbol{\beta}}}{\sigma}}_{\boldsymbol{\beta}_{\text{tram}}}\right) = \Phi\left(\theta_0 + \theta_1\boldsymbol{y}[i] - \tilde{\boldsymbol{X}}[i,]\boldsymbol{\beta}_{\text{tram}}\right) \tag{2.4}$$

where we see that the number of parameters to be estimated simultaneously is now $p+1$ which is due to $\theta_1 = \sigma^{-1}$. This means that θ_1 is not estimated independently from $\boldsymbol{\beta}_{\text{tram}}$ as it is the case in the least-squares setup.

Now, we introduce the transformation function $h(\boldsymbol{y}[i]|\boldsymbol{\theta})$ which is in this particular case $\theta_0 + \theta_1\boldsymbol{y}[i]$ and the purpose of it is doing the best it can to transform the response $\boldsymbol{y}$ to follow the distribution we want, which is here a standard normal distribution $\mathcal{N}(0,1)$ specified by $\Phi(z) = F_Z(z)$:

$$\mathbf{P}(Y[i] \leq \boldsymbol{y}[i] \mid \tilde{\boldsymbol{X}}[i,]) = F_Z\left(h(\boldsymbol{y}[i]|\boldsymbol{\theta}) - \tilde{\boldsymbol{X}}[i,]\boldsymbol{\beta}_{\text{tram}}\right) \tag{2.5}$$

Equation (2.5) describes the general specification of a transformation model as it is used in the `tram` package (Hothorn, 2020). The transformation function in (2.4) is linear and gets fitted by executing the command `tram::Lm` in R. However, we are by no means limited to this linearity and sometimes it is also necessary to use more complex transformations to assure our model is well-specified. Similarly as we see sometimes log or square-root transformed responses as an attempt to assure normality, we can use highly flexible functions such as splines to get a data-driven transformation. Such functions easily help to transform the outcome, which only has to be at least ordinal, to follow the distribution we want (not limited to normal distribution). The only restriction we must respect is that the transformation function is monotone, not strictly though. Whereas in the linear model so far we have estimated the coefficients via the least-squares method, we estimate them now by optimizing the likelihood. For more details with respect to the underlying functionalities of the `tram` package we refer the reader to Hothorn (2020) and for more theoretical issues to Hothorn *et al.* (2017).

2.4.1 An example

Figure 2.4 illustrates what a high and low condition number means in terms of model fitting. Both plots show data that is constructed by a simple equation

$$\boldsymbol{y} = 4 + 2 \cdot \boldsymbol{X}[,1] + 2 \cdot \boldsymbol{X}[,2] + \boldsymbol{\varepsilon} \cdot \sigma \tag{2.6}$$

where the explanatory variables within $\boldsymbol{X}$ are $n = 50$ realizations of a multivariate normal distribution as

$$X[i,] \sim \mathcal{N}\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \rho & 0 \\ 0 & \rho \end{pmatrix}\right)$$

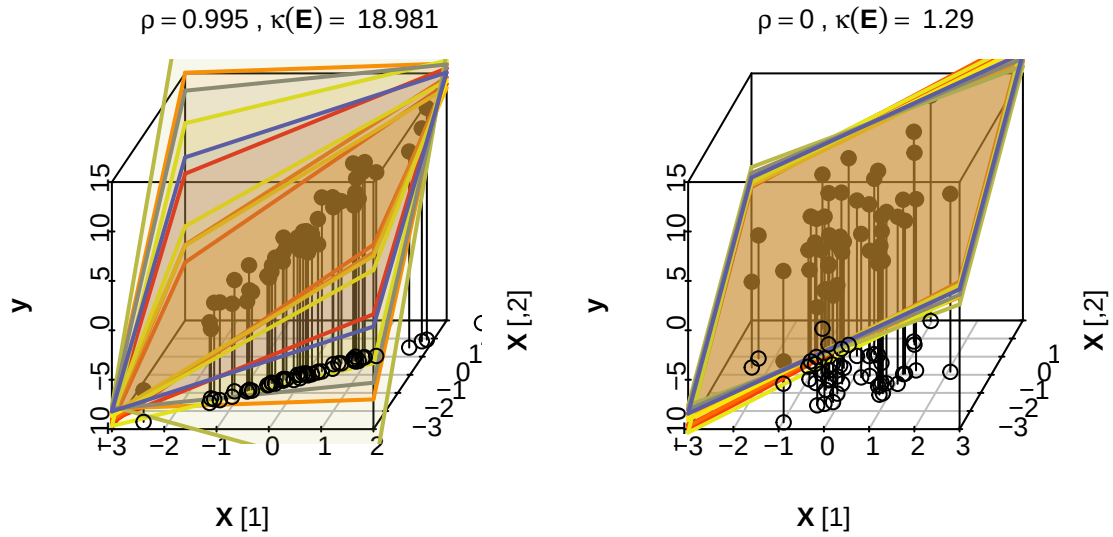

Figure 2.4: Impact of collinearity on the instability of estimates.

This allows to tune the amount of collinearity within the system by specifying ρ. On the left plot it is chosen to be high with $\rho = 0.995$ and on the right side low with $\rho = 0$. By bootstrapping the original sample 10 times and subsequent model fitting we can visualize the instability that collinearity causes. Because when we plot the planes that represent the area where the models would see $\hat{\boldsymbol{y}} = \hat{\alpha} + \hat{\boldsymbol{\beta}}[1]\boldsymbol{X}[,1] + \hat{\boldsymbol{\beta}}[2]\boldsymbol{X}[,2]$ we note on the left side with high collinearity ($\kappa(\boldsymbol{E}) = 18.981$) that the planes are quite different from each other, whereas on the right side ($\kappa(\boldsymbol{E}) = 1.29$) they seem to be very similar and thus stable. Table 2.2 shows the corresponding variance decomposition proportion matrices $\boldsymbol{\Pi}$ and the least-squares model results of the original data sets.

Table 2.2: Variance decomposition matrices as introduced by Belsley in the first row and summary output of the multiple linear regression models on the second row. Left side corresponds to the example with higher collinearity and the right table for the lower.

mu	cond_ind	const	$\boldsymbol{X}[,1]$	$\boldsymbol{X}[,2]$
1.412	1.000	0.000	0.003	0.003
1.000	1.412	0.939	0.000	0.000
0.074	18.981	0.061	0.997	0.997

	$\hat{\beta}$	se $\left(\hat{\beta}\right)$	t-value	p-value
Intercept	4.17	0.17	24.89	< 0.0001
x1	2.35	1.61	1.46	0.15
x2	1.85	1.58	1.17	0.25

mu	cond_ind	const	$\boldsymbol{X}[,1]$	$\boldsymbol{X}[,2]$
1.144	1.000	0.294	0.230	0.206
0.951	1.202	0.003	0.448	0.613
0.887	1.290	0.703	0.323	0.181

	$\hat{\beta}$	se $\left(\hat{\beta}\right)$	t-value	p-value
Intercept	4.17	0.17	24.89	< 0.0001
x1	2.16	0.18	12.19	< 0.0001
x2	2.13	0.15	14.08	< 0.0001

But why going trough all the trouble with collinear variables and the detrimental effects that come with it and not simply drop one or some of the affected variables? Figure 2.5 visualizes the model fits when the variable $\boldsymbol{X}[,2]$ is neglected although truly it has very well an effect on $\boldsymbol{y}$ as is visible in Equation (2.6). We see on the right plot for low collinearity the 95% confidence interval for $\hat{\boldsymbol{\beta}}[1]$ does cover the true effect of 2 whereas for the case with high collinearity this seems to be not the case. This demonstrates that it is not so easy to simply get rid of some variables as this may introduce bias to some extent.

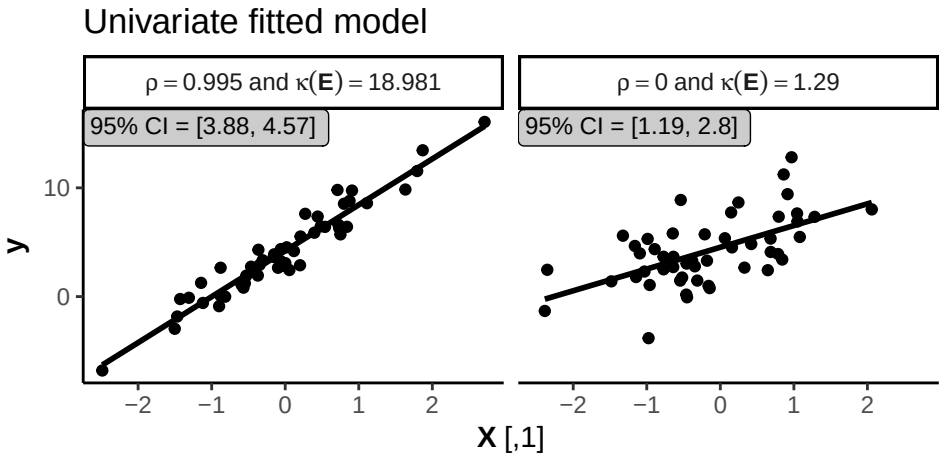

Figure 2.5: Univariate fitted model (`y~x1`) of the same data sets as in Figure 2.4. The slope of the line represents $\hat{\boldsymbol{\beta}}[1]$ which would be truly 2 and the confidence interval thereof is given in the box. Obviously, only the right plot with low collinearity seems to capture the true effect whereas with higher collinearity the estimate is biased.

2.5 Differences between `lm` and `tram::Lm`

The parametrization and the chosen estimation approaches differ between `lm` and `tram::Lm` and in this section we are going to compare what these differences mean from a theoretical perspective.

2.5.1 Maximum-Likelihood estimation for the linear regression model

We can show that independent of the estimating procedure, with the parametrization as specified in Equation (2.3) we will end up at the very same optimization problem if we go over the profile likelihood. The approximate log-likelihood of a sample that is treated as exact is

$$\begin{aligned}
\ell(\boldsymbol{\beta}, \sigma | \boldsymbol{y}) &= -N \log(\sigma) - \frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \left(-\frac{\alpha}{\sigma} + \frac{1}{\sigma} \boldsymbol{y}[i] - \tilde{\boldsymbol{X}}[i,] \frac{\tilde{\boldsymbol{\beta}}}{\sigma} \right)^2 \\
&= -N \log(\sigma) - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^{N} \left(\boldsymbol{y}[i] - \alpha - \tilde{\boldsymbol{X}}[i,] \tilde{\boldsymbol{\beta}} \right)^2 \\
&= -N \log(\sigma) - \frac{N}{2} \log(2\pi) - \frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \qquad (2.7)
\end{aligned}$$

We can now employ the profile likelihood where we treat σ as the nuisance parameter:

$$\begin{aligned}
\frac{d\ell(\boldsymbol{\beta}, \sigma | \boldsymbol{y})}{d\sigma} \Big|_{\hat{\sigma}} = -N\sigma^{-1} + \hat{\sigma}^{-3} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &\stackrel{!}{=} 0 \\
\hat{\sigma}^{-3} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) &\stackrel{!}{=} N\hat{\sigma}^{-1} \\
\hat{\sigma}^2 &\stackrel{!}{=} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) / N
\end{aligned}$$

Plugging $\hat{\sigma}$ into (2.7), we see that $\hat{\sigma}$ vanishes from the equation which is handy:

$$\begin{aligned}
\frac{d\ell(\boldsymbol{\beta}, \hat{\sigma} | \boldsymbol{y})}{d\boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} = -N \log\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right) \frac{d}{d\boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} &\stackrel{!}{=} 0 \\
\log\left((\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \right) \frac{d}{d\boldsymbol{\beta}} \Big|_{\hat{\boldsymbol{\beta}}} &\stackrel{!}{=} 0
\end{aligned}$$

Since the log is a monotone function, the maximum likelihood is also found by minimizing the term $(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^\top (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ and thus the maximum-likelihood estimator $\hat{\boldsymbol{\beta}}$ is the very same as for the least-squares estimator described in

2.5.2 Maximum-Likelihood estimation for the transformation model equivalent (`tram::Lm`)

The approximate log-likelihood with the parametrization used for the `tram::Lm` model specified in Equation (2.4) is

$$\ell(\boldsymbol{\beta}_{\text{tram}}, \theta_0, \theta_1 | \boldsymbol{y}) = +N \log(\theta_1) - \frac{N}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{N} \left(\theta_0 + \theta_1 \boldsymbol{y}[i] - \tilde{\boldsymbol{X}}[i,] \boldsymbol{\beta}_{\text{tram}} \right)^2 \qquad (2.8)$$

which has one parameter (θ_1) more to simultaneously estimate.

The design matrix in this setup is different. For `lm`, $\boldsymbol{X}$ contained the variables that will be used to explain the outcome $\boldsymbol{y}$. But this can also be reformulated in terms of using the *variables including the outcome* to explain the *error* $\boldsymbol{\varepsilon}[i] \sim \mathcal{N}(0, 1)$. Since for `tram::Lm` the parameter θ_1 is attached to the outcome $\boldsymbol{y}$, the collinearity constellation is not only restricted to the $\boldsymbol{X}$ space but extends onto $[\boldsymbol{y}, \boldsymbol{X}]$. This basically implies that the better the outcome $\boldsymbol{y}$ is explainable by $\boldsymbol{X}$, the higher the collinearity and thus the larger the effects caused by it. This is important to keep in mind.

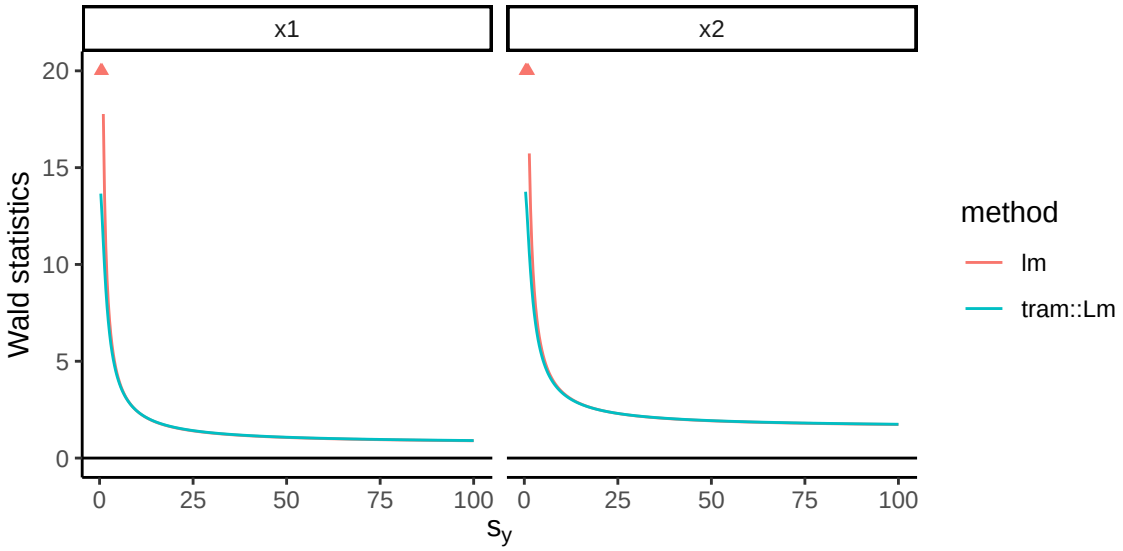

Figure 2.6: Simulating data as $\boldsymbol{y} = 10 + 2\boldsymbol{x}_1 + 2\boldsymbol{x}_2 + s_y \cdot \boldsymbol{\varepsilon}$ with $(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{\varepsilon}) \sim \mathcal{N}_{3n}(0, 1), n = 100$. The scaling factor s_y is iterated on a grid between 0.3333333 and 100 where a low scaling factor means that the outcome $\boldsymbol{y}$ is well explainable and thus collinearity for `tram::Lm` is higher. Wald statistics are plotted restricted to have maximum values of 20 and points laying above are illustrated as triangles.

Appendix A

Appendix

A.1 Approximate likelihood

Since real-life data is always observed in intervals $\boldsymbol{D} = (\underline{y}, \overline{y}]$ and is never exact (although treated as if), the likelihood contribution of one observation is:

$$l(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D}) = \mathbf{P}(\underline{y} < Y \leq \overline{y}|\boldsymbol{X} = \boldsymbol{x}) = F_Z\left(h_Y(\overline{y}|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) - F_Z\left(h_Y(\underline{y}|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right)$$

which is the exact likelihood as originally introduced by Fisher. The approximated likelihood for a continuous response is obtained by making the interval around the "observed" value y negligibly small $\boldsymbol{D} = (y - \epsilon, y + \epsilon]$ and thus the likelihood is approximated as

$$\begin{aligned}
l(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D}) &= F_Z\left(h_Y(y + \epsilon|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) - F_Z\left(h_Y(y - \epsilon|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) \\
&= \int_{y-\epsilon}^{y+\epsilon} F_Z'\left(h_Y(u|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) h_Y'(u|\boldsymbol{\theta}) du \\
&\approx f_Z\left(h_Y(y|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) h_Y'(y|\boldsymbol{\theta}) \cdot 2\epsilon \\
&\propto f_Z\left(h_Y(y|\boldsymbol{\theta}) - \tilde{\boldsymbol{x}}\boldsymbol{\beta}_{\text{tram}}\right) h_Y'(y|\boldsymbol{\theta})
\end{aligned}$$

The joint likelihood for several observations assuming independence is:

$$L(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D_1}, ..., \boldsymbol{D_N}) = \prod_{i=1}^{N} l(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D_i})$$

where it is theoretically and computationally convenient to operate on the log scale

$$\ell(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D_1}, ..., \boldsymbol{D_N}) = \sum_{i=1}^{N} \log\left(l(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D_i})\right)$$

The resulting maximum log-likelihood estimator is then:

$$\hat{\boldsymbol{\beta}}_{\text{tram}}, \hat{\boldsymbol{\theta}} = \operatorname{argmax} \ell(\boldsymbol{\beta}_{\text{tram}}, \boldsymbol{\theta}|\boldsymbol{D_1}, ..., \boldsymbol{D_N})$$

A.2 Computational reproducibility

```
sessionInfo()

## R version 4.2.2 Patched (2022-11-10 r83330)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 20.04.5 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.9.0
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.9.0
##
## locale:
##  [1] LC_CTYPE=de_CH.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=de_CH.UTF-8        LC_COLLATE=de_CH.UTF-8
##  [5] LC_MONETARY=de_CH.UTF-8    LC_MESSAGES=de_CH.UTF-8
##  [7] LC_PAPER=de_CH.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=de_CH.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] tram_0.8-0           mlt_1.4-3            basefun_1.1-2
##  [4] variables_1.1-1      forcats_0.5.1        stringr_1.4.0
##  [7] dplyr_1.0.9          purrr_0.3.4          readr_2.1.2
## [10] tidyr_1.2.0          tibble_3.1.7         ggplot2_3.4.0
## [13] tidyverse_1.3.1      RColorBrewer_1.1-3   xtable_1.8-4
## [16] biostatUZH_2.0.2     MASS_7.3-58          survival_3.4-0
## [19] tableone_0.13.2      Collinearity_1.1.2   mvtnorm_1.1-3
## [22] scales_1.2.0         scatterplot3d_0.3-41 knitr_1.39
##
## loaded via a namespace (and not attached):
##  [1] nlme_3.1-160        fs_1.5.2            cmprsk_2.2-11
##  [4] lubridate_1.8.0     httr_1.4.3          numDeriv_2016.8-1.1
##  [7] tools_4.2.2         backports_1.4.1     utf8_1.2.2
## [10] R6_2.5.1            mgcv_1.8-41         DBI_1.1.2
## [13] colorspace_2.0-3    withr_2.5.0         tidyselect_1.1.2
## [16] compiler_4.2.2      orthopolynom_1.0-6  cli_3.4.1
## [19] rvest_1.0.2         alabama_2022.4-1    xml2_1.3.3
## [22] sandwich_3.0-1      labeling_0.4.2      quadprog_1.5-8
## [25] digest_0.6.29       minqa_1.2.4         pkgconfig_2.0.3
## [28] lme4_1.1-29         dbplyr_2.1.1        highr_0.9
## [31] rlang_1.0.6         readxl_1.4.0        rstudioapi_0.13
## [34] farver_2.1.0        generics_0.1.2      zoo_1.8-10
## [37] jsonlite_1.8.0      magrittr_2.0.3      polynom_1.4-1
## [40] Formula_1.2-4       coneproj_1.16       Matrix_1.5-1
## [43] Rcpp_1.0.8.3        munsell_0.5.0       fansi_1.0.3
## [46] lifecycle_1.0.3     stringi_1.7.6       multcomp_1.4-19
## [49] BB_2019.10-1        grid_4.2.2          crayon_1.5.1
## [52] lattice_0.20-45     haven_2.5.0         splines_4.2.2
## [55] hms_1.1.1           pillar_1.7.0        boot_1.3-28
```

```
## [58] codetools_0.2-18    reprex_2.0.1        glue_1.6.2
## [61] evaluate_0.15       mitools_2.4         modelr_0.1.8
## [64] vctrs_0.5.1         nloptr_2.0.1        tzdb_0.3.0
## [67] psy_1.2             cellranger_1.1.0    gtable_0.3.0
## [70] assertthat_0.2.1    xfun_0.31           broom_0.8.0
## [73] survey_4.1-1        TH.data_1.1-1       ellipsis_0.3.2
```

Bibliography

Hernan, M. and Robins, J. (2023). *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press. 3, 4, 5, 6

Hothorn, T. (2020). Most likely transformations: The mlt package. *Journal of Statistical Software*, **92**, https://doi.org/10.18637/jss.v092.i01. 8

Hothorn, T., Möst, L., and Bühlmann, P. (2017). Most likely transformations. *Scandinavian Journal of Statistics*, **45**, 110–134. 8