More than just a number: modelling biological aging and vulnerability

Sluiskes, M.H.

Chapter 3

A word of caution on the regression to the mean phenomenon in (biological) age prediction

Marije Sluiskes, Jelle Goeman, Marian Beekman, Eline Slagboom, Hein Putter and Mar Rodríguez-Girondo

Based on: A word of caution on the regression to the mean phenomenon in (biological) age prediction. *PubPeer* (2023).
https://pubpeer.com/publications/A9B0D98C173B74C41F55C340EF4945#1

3.1 Introduction

In regression models, predicted values tend towards the mean of the outcome variable in the data set on which the model was fitted. This phenomenon, which occurs whenever outcome and predictor(s) are not perfectly correlated, is known as 'regression to the mean' [1]. It was first observed by English polymath Francis Galton in the 19$^{\text{th}}$ century [2]: while studying human height, Galton found that tall parents tend to tall children, but not as tall as their parents. Similarly, short parent generally have short children, but not as short as their parents. Galton called this phenomenon "regression to mediocrity", now known as regression to the mean.

Due to regression to the mean, in a regression model true low outcome values will on average be overestimated and true high values will be underestimated. Though this phenomenon has primarily been discussed in the context of (multiple) linear regression, it holds for all types of regression models.

Regression to the mean is relevant in the context of cross-sectional biological age prediction. Biological age is supposed to be a measure of one's true underlying aging status, instead of merely measuring the number of years alive since birth [3]. A commonly used approach to predict biological age is to perform linear regression on cross-sectional data [4]. Typically, chronological age C is taken as the outcome variable and regressed on a set of (candidate) biomarkers of aging X that were measured at the same point in time as chronological age. The model's predicted chronological age is considered to be informative of one's biological age, i.e. first a model on chronological age is fitted, $C = \beta_0 + \sum_{i=1}^{m} \beta_i x_i$, where m denotes the number of included markers. This model is used to obtain chronological age predictions $\hat{C}$, which are interpreted as predictions of biological age B. This means that the residuals (the differences between C and $\hat{C}(= B)$) of this model are interpreted as meaningful quantities in their own right.

The correlation between chronological age and the markers is not perfect: there is a (often significant) significant source of random error. Hence, there will be regression to the mean: predicted values for C will tend to regress toward the training data sample's mean chronological age. Individuals younger than this sample mean age will obtain predicted ages that are on average too high, and individuals older than this mean age will receive predicted ages that are too low.

That the regression to the mean phenomenon is a factor to consider when predicting biological age has been known for decades: see e.g. Dubina et al. [5] or Hochschild [6]. Nevertheless, it is not always recognized as such, leading to premature or invalid

conclusions. The risk of this is especially high when the size of the absolute differences between predicted and true C (i.e. the model's residuals) is directly used and interpreted as an indication of accelerated or decelerated aging, and compared between groups that have different ages. This short chapter therefore aims to explain the regression to the mean phenomenon in simple terms and provide an illustrative example of what can go wrong. It is meant to help aging researchers to better understand regression to the mean and to help them spot analyses and scenarios in which this phenomenon might affect conclusions.

3.2 Regression to the mean

Regression to the mean is a statistical phenomenon that occurs whenever two variables of interest, X and Y, are imperfectly correlated. On average, when a specific measurement of X deviates from the mean of X, the corresponding measurement of Y will be less far away from the mean of Y. This occurs whenever the correlation coefficient between X and Y is smaller than 1.

It is easiest to understand this phenomenon by considering two extreme cases. Assume that X and Y are normally distributed with mean μ and standard deviation σ. (This is for simplicity of the subsequent formulas, but these can be adjusted without loss of generality.) Denote the correlation coefficient between X and Y by r.

For a given subject i, draw measurement x_i from X. The expected value for measurement y_i is then $\mu + r(x_i - \mu) = rx_i + (1 - r)\mu$. If there is a perfect correlation between X and Y (i.e. $r = 1$), $E[y_i] = x_i$. If they are completely independent (i.e. $r = 0$), $E[y_i] = \mu$.

In reality, for any X and Y, r will generally be somewhere in between these two extreme cases. Hence, if subjects with extreme measurements of X are selected, the average of the measurements for Y for these subjects will be closer to the mean.

3.3 Illustration

The illustration below was inspired by a recent paper from Daunay et al. [7], in which four previously proposed epigenetic clocks based on a small number of CpG sites are evaluated and compared in three sets of samples: semi-supercentenarians (mean age ± sd: 101.3 ± 1.4), offspring of nonagenarians and centenarians (61.2 ± 6.1), and individuals from the general population (56.0 ± 4.7). The authors predicted age, called

the predictions "methylation age" and interpreted the difference between predicted and true chronological age. They found larger negative differences for the semi-centenarians and the offspring than for the individuals from the general population, and concluded that there is therefore an indication of decelerated epigenetic and biological aging in the first two groups (semi-supercentenarians and offspring of nonagenarians and centenarians) as compared to the last (general population controls).

The simple simulated data example in this section illustrates that the possibility cannot be ruled out that the conclusion of Daunay et al. [7] is due to regression to the mean instead of a true underlying biological phenomenon.

We assume the model $C = \beta X + \epsilon$, where X is some marker of chronological age and C (some scaled version of) chronological age. X and ϵ are two independent normally distributed variables with mean equal to 0 and standard deviation equal to 6. The coefficient β is equal to 2.

We generate a training data set and a test data set, both of size $n = 10,000$, and fit two models on the training data: a simple linear regression model (LR) and a generalized boosted regression model (GBM). To fit the GBM, we use the function `gbm()` from the package `gbm` (version 2.2.2), using the default settings but specifying as distribution `gaussian`, corresponding to a squared error loss.

In Figures 3.1 and 3.2 the residuals $(\hat{C} - C)$ of both models are plotted against C. In both figures there is a clear downward sloping pattern visible: people with higher chronological ages (higher C) on average have lower predicted chronological ages (lower $\hat{C}$). This is due to regression to the mean.

Next, we split the test set in two groups, based on their observed value for C. This would be similar to comparing groups from the general population defined by their chronological age (e.g. young and old individuals from the same population), in line with the approach of Daunay et al. [7]. The models fitted on the training data are used to obtain predictions for C for these two groups.

We now compare the values of the residuals $(\hat{C} - C)$ between these two groups. Ideally, they should be similarly distributed and centered around zero: after all, the random noise distribution is the same for the two groups. However, Figures 3.3 and 3.4 show that this is in fact not the case: the mean value of the residuals in the young group is slightly above 0, in the old group slightly below 0. If interpreted without care, this would lead to the conclusion that members of the old group are on average biologically young for their age, while members of the young group are on average biologically old for their age. As we know from the data-generating mechanism, this is not true: the observed pattern is only due to the regression to the mean

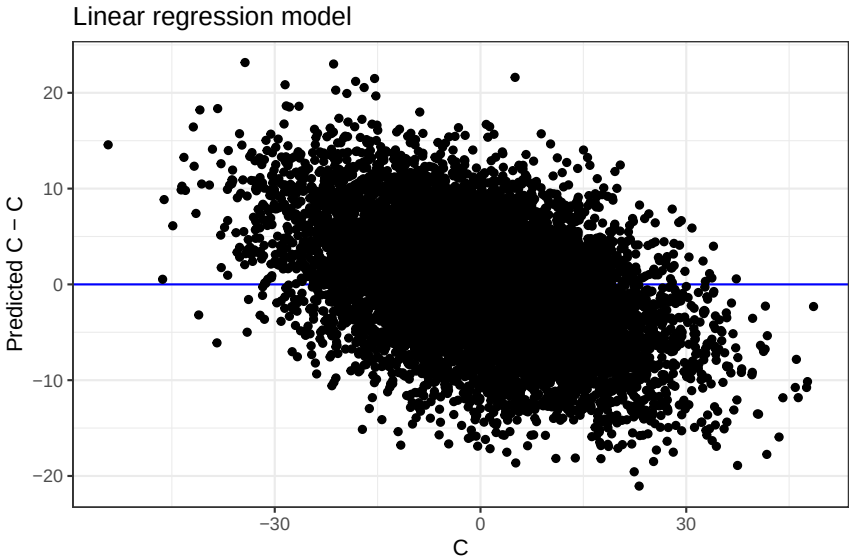

Figure 3.1: Plot of residuals (predicted C minus true C) against C. Predictions were obtained using a linear regression model.

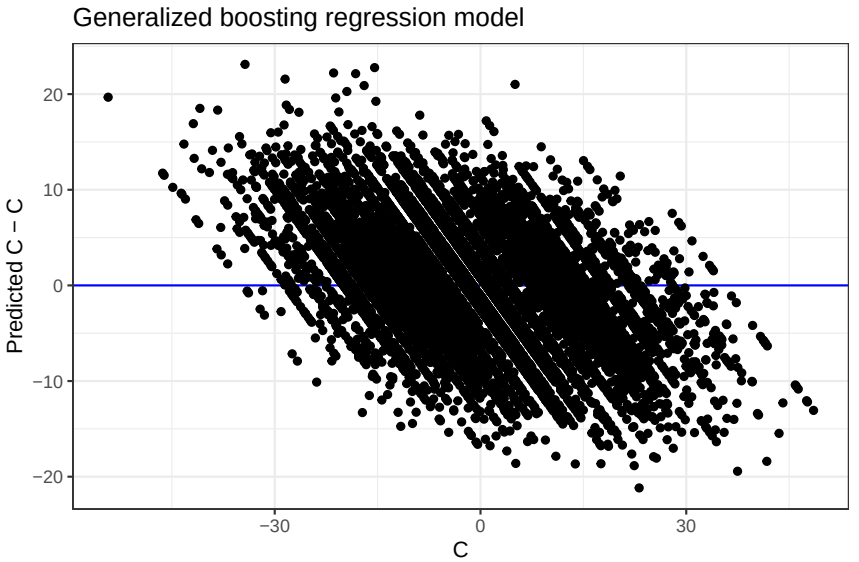

Figure 3.2: Plot of residuals (predicted C minus true C) against C. Predictions were obtained using a generalized boosting model.

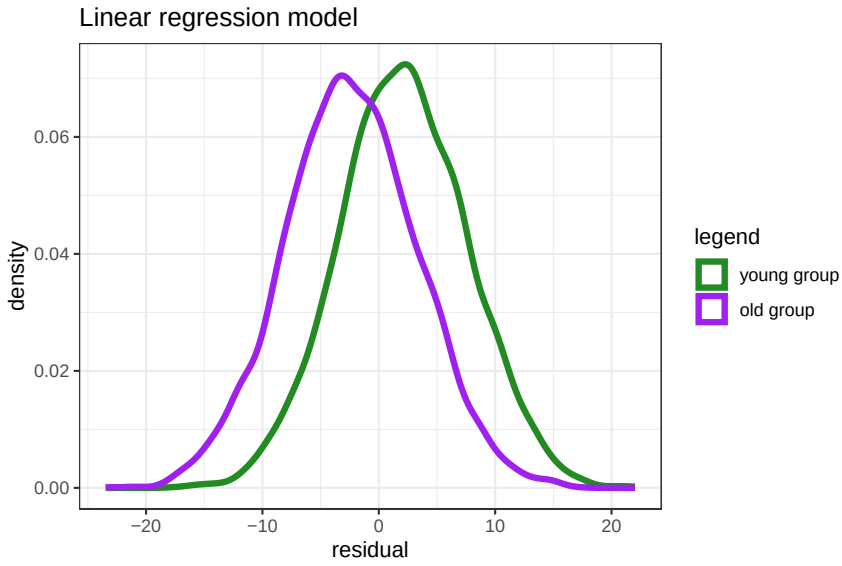

Figure 3.3: Density plot of residuals (predicted C minus true C). Predictions were obtained using a linear regression model.

phenomenon.

3.4 Discussion

In this short chapter we have shown that regression to the mean is a factor of importance when predicting biological age with regression models that use chronological age as the outcome of interest. As a consequence, when the difference between true and predicted chronological age is directly interpreted as an indicator of one's biological aging status, individuals younger than the training sample mean will on average always have a positive difference (indicating accelerated aging) and those older than the sample mean a negative difference (indicating decelerated aging). This is particularly problematic when directly interpreting the model's residuals and using them to compare individuals from different age groups, as we have illustrated with our synthetic data example.

There are several approaches that alleviate the problem, although none of them are perfect. The simplest approach is to only compare predicted age of individuals if these individuals are matched by age, which is the approach taken by e.g. Horvath et al. [8]. However, any differences between groups of the same age should still only

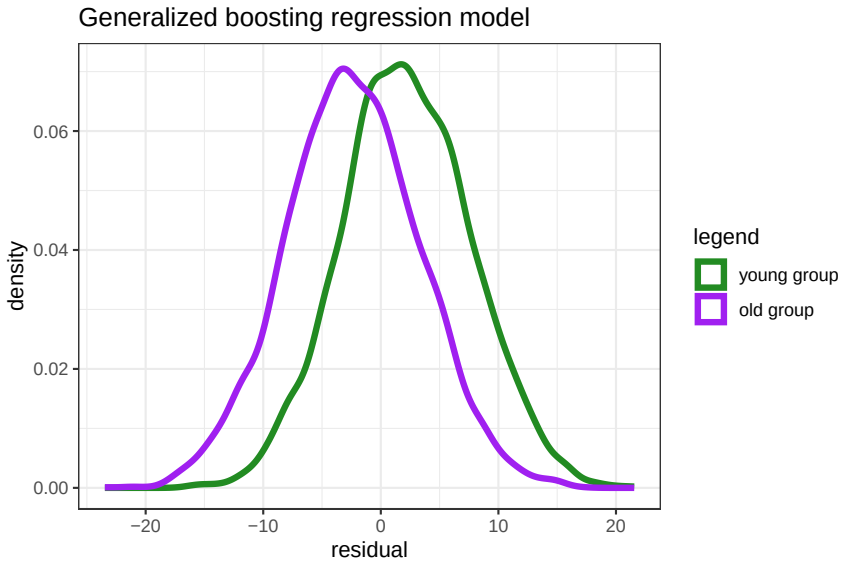

Figure 3.4: Density plot of residuals (predicted C minus true C). Predictions were obtained using a generalized boosting model.

be interpreted in relative terms, not in absolute ones. Another approach is to 'regress out' the effect of chronological age by regressing C on the residuals $\hat{C} - C$, such that the resulting residuals are in fact no longer correlated with chronological age. This is a popular approach, but is rather ad hoc. A final alternative is to rely on models that are not affected by regression to the mean. One such method is the Klemera-Doubal method [9], though this method comes with its own issues, as it relies on a strong and untestable assumption regarding the underlying biological age model and cannot easily be scaled to settings with many predictor variables. It should be noted that with all three these alternative approaches, the fundamental problem with any cross-sectional biological age prediction approach, as discussed in detail in Chapter 2, is not yet solved.

The most appealing way forward is to consider models built with longitudinal data, considering e.g. time-to-mortality as the outcome of interest. Regression to the mean can be a factor here as well, as most longitudinal models still rely on some form of regression, but the risk of misinterpretation is smaller because the residuals of longitudinal models generally cannot be directly be interpreted on an age-scale. It is therefore common practice to first regress out the effect of chronological age and consider the resulting residuals, which are no longer correlated with age, as a sort of

unitless measures of biological aging.

In conclusion, the regression to the mean phenomenon should be well-known among aging researchers to avoid misinterpretation issues. As we have illustrated, regression to the mean is not only a factor of importance in the context of (multiple) linear regression, but is a matter of concern for any regression model, including machine learning-type models such as boosting methods.

3.5 Bibliography

[1] J Martin Bland and Douglas G Altman. Statistics notes: some examples of regression towards the mean. *BMJ*, 309(6957):780, 1994.

[2] Francis Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.

[3] Juulia Jylhävä, Nancy L Pedersen, and Sara Hägg. Biological age predictors. *EBioMedicine*, 21:29–36, 2017.

[4] Marije H Sluiskes, Jelle J Goeman, Marian Beekman, P Eline Slagboom, Hein Putter, and Mar Rodríguez-Girondo. Clarifying the biological and statistical assumptions of cross-sectional biological age predictors: an elaborate illustration using synthetic and real data. *BMC Medical Research Methodology*, 24(1):58, 2024.

[5] TL Dubina, A Ya Mints, and EV Zhuk. Biological age and its estimation. III. Introduction of a correction to the multiple regression model of biological age and assessment of biological age in cross-sectional and longitudinal studies. *Experimental Gerontology*, 19(2):133–143, 1984.

[6] Richard Hochschild. Improving the precision of biological age determinations. Part 1: a new approach to calculating biological age. *Experimental Gerontology*, 24(4):289–300, 1989.

[7] Antoine Daunay, Lise M Hardy, Yosra Bouyacoub, Mourad Sahbatou, Mathilde Touvier, Hélène Blanché, Jean-François Deleuze, and Alexandre How-Kit. Centenarians consistently present a younger epigenetic age than their chronological age with four epigenetic clocks based on a small number of CpG sites. *Aging (Albany NY)*, 14(19):7718, 2022.

[8] Steve Horvath, Chiara Pirazzini, Maria Giulia Bacalini, Davide Gentilini, Anna Maria Di Blasio, Massimo Delledonne, Daniela Mari, Beatrice Arosio, Daniela Monti, Giuseppe Passarino, et al. Decreased epigenetic age of PBMCs from Italian semi-supercentenarians and their offspring. *Aging (Albany NY)*, 7(12):1159, 2015.

[9] Petr Klemera and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mechanisms of Ageing and Development*, 127(3):240–248, 2006.