House of Cards

Table of abbreviations

AA	"Accelerated ageing" of an individual, otherwise known as individual-specific "prediction error" or "residual" arising from a estimated model.
AR	Age residuals from a predictive model, synonomymous with AA. Our prefered term because (this name respects longstanding statistical convention: "acceleration" is a 2nd order quantity while AA is not).
MS	methylation state of an individual.
TS	transcriptome state.

Table of Variables

t	chronological age
$\hat{t}(X)$	predicted chronological age using X
X	possibly transformed predictors, MS
Z	causes of X variation
V	causes of AR or RL

Box: Questions about AR

- Predictive accuracy varies with the predictive quality of the predictors, for example it varies over different tissue types. Does this mean that there is more heterogeneity in the biological age of these tissue types? It is a prediction of this model, that they would be more heterogeneity in the biological age of tissues which demonstrate less reliability as predictors of chronological age. Q: does this explain why the success of age residuals in explaining organ function (Table 1 of HR) is greater in tissues with lower predictive accuracy. The scale of AR variation is artifactual.
- By construction, AR variation is unrelated to any t variation "explained" by MS (namely t_hat(X)). This is most clear in the case of a simple regression where the component variation of AR is orthogonal to the subspace explained by methylation. Presumably this means that Z.
- The usual interpretation of residuals, if there is noise. There is a well-known problem associated with exclamation bias due to noisy predictors.
- Does MS predict age-at-death, the least ambiguous measure of organ function. It is conceptually possible for an individual to have positive AR but to live longer. What then is AR?
- Residuals indicate the quality of a model. AR is used in various regression diagnostics for determining influential observations. If the empirical average value of residuals is not close to 0, it implies that the

model is systematically biased toward either over-prediction or under-prediction. If residuals contain a pattern, then the model is failing to explain some relations within the data.

- Does t_hat (or AR) predict RL?
- Epigenetic age as counterfactual age, i.e. the age you should have, were it not for other biological factors. If such biological factors can be found to explain the difference between your actual age and your counterfactual age, they are implicated in the ageing process. is based on fallacy that the subject counterfactual is the same thing as the population mean.
- t - t_hat is not causal
- t – t_hat is a counterfactual contrast (had my environment/genomics conferred average).
- In statistics the residuals are distinguished from the errors. Residual tests are used to check properties of the true error term. here.
- The interpretion of AR depends on the interpretation of the predictive model. Predictive models have famously poor interpretatability: that is not their goal.
- residual analysis: the task of identifying factors missing from the model which features - variables or transformations - when included in the model will improve the model accuracy. If the residual plot shows a non-random pattern wrt an independent variable, "transform" the data to use a linear model with nonlinear data.
- The most common residual plot shows yhat on the horizontal axis and the residuals on the vertical axis. If the assumptions regarding the error term, e., are satisfied, the residual plot will consist of a horizontal band of points. If the residual analysis does not indicate that the model assumptions are satisfied, it often suggests ways in which the model can be modified to obtain better results.

A very general way of describing what statistical tools do is: they separate the data (the variation we are interested in) into two parts, one that we are able to summarize (structure, explain, describe, and another part that is not (yet?) summarized (structured, explained, described), i.e. the remaining residual part of the variation. Generally analyzing the residuals produced by any statistical tool, provides us with a general approach to assess the quality and diagnose possible problems with the summary, as whenever you are able to discover any kind of systematic aspect (groups, outliers, any structure) you will know that the summary did not succeed in catching all systematic elements in the data.

here Below we will consider residuals from regression; you should however be aware that the general principles can be applied to any statistical tool; no matter what tool you are using, a residual is always:

residual = data - summary

well HR acknowledge that

new here

The analysis of residuals plays an important role in validating the regression model. If the error term in the regression model satisfies the four assumptions noted earlier, then the model is considered valid. Since the statistical tests for significance are also based on these assumptions, the conclusions resulting from these significance tests are called into question if the assumptions regarding e are not satisfied.

The ith residual is the difference between the observed value of the dependent variable, yi, and the value predicted by the estimated regression equation, yi. These residuals, computed from the available data, are treated as estimates of the model error, ϵ. As such, they are used by statisticians to validate the assumptions concerning e. Good judgment and experience play key roles in residual analysis.

Graphical plots and statistical tests concerning the residuals are examined carefully by statisticians, and judgments are made based on these examinations. The most common residual plot shows yhat on the horizontal axis and the residuals on the vertical axis. If the assumptions regarding the error term, e, are satisfied, the residual plot will consist of a horizontal band of points. If the residual analysis does not indicate that the model assumptions are satisfied, it often suggests ways in which the model can be modified to obtain better results.

new

"Residuals are differences between the one-step-predicted output from the model and the measured output from the validation data set. Thus, residuals represent the portion of the validation data not explained by the model."

Let, $y_i = \mathbf{x}_i'\beta + u_i$ Then, $\mathrm{Corr}(y_i, \hat{u}_i) = \sqrt{1 - E\mathbf{x}_i' \left(\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j'\right)^{-1} \mathbf{x}_i}$

If N is much bigger than p, hence a lot of h_{ii} would be close to the zero, meaning that the correlation between the residual and the response variable would be close to 1 for the bigger part of observations.

The term h_{ii} is also used in various regression diagnostics for determining influential observations.

It is also worth checking, whether residuals are homoscedastic (their variance does not change over time). If residuals are heteroscedastic, this means that the predictive power of the model is different for different sections of the data and, perhaps, it is worth thinking about dividing the dataset into two (or more) subsets in order to train two (or more) models, such that each model is specializing on the corresponding subset.

HAN rightly point out that in order to conclude that residual age or age ratio can be interpreted as accelerated ageing, one must conclude that they are not simply measurement error or intrinsic variability. For them to be established as genuine biological differences in the state of the individual one must first establish their association to relevant clinical factors including gender and BMI. we share with that the discrepancy between predicted agent age can be explained by other clinical factors like gender and BMI. Similarly this discrepancy is influenced by some SNPs. Puzzlingly, they go on to assert that if individuals differ in how quickly they age, then their methylomes should correspondingly diverge over time.sk

Notes

Simulation study

Zt+1 -> Xt ˆ | Zt -> Xt

Simplest case (no vertical dependence).

Use cross-validation.

#Waffle

Like neuroimaging the sheer multiplicity of high throughput data makes genomics susceptible to false positives.

The field of biological ageing research deserves conceptual clarity that is not offered by black-box machine learning and predictive algorithms. Age at death (or age at organ failure) is the least ambiguous measure of organ function or "biological age", c.f. survival analyses.

We study the quantity called accelerated ageing (AR). We find that most AR variation is unlikely to reflect individual differences in methylation status or biological age/organ function, as opposed to a myriad of non-biological factors. It remains open whether any component of AR reflects MS or BA Z(?). Neither do we see any reason to interpret AR counterfactually as the difference between an individual's actual biological age and their counterfactual biological age, had they been an average person (with average genome and exposure). This is disappointing given the promise that counterfactual accounts hold for formalizing causal hypotheses. Finally, the coefficients on AR in secondary model predicting health are likely biased. This is because AR is residual and therefore essentially random error. For simple linear regression the effect of AR is an underestimate of the coefficient, known as the attenuation bias. In non-linear models the direction of the bias is likely to be more complicated.[4][5]

A clever way to convert predictive failure to explanatory success. The strategy is doomed to be a victim of its own success. The more accurate the age predictor, the smaller the residual prediction error. There is at best a tension between estimating a predictive model, whose goal is to minimize error, and scientific interest in this error.

the residuals are your unexplained variance. the residuals from your model should be random, meaning they should not be correlated with independent variables (although they can be correlated with dependent variable).

Detailed residual analysis plays a crucial role in model assessment for basic linear models. Nonetheless much is known about it's behavior. The relation of AR to model quality, to the explanatory variables or the outcome variable is opaque for more elaborate machine learning models.

There seems to be an implicit interpretation of a a as a counterfactual and therefore causal quantity. Namely AR is the causal effect of an individual's environment and genome on their organ function.

Body

HR highlights two distinct (classes of) interpretations of the term "epigenetic clock". The first, t_hat is simply a predictive model which uses MS to predict age t. The second is as a latent causal variable: a "collection of innate biological mechanisms" z who is activity causes the age-dependent MS underpinning the predictive success of t_hat. As such one can think of Z as a low dimensional latent cause of observed MS variation (think PCA). In addition, it seems reasonable to say that this Z or X may or may not intersect with variables V which influence age-dependent organ function (Table 1 of HR).

HR biological age is "an ambiguous concept held to be dependent on the biological state of the individual".

Because accelerated ageing is defined as a residual, we prefer the more neutral term residual. Because residuals depend crucially on the choice of predictors (and on .....). By choosing worse predictors we artificially generate greater heterogeneity between individuals in (heterogeneity).

Section

Predictive models and modelling have an important place in biology, and in science more broadly. However it has become popular in some biological fields to categorically misinterpret predictive models. In particular, prestigious journals have published substantive research which are at best founded on an egregious statistical fallacy, and at worse meaningless or "not even wrong". We discuss one such error, based on the so-called delta-age variable, which purports to use predictive modelling to identify the arithmetic difference between two distinct definitions of "age": a biological age and chronological age. We will first give a very high level, superficial description of the aims and methods of supervised predictive modelling, before illustrating how they may be subject to misinterpretation.

The objective of predictive models in supervised machine learning and statistics is to best predict out-of-sample variation in a designated outcome variable y. It is not to predict hidden variables, even if these hidden variables are caused by, or correlated with, the outcome variable. Of course, a successful predictive model of designated outcome y, will also be successful at predicting variables z which are correlated with y. By mathematical necessity however, prediction of these z will be worse not better than the predictions for the train to variable y. In particular predictions of z will contain more random, i.e. non-systematic, prediction error (owing to z being only statistically and not deterministically related to the variable y for which the original model estimation was optimized).

So much for the objective of predictive modelling, we turn now to the methods of predictive modelling, which often feature a quantity called the prediction error. The prediction error, or generalization error, quantifies are failure to predict. For continuous valued y, the prediction error is typically (a function of) the arithmetic difference yhat − y between the predicted and the observed outcome. This predictive failure arises either because of a discrepancy between the in sample and out of sample data generating process, or due to a misspecification of the predictive model. (Such misspecification commonly arises because the predictive model is too complicated or too simple relative to the amount of data and/or the true data generating process.) Famously, prediction error is higher for models that are too simple or too complex and is therefore used as a model selection or model evaluation criteria (as an Occam's razor to guard against overfitting). The goal in prediction is not to minimize in sample prediction errors, but the maximize the so called expected empirical risk, which is typically some form of long term average prediction error over repeated draws from the data generating process.

We have just offered a informal but accurate description of the goals and methods of standard predictive modelling. We will now articulate a highly non-standard interpretation of predictive modelling that has apparently become popular in some biological fields. In such fields, not only do we assume that yhat estimates a hidden variable z (correlated to y), but we assume that yhat is closer to z than it is to y. No formal justification is given for the former, and the latter contradicts the forgoing statistical reality that yhat must predict any correlate z worse not better than it predicts y. Furthermore, the reality that prediction error is in part determined by an artifactual modeling choice – the choice of an overly complex or simple model – is entirely absent from this interpretation. Nonetheless, under these (false) assumptions z-y = yhat -y, which in turn can be (falsely) interpreted as a biologically meaningful quantity which characterizes each individual sample. In the case of delta-age, this (unwarranted) interpretation of yhat-y purports to give it a scientific interpretation as "accelerated ageing". In summary, having trained a model to predict chronological age from transcript data, it is falsely presumed that the estimated model magically offers a superior characterisation and prediction of "biological age" z than does the chronological age y.

On residuals

The residuals are nothing but the difference between actual and predicted values. So the values of the residuals are a function of how good or bad the prediction method is. You probably asked your question in the context of linear regression. The linear regression algorithm adjusts the slopes (i.e. the weights of the explanatory variables) and the intercept in a way so the sum of the residuals becomes zero and their correlation with the explanatory variables also becomes zero. This behavior is an outcome of the way linear regression algorithm works and follows from the theory (as described in the above link). Basically, the algorithm tilts and moves the line until the residuals behave they way as they do.

It is completely conceivable that a regression method optimizes using some algorithm that does not guarantee that sum(residuals) and correlation of the residuals with explanatory variables would be zero. In fact, if we force the intercept in linear regression to be zero (force the line to pass through the origin), all other things being the same, neither the sum of residuals nor their correlation with the explanatory variables is likely to be zero. This happens because with zero intercept, the line can be tilted but not moved. So, it entirely depends on the regression algorithm how the residuals behave. If a predictor always returns a constant irrespective of what the explanatory variable values are, sum(residuals) will naturally not be zero. And, if there are two explanatory variables x1 and x2 and the predictor (perhaps erroneously) assigns a very high weight to x1 and a very low weight to x2, the correlation between x2 and the residuals may be nearly same as the correlation between x1 and x2 (because the predicted value will mostly follow x1). So, in summary, sum(residuals) is zero because learning algorithms try to minimize the error after all. And, correlation between residuals and explanatory variables is zero as a result of —at least in the case of linear regression— how the algorithm works.

HR "minimizes the error associated with estimating chronological age".

If the weighted-average of CpGs is dimension reduction, which simply summarizes the status of methylation.

This is unproblematic.

It would not be the first time that entire Fields have collapsed (VUL).

TABLE 1: Hyperbole

It is not conventional to name a prediction after it predictors. When predicting temperature from a thermometer we don't refer to it as thermometer temperature, this is usually just emphasize that it is an in accurate measure of temperature. Similarly for epigenetic age. Thermometer temperature has only a methodological and not a scientific interpretation, as does the PE. "Acceleration" has a physical meeting as the second derivative not a PE (which is a zeroth order quantity, note even a rate).

HR = Horvath, S., & Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics, 19(6), 371.

VUL = Vul E, Harris C, Winkielman P, Pashler H. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspectives on psychological science.

HAN = Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., ... & Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell, 49(2), 359-367.