Cross-validating models of continuous data from simulation and experiment by using linear regression and artificial neural networks

Zohreh Zakeri[1, *, +], Neil Mansfield[1], Caroline Sunderland[2], Ahmet Omurtag[1, +]

[1]Department of Engineering, School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, NG11 8NS, UK

[2]Department of Sport Science, School of Science and Technology, Nottingham Trent University, Clifton Lane, Nottingham, NG11 8NS, UK

*Correspondence: zohreh.zakeri@ntu.ac.uk

[+]These authors contributed equally to this work

Abstract

We are increasingly surrounded by sensors gathering massive amounts of data, and patterns in continuous variables are often discovered by using artificial neural networks (ANN), while linear regression (LR) is useful for detecting linear relationships. LR also provide preliminary estimates of potentially complex associations, and serve as a benchmark for the performance of ANNs. We show that while cross-validation (CV) is indispensable for insuring the robustness of the discovered patterns, it systematically leads, when combined with LR, to specific artefacts that underestimate the extent of the associations between predictor and target variables. We explain how this previously unnoticed type of artefact arises specifically from the combination of CV with LR and does not affect non-linear methods such as ANN. We also demonstrate through simulations that ANN were able to discover a wide range of complex associations missed by LR. The results were confirmed by the analysis of physiological, behavioural and subjective data collected from N=31 human subjects performing laparoscopy training experiments.

Keywords

Cross-validation, linear regression, artificial neural networks.

Introduction

Linear regression (LR) can be useful not only for discovering patterns in experimental data, but also as a baseline for benchmarking and validating new analysis techniques, especially novel or unfamiliar ones. For example an artificial neural-network (ANN) may reveal multidimensional, nonlinear relationships that are inaccessible by linear regression (LR). However it may have features that need further study and validation by comparing to an alternative method. As a well-established and relatively transparent method, LR is particularly suited for this role (e.g. [1,2]).

Cross-validation (CV) is commonly employed for developing and assessing the accuracy of ANNs [3]. It involves reserving part of the available data for training the ANN, and the rest for testing its accuracy. Given a fixed set of data, repeated CV with different training/test groups can minimize overfitting. In benchmarking performance, it is desirable to assess all the analysis techniques on an equal footing. This may involve using the same data set, as well as the same way of measuring accuracy. Thus it may appear sensible to use CV to measure the accuracy of both the ANN and LR. However this is not necessarily the case, as we show below.

The above scenario is not the only reason for wanting to combine LR with CV. CV is a good way of assessing the predictive power of data models. The relationships discovered by LR applied to the entire data set may be driven by a small number of outliers and the LR may be overfitted; that is, latch on to the patterns that exist only in the current data. Repeated LR using different subsets of data may be a better indicator of how well the results will hold up when new data arrive. In fact hold out methods, where the data set is split into training and testing sets, are recommended for LR in the literature [4–8].

However CV in combination with LR has an inevitable and underappreciated defect associated with finite samples: excluding part of the data generates spurious negative correlations between the actual and predicted targets. This effect is stronger smaller the data set. In this paper we illustrate how this bias comes about, use experimental and simulated data to explore its effects under various types of predictor-target dependence, and recommend alternative metrics for quantifying accuracy. The experimental data came from our previous study on measuring the cognitive load experienced by novice trainees during laparoscopic surgery training [2].

Method

Given a set of continuous valued data $\{x_n , y_n\}, n = 1, \dots, N_D$, various techniques may be employed to discover relationships among the predictors x_n and the targets y_n. An important goal is to be able to predict target values if predictor values become available that are not in the current data set. It is therefore necessary to find relationships that are generalizable. For this purpose, cross-validation (CV) is the method of choice. K-fold CV consists of randomly partitioning data into disjoint groups, holding out one of these as the testing partition, developing the prediction model using all of the remaining partitions (training set), having the model predict the targets in the testing set, then comparing the predictions with the actual values. This procedure is repeated k times and the resulting accuracy is averaged. Since partitioning in a large data set may be done in numerous different ways, the k-fold CV may then be repeated a number of times for better statistical validity. When $k = N_D$, this approach is referred to as leave-one-out cross-validation (LOOCV) and each partitioned test data set has a size of $N_D = 1$.

With continuous valued targets, a good way to determine the accuracy of prediction is to calculate the Pearson correlation r between the predicted, $\bar{y}$, and actual target values. In this paper we used linear regression (LR) and artificial neural networks (ANN) for prediction. The cascade-forward ANN was used with two hidden layers and the Levenberg-Marquard backpropagation algorithm. In what follows, we use the acronym LR-FULL to denote LR applied to the full data set; LR-CV for cross-validated linear regression; and ANN-CV for cross-validated ANN.

We conducted simulations where targets were calculated from predictors according to the formula $y^{(i)} = \sum_{n=1}^{Np} \left[a x_n^{(i)} + b(x_n^{(i)})^2 \ \ + c\, f(x_n^{(i)})\right)] + v\,\phi^i$, with $i = 1, \dots, N_D$. The predictors and noise (ϕ^i) were chosen as random deviates from the standard normal distribution or a uniform distribution over the unit interval. Size of the data set was N_D, and the amplitude of the noise was controlled by v. The other coefficients were chosen to generate different types of dependence of the target on the predictors: independence, $a = b = c = 0$; linear dependence, $a = 1$, $b = c = 0$; quadratic dependence, $a = c = 0$, $b = 1$; and an additional nonlinear dependence, $a = b = 0$, $c = 1$), given by $f(x) = \cos(6x^3)$. Pearson correlation between predictor and targets was denoted r_0 . The noise amplitude $v = \sqrt{r_0^{-2} - 1}$ was used to create linear dependence with $r_0 > 0$.

For multivariate x_n , the coefficient of determination tends to increase with the number of predictors, regardless of the actual dependence between x_n and y_n. This is a well-known artefact of multiple regression. Consequently it is advisable to use, R^2_{adj} , the adjusted R-squared which compensates for this increase, instead of the coefficient of determination [9]. For multivariate predictors, we used R_{adj} , the square root of the absolute value of adjusted R-squared (which can become negative). This quantity, which we refer to as adjusted R, provided a value that was commensurate with r .

To confirm our results in an experimental setting, we used the data collected from 31 participants in a previous study that investigated the ability of physiological variables to track the cognitive load associated with laparoscopic surgery (LS) training [2]. The study measured the responses described by the heart rate and prefrontal cerebral oxygenation, derived from concurrent near-infrared spectroscopy (fNIRS) recorded from novice subjects. As they performed tasks on a LS trainer box, the subjects in these experiments responded to randomly timed auditory stimuli by pressing a pedal and their reaction times were recorded. The non-response rate was calculated as the rate at which a subject failed to respond before the next stimulus onset. They reported their experience by filling out a NASA-TLX questionnaire after each session. Fingertip blood samples were taken at baseline and immediately after completion of all three LS tasks to determine the serum cortisol and brain-derived neurotropic factor (BDNF) concentrations. The study postulated that the physiological metrics (predictors) would track the overall effort as indexed by the traditional secondary task and subjective methods (targets).

Three tasks were performed by each participant (the final one being a repetition of the first). The data from the 3 tasks were pooled, so that every distinct predictor-target analysis contained N=93 data points. To summarise the predictor variables described subjects' physiology: plasma cortisol concentration (ng/ml), brain derived neurotropic factor (BDNF) concentration (pg/ml), heart rate (HR) (min^{-1}), and left prefrontal oxygenated haemoglobin concentration changes (mM) denoted HbO. The target variables described the subjects' behaviour or subjective status: task completion time (CT) (min), error rate (ER) (min^{-1}), reaction time (RT) (s), Non-response rate (NRR) (s^{-1}), and the average NASA-TLX score.

Results

In this section we first present the explanation of the underlying mechanisms that create spurious correlations when CV and linear regression are combined (Figure 1). We show simulations with various types of predictor-target dependence in Figure 2. Next, we explore multivariate linear regression (Figure 3) and, finally, apply the concepts discussed to experimental data (Table *1*).

Although the correlation r between the actual and predicted targets is generally a good indicator of the strength of the discovered patterns, it becomes unreliable when LR and CV are used in combination. Our first task was to elucidate the underlying reasons for this problem in a small data set consisting of only $N_D = 10$ predictor-targets pairs. The linear dependence of the targets on the predictors with added noise was generated in accordance with $y = x + vQ$, a special case of the general formalism described in the Methods section. This equation follows straightforwardly from a general noisy linear dependence by rescaling and shifting the predictor and target variables.

We chose to work with it since it has the advantage of having only one parameter, v , that determines the strength of the noise relative to the predictor's variability. The correlation r_0 between the predictor and target could be adjusted by changing v.

Figure 1a shows a data set generated from uncorrelated targets and predictors. In this case the regression (thick solid black) line showed a negative correlation due to small sample effect. In order to investigate the effects of cross validation, we held out individual (test) points and replotted the resulting regression line obtained from the remaining (training) points. The figure shows that a new, slightly shifted, regression line (red, dotted) was obtained by holding out the point labelled 1. Another shift (blue, dashed) in the regression line occurred when point 2 was held out. Similar shifts occurred when any other part of the data were held out. It was the general pattern of these shifts that collectively distorted the value of r, as we explain next.

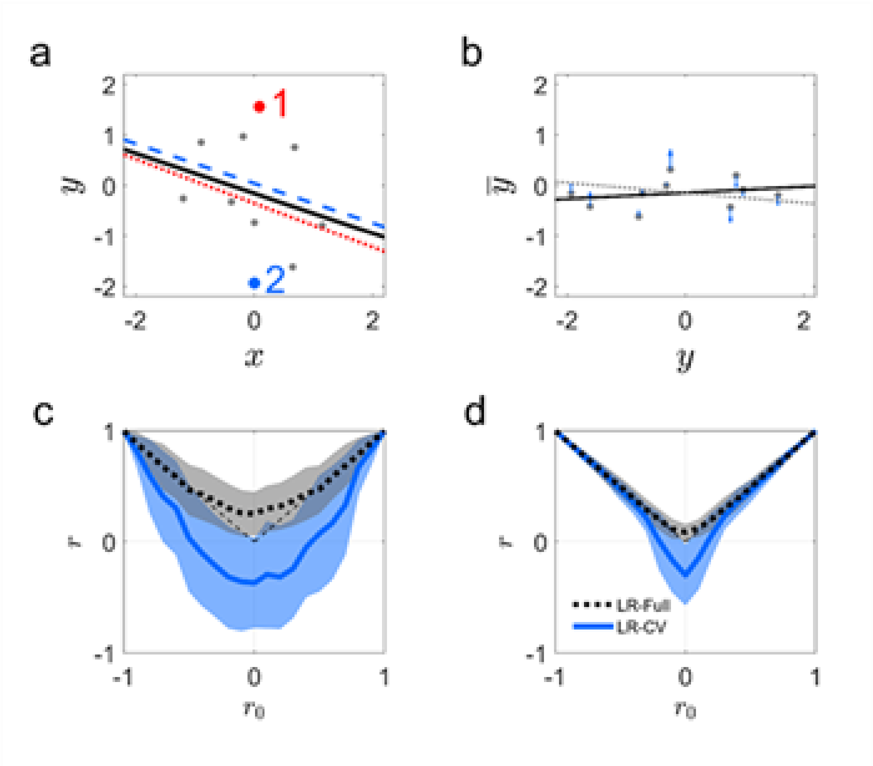

Figure 1. The effects of cross-validation on linear regression with simulated predictor-targets. (a) Linear regression with $N_D = 10$ data points. The solid line shows the regression to the full data set. The dotted (red) line shows the regression line when the point labelled 1 is held out. The dashed (blue) line shows the regression line when the point 2 is held out. (b) Each point represents a target y (from subplot (a)) and the corresponding value $\bar{y}$ predicted by regression on the full set of data. The arrows represent how $\bar{y}$ changes when prediction is done by leave-one-out cross-validation. $N_D = 10$. (c) The correlation r between predicted and actual targets versus the correlation r_0 between linearly dependent predictor and targets. The thick solid blue curve is for cross-validated linear regression. The thick dotted black curve is for linear regression on the full data set. The shaded region around these curves represent the standard deviation of the variability obtained by 100 repetitions of the simulation. The dashed thin black line shows the locus of $r = \pm r_0$. $N_D = 10$. (d) Same as in (c) but with a larger data set, $N_D = 100$.

Note that when LR was used on the full data set, the predicted value, $\bar{y}_{FULL}$, of a target was given by a point on the thick black regression line that corresponds to x. But when LR was combined with CV, the predicted target value $\bar{y}_{CV}$ of the same target was given by a point on a different regression line. Thus in going from the full to the training data set, the prediction was changed by an amount $\Delta\bar{y} = \bar{y}_{CV} - \bar{y}_{FULL}$. The key issue of interest was that this shift was not independent of the target's value; in fact $\Delta\bar{y}$ was negatively correlated with y.This fact is illustrated in Figure 1b which shows each $\bar{y}_{FULL}$ against the corresponding y as a point. How the prediction changed with CV (i.e. $\Delta\bar{y}$) is illustrated by an arrow. This change occurred because removing a low target value slightly raised the mean of the remaining set of targets, causing an upward shift in the regression line.

And removing a high target value caused an opposite shift. Thus lower values of the target had their predictions raised while higher values had them lowered. This inevitably led to a downward bias in r, the correlation between actual and predicted targets. This downward bias is shown by the difference between the dotted line in Figure 1b that was fitted to the pairs y & $\bar{y}_{CV}$, as compared with the solid line that was fitted to the pairs y & $\bar{y}_{FULL}$.

What was described so far was based on uncorrelated predictors and targets. We next investigated how the downward bias was affected by various degrees of the true correlation r_0 between the predictors and targets. Figure 1c shows that regression on the full data set yielded values of r that closely reflected r_0, when the magnitude of r_0 was sufficiently high (except for a small positive bias when $r_0 \approx 0$). This was consistent with the acknowledged utility of r as a metric that indicates the effectiveness of the prediction method. The most salient feature in Figure 1c is, however, the negative bias in r that resulted from cross-validated regression, as shown by the thick solid blue curve. The figure shows that this negative bias increased with decreasing magnitude of r_0. Since Figure 1a-c were for a small data sets ($N_D = 10$) only, we performed simulations with larger data sets and found that the biases in (c) were reduced with increasing N_D. Figure 1d shows an example of this reduction with $N_D = 10$.

The results described so far were for linear dependence of targets on predictors. In Figure 2 we show how the mechanism explained above created similar results under various other types of dependence. Each column of the figure is for a different type of dependence, with the leftmost column showing independent predictor-targets. The top row in the figure shows the scatter plots of the predictor and targets. In each subplot, the Pearson correlation is shown in the top left. When the targets were independent (Figure 2a), the regression on the full data set showed no correlation between the actual and predicted targets (LR-FULL, Figure 2b).

When the same data points were used in cross-validated linear regression, a strongly negative correlation was found (LR-CV, Figure 2c)

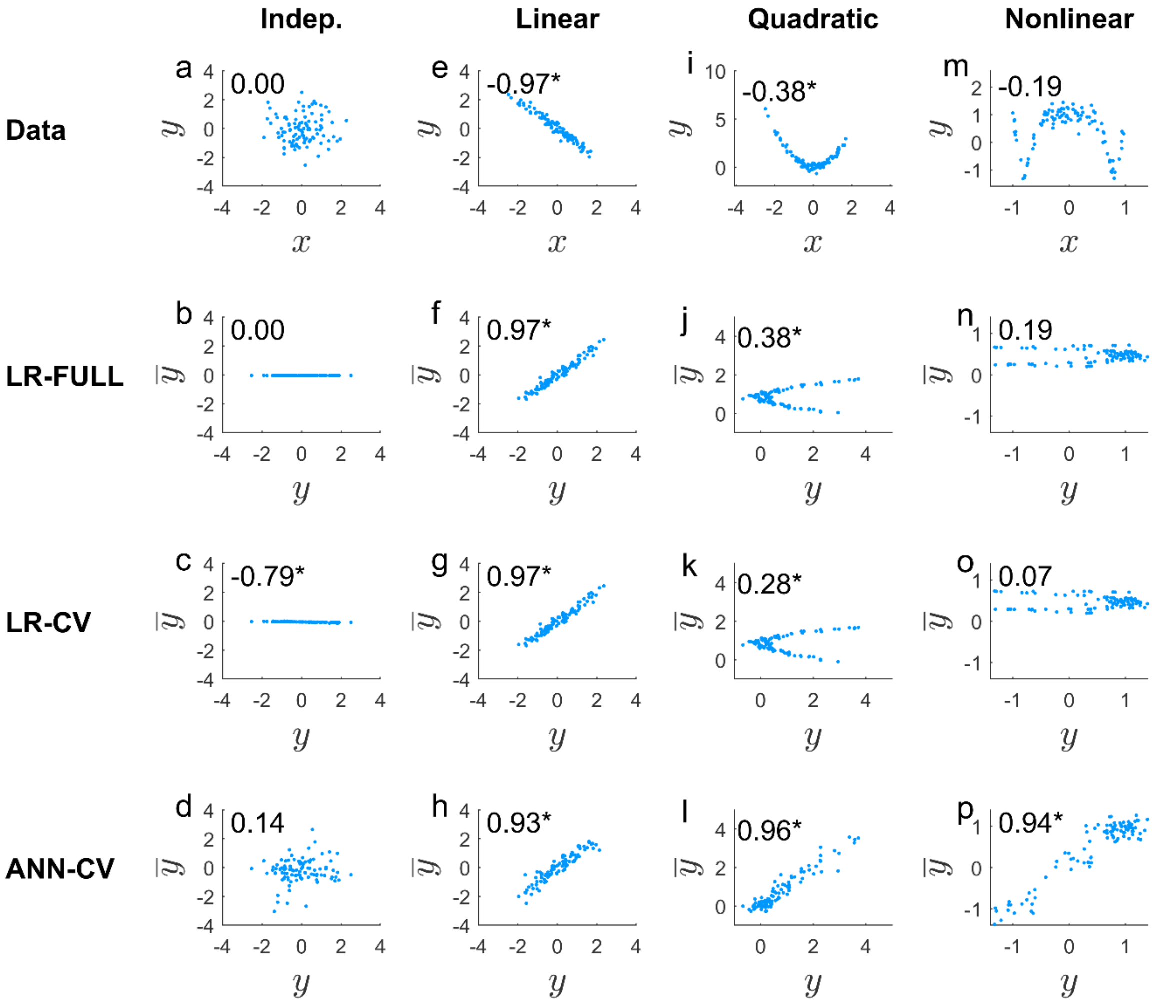

Figure 2. Simulated univariate data with various types of dependence. Columns correspond to no dependence **(a-d)**, linear **(e-h)**, quadratic **(i-l)**, and other nonlinear **(m-p)** dependence. The top row shows scatter plots of the predictor and targets. The second row shows the actual targets v predicted values determined using linear regression on the full data set (LR-FULL). The third row is for cross-validated linear regression (LR-CV). The bottom row is for the cross-validated artificial neural network (ANN-CV). Cross-validation was repeated 100 times with different partitioning. $\nu = 0.23$, $N_D = 100$.

This finding might have been puzzling were it not for the explanation offered in Figure 1. Figure 2d confirmed that ANN was free of this artefact. ANN found a small, positive correlation that was spurious, possibly a finite-sample effect.

The second column of Figure 2 (linear dependence) showed that the negative bias of LR-CV was minimised when the target was strongly correlated with the predictor (as expected from

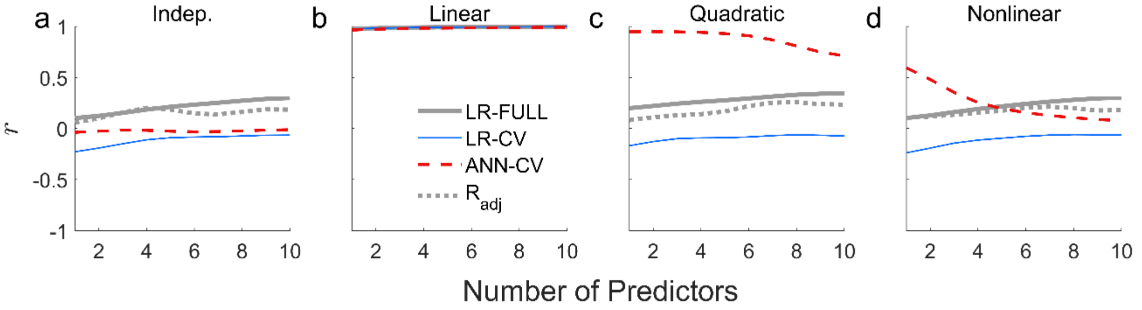

Figure 3. Prediction metrics for simulated multivariate data as a function of the number of predictors. **(a)** No dependence between predictor and target; **(b)** linear dependence; **(c)** quadratic dependence; **(d)** **o**ther nonlinear dependence. Thick gray curve (LR-FULL) is the Pearson correlation between actual targets and their values predicted by regression on the full data set. The dotted gray curve is the adjusted R. The thin (blue) and thick (red) curves are, respectively, for the cross-validated linear regression and ANN. $\nu = 0.23$, $N_D = 100$.

Figure 1c). In the case of quadratic and other nonlinear dependences (third and rightmost columns of Figure 2), LR-CV compared to LR-FULL showed a small negative bias. In those cases, LR was not able to discover the targets' relationship to the predictor, while ANN-CV did so with a nearly perfect correlation between the actual and predicted targets.

Next we investigated how the effects explained in Figure 1 play out in the case of multivariate prediction, with multiple types of simulated dependence. Figure 3a shows that, for independent predictor-targets, LR-CV (thin blue curve) had a negative bias that tended to diminish with increasing number of predictors N_P while ANN-CV (dashed red) yielded $r_0 \approx 0$ for any N_P. LR-FULL (thick gray) increased with . Although the figure shows up to 10 predictors, we verified that $r \longrightarrow 1$ as $N_P \longrightarrow N_D$. The adjusted R (dashed gray) did not increase, but was not as close to zero as was ANN-CV. Figure 3b shows that all of the methods under discussion successfully discovered the linear relationship. In the case of quadratic dependence (Figure 3c), LR-CV showed a strong negative artefact, LR-FULL displayed a positive bias that increased with N_P, and adjusted R remained close to zero indicating the inability of LR to find the quadratic dependence. ANN-CV was able to indicate the dependence, although it found it increasingly difficult to do so with increasing N_P. A similar situation obtained with the other nonlinear dependence (Figure 3d), although the ANN was even less successful with increasing N_P.

Having examined the adverse effects of combining CV with LR through simulations, we next turned to the experimental data set explained in the Methods section. Table 1 shows that for the pair Cortisol-Error rate, LR-FULL was 0.01 suggesting that the two variables were independent. This may have led to the expectation that cross-validated LR-CV would also be nearly zero. However LR-CV was -0.80. Numerous other similar cases were exemplified in the table, for example for the pair HbO and Reaction time, and the pair Heart Rate and Completion Time. In fact, LR-CV was smaller than LR-FULL for every predictor-target pair. Such large differences were not observed in the case of the ANN-CV. The mean difference between LR-CV and LR-FULL was -0.28, while the mean difference between ANN-CV and LR-FULL was -0.08.

Table 1. Metrics that quantify the prediction of experimental data. For each predictor-target pair, the table shows the Pearson correlation (r_0) between the predictor and target, the Pearson correlation between the actual and predicted target for linear regression on the full data set (LR-FUL), cross-validated linear regression (LR-CV), and cross-validated neural network (ANN-CV). There were $N_D = 93$ data points in each set.

		TARGET				
		Completion time	**Error rate**	**Reaction time**	**Non-response rate**	**NASA-TLX avg.**
Cortisol	r_0	0.04	0.01	0.19	0.13	0.13
	LR-FULL	0.04	0.01	0.19	0.13	0.13
	LR-CV	-0.39	-0.80	0.09	0.01	-0.04
	ANN-CV	0.00	-0.21	0.06	-0.04	-0.04
BDNF	r_0	0.26	0.05	-0.01	0.13	0.29
	LR-FULL	0.26	0.05	0.01	0.13	0.29
	LR-CV	0.19	-0.36	-0.81	0.03	0.23
	ANN-CV	0.14	-0.20	-0.02	0.02	0.34
Heart Rate	r_0	0.09	-0.19	-0.29	-0.29	0.18
	LR-FULL	0.09	0.19	0.29	0.29	0.18
	LR-CV	-0.26	0.07	0.24	0.18	0.04
	ANN-CV	0.04	-0.05	0.27	0.36	0.09
HbO	r_0	0.05	-0.16	-0.02	0.01	0.24
	LR-FULL	0.05	0.16	0.02	0.01	0.24
	LR-CV	-0.28	0.06	-0.59	-0.53	0.16
	ANN-CV	0.09	0.06	-0.04	0.18	0.18

Discussion and Conclusions

The correlation r between the actual and predicted values, is often used as a measure of the quality of the prediction. In this paper we showed that r is negatively biased when the prediction method is linear regression with cross-validation. The bias was greater for smaller data sets and when the true correlation between the predictor and target was low. This bias was caused by the fact that removing part of the data (test set), as part of the process of CV, caused the statistics of the remaining data (training set) to change, and this resulted in changes to the predicted values which are anti-correlated with the actual values of the target. Both linear regression and cross-validation are widely used techniques, and their combination appears to be recommended in some textbooks (e.g. [4,7]) as a way of mitigating overfitting. We have shown, for the first time in peer-reviewed literature to our knowledge, that this combination can create serious problems in the analysis of experimental data (e.g. Table 1). We also explained the underlying reasons (Figure 1) and demonstrated its effects through simulations (Figure 2 and Figure 3).

The most obvious way to reduce the bias is to use larger data sets (Figure 1d); however in practice this is often not possible, as data may be difficult or expensive to obtain. Another well-known approach in cross-validation is to use stratification during partitioning. This way the membership of every partition is chosen, to the extent possible, to represent the population. For continuous data, however, stratification is not well-defined (Witten et al., 2016). We implemented stratification (after arbitrarily binning the targets) but this did not eliminate the problem. In a further attempt to mitigate the bias, we devised a modified CV procedure that involved choosing members of a partition in pairs, so that the mean of each pair was as close as possible to the population mean. For example the points 1 and 2 in Figure 1a could form such a pair. This paired-CV procedure reduced but did not eliminate the problem, hence was not further pursued. Nonetheless the reduction of the bias through paired-CV helped further confirm of the explanation given in Figure 1.

Our results suggest that cross-validation should be avoided when prediction is done by linear regression. We recommend, instead, that regression be applied to the full data, after implementing alternative ways of avoiding overfitting (such as removing outliers). Figure 1c-d indicate that regression on the full data set generates a bias when r_0 is small, in this case a small positive value of r. In order to understand the origin of this bias, note that although the predictor and targets may be uncorrelated ($r_0 = 0$) a finite sample will in general contain a (usually small) positive or negative correlation between them. Regardless of whether this predictor-target correlation is positive or negative, it will result in a positive correlation between the actual and predicted targets; hence r will tend to deviate upward from zero.

For multivariate predictors, the correlation between actual and predicted targets in the full data also cannot be used, because this quantity spuriously increases as the number of the type of predictors increases. In this case, the adjusted-R is the preferable metric for quantifying the outcome of linear regression (Figure 3). Our results indicate that, provided these precautions are heeded, linear regression remains an effective way of discovering patterns in data as well as of providing a performance benchmark for more sophisticated analysis techniques.

Declaration of competing interest

None declared.

Acknowledgments

The authors would like to thank Bethany Twigge for her invaluable help in performing the experiment. This research was funded by the Higher Education Funding Council for England (HEFCE) at Nottingham Trent University, the School of Science and Technology.

References

[1] Friedman N, Fekete T, Gal K, Shriki O. EEG-Based Prediction of Cognitive Load in Intelligence Tests. Front Hum Neurosci 2019;13:191. https://doi.org/10.3389/fnhum.2019.00191.

[2] Zakeri, Z., Mansfield, N., Sunderland, C., and Omurtag A. Physiological correlates of cognitive load in laparoscopic surgery. n.d.

[3] Witten IH, Frank E, Hall MA, Pal CJ. Data Mining: Practical Machine Learning Tools and Techniques. Elsevier Inc.; 2016. https://doi.org/10.1016/c2009-0-19715-5.

[4] Kleinbaum, D.G., Kupper, L.L., Muller, K.E., and Nizam A. Applied regression analysis and other multivariable methods. Duxbury Press Belmont, CA; 1988.

[5] Mark J, Goldberg MA. Multiple Regression Analysis and Mass Assessment: A Review of the Issues. Appraisal J 1988;56:89.

[6] Martinez, W.L., and Martinez AR (2015). C. Computational statistics handbook with MATLAB. Chapman and Hall/CRC; 2015.

[7] Myers, J.L., Well, A.D., and Lorch Jr RF. Research design and statistical analysis. Routledge; 2013.

[8] Tatsuoka MM, Lohnes PR. Multivariate analysis: Techniques for educational and psychological research. Macmillan Publishing Co, Inc; 1988.

[9] Chatterjee, S., and Hadi AS. Regression analysis by example. John Wiley & Sons; 2015.