Proceedings of Machine Learning Research 58 (2016) 13-25

Simple Regression Models

Jan Malte Lichtenberg LICHTENBERG@MPIB-BERLIN.MPG.DE
Özgür Şimşek OZGUR@MPIB-BERLIN.MPG.DE
Center for Adaptive Behavior and Cognition
Max Planck Institute for Human Development
Lentzeallee 94, 14195 Berlin, Germany

Editor: Tatiana V. Guy, Miroslav Kárný, David Rios-Insua, David H. Wolpert

Abstract

Developing theories of when and why simple predictive models perform well is a key step in understanding decisions of cognitively bounded humans and intelligent machines. We are interested in how well simple models predict in regression. We list and review existing simple regression models and define new ones. We identify the lack of a large-scale empirical comparison of these models with state-of-the-art regression models in a predictive regression context. We report the results of such an empirical analysis on 60 real-world data sets. Simple regression models such as equal-weights regression routinely outperformed state-of-the-art regression models, especially on small training-set sizes. There was no simple model that predicted well in all data sets, but in nearly all data sets, there was at least one simple model that predicted well.

Keywords: Regression, Simple Heuristics, Improper Models

1. Introduction

The study of simple predictive models is an important topic in decision making and machine learning. High predictive accuracy seems to be the main focus of most current research. Yet low time complexity, robustness to small sample sizes, and interpretability are also desirable properties of a useful model. In this article, we study simple models which are much faster to estimate and easier to understand than their current state-of-the-art peer algorithms. Their advantage in computation time and interpretability will be obvious. We are mainly interested in how much predictive power is lost when using simple models, if any.

Predictive models can be simple in many different ways. One seemingly extreme approach is to take only one predictor into account. Another approach is to take all predictors into account but to combine them in simple ways, for example, by giving them equal or random weights. Such simple models have been shown to predict remarkably well in tasks such as classification (Holte, 1993), paired comparison (Czerlinski et al., 1999; Brighton, 2006; Şimşek and Buckmann, 2015), and portfolio optimization (DeMiguel et al., 2009). Less attention has been given to simple models in a predictive regression context, that is, when the problem under consideration is to estimate the value of a continuous response variable on previously unseen data.

Simple regression models such as equal weights and random weights regression have been examined empirically and theoretically in the psychological literature. Collectively, such models were termed *improper* linear models to distinguish them from *proper* linear models whose weights are obtained by optimizing some objective function. For example, ordinary least squares are obtained by minimizing the residual sum of squares.

Wainer (1976) argued that "it don't make no nevermind" if optimal weights are replaced by equal weights, as the loss in explained variance is small if predictors are directed properly (see also Laughlin, 1978; Wainer, 1978). Similar findings, commonly known as the *flat maximum effect*, show that the space of nearly optimal weights for linear models is large (Ehrenberg, 1982; von Winterfeldt and Edwards, 1982; Lovie and Lovie, 1986). Dawes and Corrigan (1974) and Dawes (1979) concluded that optimal weighting of predictors would therefore be less important than choosing the right predictors and knowing their directional relationship with the response.

Moreover, Einhorn and Hogarth (1975) argued that improper models suffer from smaller estimation error compared to proper models (or no estimation error) because the weights of improper models do not have to be estimated from the data. Because they combine a small loss in accuracy with increased robustness due to smaller estimation error, the common message of these studies was that improper models would be superior to multiple linear regression in some situations and not greatly inferior in others when the aim is out-of-sample prediction.

These surprising results showed that improper models can match the performance of more complex models or even outperform them. However, existing work on simple regression models did not study these models in a regression context. Some of them used loss functions that are not regression adequate (Wilks, 1938; Dawes and Corrigan, 1974; Dawes, 1979; Einhorn and Hogarth, 1975; Dana and Dawes, 2004). For example, Dawes (1979) used the correlation coefficient between predicted and true response values to assess the prediction performance of different models. Other studies evaluated improper regression models in a fitting rather than in a prediction context (Wainer, 1976; Waller and Jones, 2009). Furthermore, many results were presented relative to the performance of multiple linear regression (Einhorn and Hogarth, 1975; Wainer, 1976; Graefe, 2015), which is known to have severe estimation issues under a large variety of conditions. It is unclear whether the results still hold relative to more recent proper models, such as the elastic net (Zou and Hastie, 2005) or other regularized linear models.

In this article, we examine how well simple regression models predict in a regression context when compared with state-of-the-art statistical models in a large, diverse collection of real-world data sets. In doing so, we complement and contrast findings from other domains such as classification and paired comparison, building toward a more general theory of when and why simple models perform well.

Our results show that simple regression models such as equal weights regression routinely outperformed not only multiple linear regression but also state-of-the-art regression models, especially on small training-set sizes. There was no simple model that predicted well in all data sets, but for nearly all data sets, there was at least one simple model that predicted well.

In Sections 2 and 3 we review existing simple regression models, define new ones, and describe how to estimate their parameters. In Section 4 we report the results of a large empirical study that compared simple models with state-of-the-art regression algorithms on 60 data sets using a regression-adequate loss function.

2. Simple regression models

Let us assume that we have some data $(y_i, \boldsymbol{x}_i),\ i = 1, \dots, n$, where $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$ is a p-dimensional vector of predictors and y_i a real-valued response for the ith observation. A *regression model* f is a model that makes a prediction $\hat{y}$ of y for a potentially new input vector $\boldsymbol{x}$, that is,

$$\hat{y} = f(\boldsymbol{x}).$$

The simple models we consider are special instances of the linear regression model[1]

$$\hat{y} = \beta_0 + \gamma \sum_{j=1}^{p} x_j \alpha_j,$$

and share the following properties: (a) weights α_j are chosen heuristically (for example, equal weights), and (b) weights α_j can be estimated or determined independently of the location parameter β_0 and the scale parameter γ. Intuitively, the weighted sum determines the nature of how the predictors are combined or selected. The two parameters β_0 and γ then determine the location and scale of this weighted sum, respectively. Different simple models correspond to different ways of determining α_j. Estimation of β_0 and γ is the same for all considered simple models and is explained in the following section.

We assume that predictors and responses are *centered*, that is,

$$\bar{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i = 0 \quad \text{and} \quad \bar{\boldsymbol{x}}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} = 0 \text{ for all } j = 1, \dots, p,$$

and *scaled*, that is,

$$s_{\boldsymbol{y}} = \frac{1}{n} \sum_{i=1}^{n} y_i^2 = 1 \quad \text{and} \quad s_{\boldsymbol{x}_j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2 = 1 \text{ for all } j = 1, \dots, p.$$

A centered and scaled variable is called *standardized.* Scaling the response is not necessary for the simple models to function well but simplifies the analysis across different data sets. Furthermore, predictors are said to be *directed* if they correlate non-negatively with the response. We now define each of the considered simple models. Table 1 points to existing literature for each model.

Mean prediction. This is the simplest available model and always predicts the mean value of the response calculated on the training data. The corresponding model can be written as

$$\hat{y} = \beta_0.$$

Mean prediction is appropriate when no predictor is available. Typically, data sets that do not contain predictive predictors are not considered. Therefore, mean prediction does not play a role in supervised learning in general. Yet the model can still serve as a baseline.

Random weights. This is perhaps the most improper model one could imagine. Once standardized and directed, each predictor is assigned a random weight stemming from a

[1] Setting $\beta_j = \gamma\alpha_j$, we obtain the classic formulation of linear regression: $\hat{y} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$.

Model	**Literature**
Random weights	Wilks (1938); Dawes and Corrigan (1974)
Equal weights	Wesman and Bennett (1959); Schmidt (1971); Dawes and Corrigan (1974); Einhorn and Hogarth (1975); Wainer (1976); Dawes (1979); Dana and Dawes (2004); Davis-Stober et al. (2010); Graefe (2015)
Correlation weights	Dana and Dawes (2004); Waller and Jones (2009)
Single-cue regression	Dana and Dawes (2004); Davis-Stober et al. (2010)
Correlation ranks	Wesman and Bennett (1959)

Table 1: Literature on simple regression models.

uniform distribution, that is,

$$\hat{y} = \beta_0 + \gamma \sum_{j=1}^{p} \omega_j x_j,$$

where $\omega_j \sim \mathcal{U}(a, b)$. Different authors used different values for a and b. We used $a = 0$ and $b = 1$. Nearly 80 years ago, Wilks (1938) showed that the correlation of predictions of two independent random-weights models tends to 1 with an increasing number of positively intercorrelated variables. Random weights should be outperformed by equal weights because of the smaller variance of the latter (Dawes, 1979). We include random weights as a lower benchmark model in our empirical analysis.

Equal weights. This model takes all standardized predictors into account and weights them equally, that is,

$$\hat{y} = \beta_0 + \gamma \sum_{j=1}^{p} x_j. \tag{1}$$

Under the assumption that all predictors are directed, equal weights has only two free parameters, location and scale.

Equal-weighting has been discussed in a large variety of settings resulting in slightly different models: If $\beta_0 = 0$ and $\gamma = 1$ in Equation (1), the resulting model is called *unit weights* (Einhorn and Hogarth, 1975). In a paired comparison context, where equal-weights models have been called *tallying* or *Dawes's rule*, these models have been shown to outperform more complex models, especially on small sample sizes (Gigerenzer et al., 1999; Şimşek and Buckmann, 2015). An equal-weights model has been found to compete well with state-of-the-art portfolio theory models in a portfolio allocation problem, where it is called the *1/*N *rule* (DeMiguel et al., 2009).

Correlation weights. This model weights all predictors by their correlation with the response, that is,

$$\hat{y} = \beta_0 + \gamma \sum_{j=1}^{p} r_{yx_j} x_j,$$

where r_{yx_j} is the sample correlation coefficient between the response and predictor $\boldsymbol{x}_j$. Correlation weights has to estimate $p + 2$ parameters. However, these coefficients are easier to calculate than ordinary least squares (OLS) weights in terms of both computational complexity and numerical stability issues. Whereas the OLS model suffers, for example, from the multicollinearity problem, the correlation coefficients are calculated independently of each other and independently of β_0 and γ. The correlation-weights model thus scales favorably with the number of predictors when compared to the OLS model and its matrix inversions.

Single-cue regression. This model considers only the predictor that has the highest correlation with the response among all available predictors. The corresponding model can be written as

$$\hat{y} = \beta_0 + \gamma x_1,$$

where x_1 is the cue that is most correlated with the criterion y. To determine *the* single cue, the correlations between all predictors and the response are estimated and the one with the highest absolute value is chosen. Estimation of single-cue regression is not less complex than estimation of correlation weights as it is necessary to calculate all p predictor–response correlations in order to determine the single cue. Yet there may be simpler ways to (approximately) determine the single cue, and single-cue regression is simpler at decision time, where it requires only the information of one predictor.

Correlation ranks. This model does not need the exact values of the correlation weights but only their ranks, that is, their relative order. The corresponding model can be written as

$$\hat{y} = \beta_0 + \gamma \sum_{j=1}^{p} \rho_j \ x_j, \quad \text{where} \quad \rho_j = rank(r_{yx_j}).$$

The lowest correlated cue has rank 1 and the highest correlated cue has rank p. Ties are assigned the average rank.[2] Correlation ranks might be easier to estimate and thus more robust than correlation or OLS weights but still allow for differential weighting of multiple predictors, as opposed to equal-weighting or single-cue strategies. Our implementation of correlation ranks actually first estimates all correlations and then assigns ranks. However, there may be simpler ways to (approximately) determine the ranking of correlations. Models similar to correlation ranks have been compared to true-weights and equal-weights models in the context of *multiattribute decision making*[3] in Barron and Barrett (1996). We know of no study that compares the prediction accuracy of correlation-ranks models to other regression models in a regression context.

3. Parameter estimation from training data

Unless specifically stated otherwise, we assume that β_0 and γ are calculated using simple linear regression (SLR). Estimation of the weights α_j depends on the respective algorithm and is done before the estimation of β_0 and γ.

SLR is much easier to estimate than OLS regression in general as it involves no inversion of matrices but only simple estimates of scale and covariation. Defining $c(\boldsymbol{x}_i) = \sum_{j=1}^{p} x_{ij}\alpha_j$ and $\boldsymbol{c} = (c(\boldsymbol{x}_1), \ldots, c(\boldsymbol{x}_n))^T$, the SLR estimates for model (2) are given by

$$\gamma = r_{\boldsymbol{yc}} \frac{s_{\boldsymbol{y}}}{s_{\boldsymbol{c}}} \quad \text{and} \quad \beta_0 = \bar{\boldsymbol{y}} - \gamma \bar{\boldsymbol{c}},$$

where $r_{\boldsymbol{yc}}$ is the sample correlation coefficient between $\boldsymbol{y}$ and $\boldsymbol{c}$, and $s_{\boldsymbol{y}}$ and $s_{\boldsymbol{c}}$ are the standard deviations of $\boldsymbol{y}$ and $\boldsymbol{c}$, respectively.

[2] For example, the vector $(7, 4, 4, 2)$ has ranks $(4, 2.5, 2.5, 1)$.

[3] Find the alternative with the highest response value among a set of $n \geq 2$ alternatives.

ID	Name	Obs.	Predictors
1	Abalone	4,177	8
2	AFL	41	5
3	Air	41	6
4	Airfoil	1,503	5
5	Algae	340	11
6	Athlete	202	8
7	Basketball	96	4
8	Birth weight	189	8
9	Body fat	252	13
10	Bone	42	6
11	Car	93	21
12	Cigarette	528	7
13	Concrete	1,030	8
14	Contraception	152	6
15	CPU	209	6
16	Crime	47	15
17	Diabetes	442	10
18	Diamond	308	4
19	Dropout	63	17
20	Excavator	33	4
21	Fish	413	3
22	Fuel	51	5
23	Gambling	47	4
24	Highway	39	11
25	Hitter	263	19
26	Home	3,281	4
27	Homeless	50	7
28	Infant	101	3
29	Labor supply	5,320	5
30	Lake	69	10
31	Land	67	4
32	Lung	654	4
33	Mammal	58	7
34	Medical expenditure	5,574	14
35	Men	34	3
36	Mileage	398	7
37	Mine	44	4
38	Monet	430	4
39	Mortality	60	15
40	Movie	62	12
41	Mussel	44	8
42	News	39,644	52
43	Obesity	136	11
44	Occupations	36	3
45	Pinot	38	6
46	Pitcher	176	15
47	Plasma	315	12
48	Prefecture	45	5
49	Prostate	97	8
50	Reactor	32	10
51	Rebellion	32	6
52	Recycle	31	7
53	Rent	2,053	10
54	Salary	52	5
55	SAT	50	4
56	Schooling	3,010	22
57	Tip	244	6
58	Vote	159	5
59	Wage	4,360	10
60	White wine	4,898	11

Table 2: Data sets used in the empirical comparison.

Note that β_0 can be omitted (set to 0) for all models when predictors and responses are standardized. Some authors do not include the scale parameter γ when the loss function is invariant under scaling. In this article, we are interested in regression under squared error loss, which is not invariant under scaling. Inclusion of γ is therefore crucial.

4. Empirical analysis

We examined the predictive accuracy of simple regression models on a large collection of real-world data sets. We assessed the predictive accuracy of different algorithms using root mean squared prediction error (RMSE).

Data sets. We used 60 publicly available data sets from varying domains. Sources included online data repositories, statistics and data-mining competitions, packages for R statistical software, textbooks, and research publications. The number of observations ranged from 31 to 39,644, the number of predictors from 3 to 52. Table 2 shows the number of observations and the number of predictors in each data set. Many data sets are from earlier studies (Czerlinski et al., 1999; Şimşek, 2013). Detailed information about the data sets can be found in the Online Appendix.

We had difficulties finding data sets for regression where $p > n$. Although regularized linear models have been developed primarily for problems where $p > n$, such data naturally mostly seem to occur in (binary) classification contexts in biology and genomics. Nonetheless, $p > n$ situations occurred in our learning curve analysis for small training-set sizes.

For each data set, the response was standardized and all predictors were standardized and directed. Missing predictor values were mean imputed and observations with missing response values were removed from the data set.

Benchmark models. We chose to include three benchmark models, described below. Two are state-of-the-art regression models. We included the OLS model for historic reasons.

(1) The OLS model minimizes the mean squared error between predicted and true values on the training data. We used the R (R Core Team, 2015) built-in function *lm* for estimating OLS. Whenever there were p predictors, n observations, and $p > n$, we used only the $n - 1$ predictors that were most correlated with the response.

(2) The elastic net (Zou and Hastie, 2005) is a state-of-the-art regularized linear regression model. Regularized linear models were originally developed to overcome the estimation difficulties of OLS (Hoerl and Kennard, 1970). They attempt to optimize prediction accuracy by finding the happy medium between simplicity and complexity. The elastic net has two main parameters. Parameter $\lambda \geq 0$ controls the overall strength of regularization. The elastic net reduces to OLS for $\lambda = 0$. Parameter $0 \leq \alpha \leq 1$ controls the amount of *ridge* versus *Lasso* characteristics. The elastic net reduces to two other regularized linear models, ridge regression (Hoerl and Kennard, 1970) when $\alpha = 0$ and the Lasso (Tibshirani, 1996) when $\alpha = 1$. We used the R package *glmnet* (Friedman et al., 2015) to estimate the elastic net. The parameters α and λ were jointly optimized using 10-fold cross-validation on a two-dimensional grid with $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ and λ on the built-in search path of glmnet, which uses a log-spaced grid with a maximum of 100 candidate values. We also tested ridge regression and the Lasso in our empirical study and found their results to be very similar to those of the elastic net.

(3) Random forest regression (Breiman, 2001) is a non-parametric and non-linear regression model. It optimizes prediction accuracy by fitting an ensemble of regression trees. We used the R package *randomForest* (Liaw and Wiener, 2002) with `ntree = 500` trees per forest.

Results. We show three sets of results. Figure 1 shows the mean RMSE of each algorithm across 60 data sets, computed using 10-fold cross-validation. These estimates of the prediction error correspond to large training-set sizes relative to the total size of the data set (90% of available observations).

Figure 2 shows learning curves averaged across 60 data sets, as the training set varied in size from 4 to 100. The figure shows learning curves for all models except mean prediction, random weights, and OLS.[4] The test set consisted of 10% of the total number of observations in the data set and did not overlap with the corresponding training set. The estimation procedure was repeated 100 times for each training-set size and algorithm. There is a slight

[4] OLS overfits for small ratios of n/p. Resulting average RMSEs were outside the figure boundaries because some data sets had a large number of predictors p.

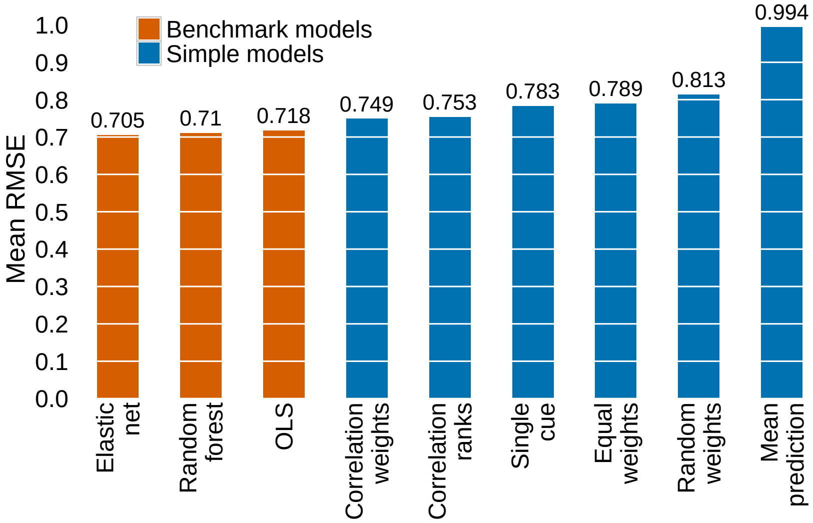

Figure 1: 100 × 10-fold cross validated root mean squared error (RMSE) across 60 data sets. OLS = ordinary least squares.

increase toward the end of the learning curve because the means were calculated on fewer data sets for higher training-set sizes. The number of data sets large enough to be eligible for a given training-set size is indicated at the top of the figure.

Finally, we present learning curves of various algorithms in individual data sets. Figure 3 shows learning curves in the data sets *Diabetes, Prostate*, and *SAT*. Figures A.1 to A.3 in the Online Appendix present the learning curves in all remaining data sets.

We first compare simple regression models to benchmark models collectively. We then comment on results within the groups of simple and benchmark models, respectively.

Averaged across 60 data sets, simple models were collectively outperformed by all benchmark models for larger training-set sizes.[5] However, equal weights and correlation ranks outperformed all competing models for training-set sizes below 15 on the mean learning curve. In addition, the learning curves in individual data sets show that for many data sets, there is at least one simple model that performed well across large parts of the learning curve. Let us define the minimum error curve as the algorithm with minimum error among all algorithms as a function of training-set size. Then, in 22 of 60 data sets, simple models occupied the entire minimum error curve except for possibly one training-set size. In another 21 data sets, simple models occupied at least half of the minimum error curve. The 17 remaining data sets were dominated by benchmark models.

On many data sets some simple models performed very well while others performed very poorly, rather than all simple models performing equally well. A good example is the

[5] The end of the learning curve shows the average across all 30 data sets that were large enough for training-set sizes of 100. The cross-validation-based analysis of Figure 1 shows the average across all 60 data sets for training-set sizes ranging from 27 to 35,679 observations, corresponding to 90% of the total size of the respective data sets. The two analyses show similar results.

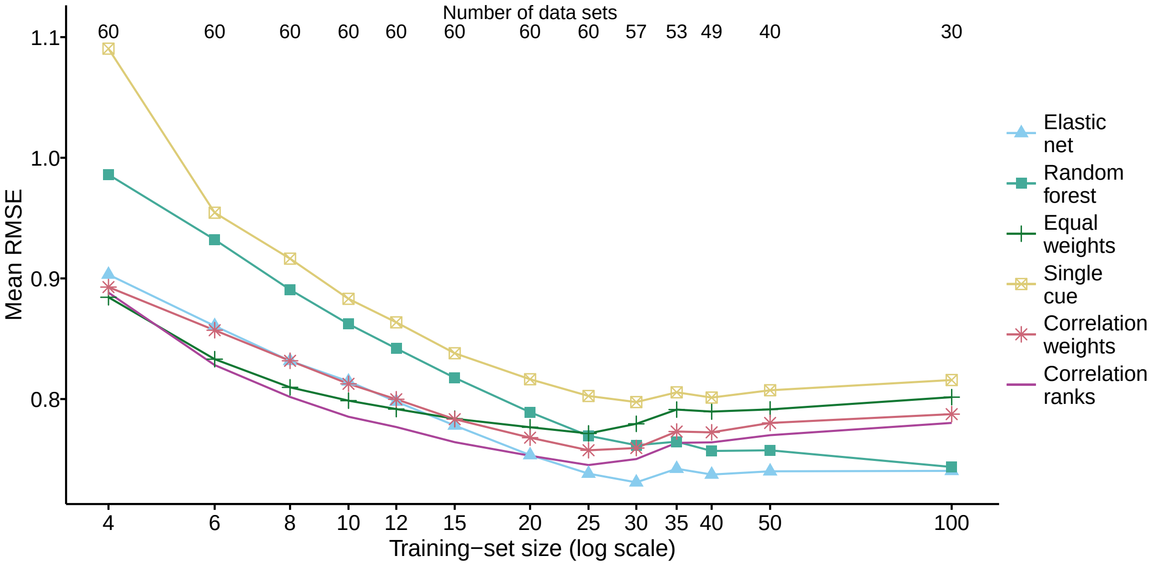

Figure 2: Learning curves across 60 data sets. The number of large-enough data sets per training-set size is indicated on top of the graph. Mean RMSEs for OLS are beyond the plot range and have not been plotted.

SAT data set shown in Figure 3, which is one of the few data sets for which both equal weights and correlation ranks perform poorly, but for which single cue is the best-performing algorithm across the entire learning curve.

The data sets *Prostate* and *Diabetes* have been used to illustrate the favorable prediction performance of the elastic net and other sophisticated regression models in the past (Tibshirani, 1996; Efron et al., 2004; Zou and Hastie, 2005). Figure 3 shows that correlation weights outperformed the elastic net in both data sets in training-set sizes smaller than 30.

On the mean learning curve, correlation ranks outperformed all other simple models across the entire curve. However, on individual data sets, correlation ranks was often outperformed by one or more of the other simple models. In fact, in almost all data sets, the learning curve of correlation ranks lay in between those of equal weights and correlation weights, independent of which of the two latter models performed better. This confirms the intuition that correlation ranks is an intermediately complex model that is able to perform well in situations of scarce information (similar to equal weights) but can also exploit the benefits of weighting predictors differently when there is enough information to reliably estimate the ranking of predictors.

Both equal-weights and single-cue regression share the property of performing either very well or very poorly on many data sets. Single-cue regression was the second-worst or worst model over the entire range of training-set sizes in 23 of the 60 data sets. Yet it was also the best-performing model across the entire learning curve in *SAT* and across large parts of the learning curve in *Body fat*. Equal weights outperformed all other models across the entire learning curve in data sets *Bone, Fuel, Pinot, Reactor, Rent*, and *Wage*. But it

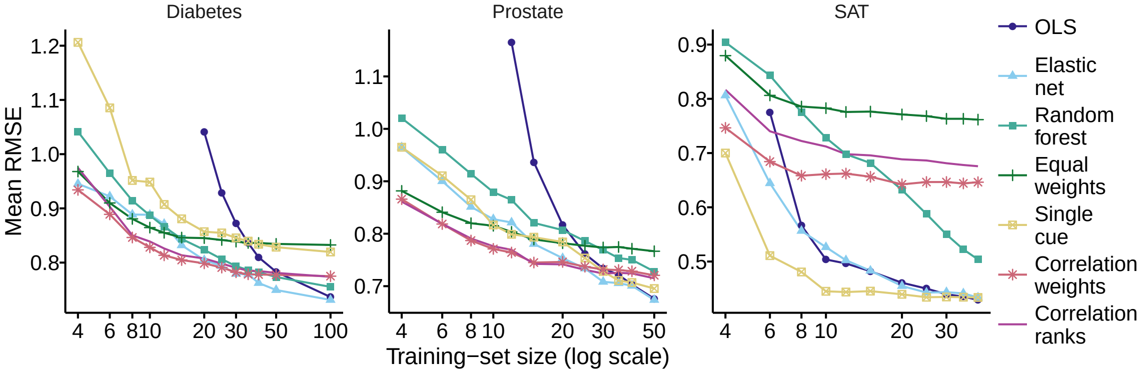

Figure 3: Individual learning curves for data sets *Diabetes*, *Prostate*, and *SAT*. Mean RMSEs above 1.2 have not been plotted. OLS = oridnary least squares.

was by far the worst model on large parts of the learning curve in *Diamond, Excavator, Fish*, and *SAT*.

Among benchmark models, OLS was outperformed by random forest regression and the elastic net, both on average and individually on most of the data sets. Elastic net generally outperformed random forest regression, especially on small training sets.

5. Discussion

Our analysis shows that simple regression models, for example, equal-weights regression, routinely outperform not only multiple linear regression but also state-of-the-art regression models, especially on small training sets.

None of the simple models we examined predicted well in all data sets. But in nearly all data sets, there was at least one simple model that predicted well.

Because OLS has severe estimation difficulties with small training sets, it would be reasonable to expect simple regression models to perform better than OLS on small training sets. However, we did not expect the simple models to be able to compete with state-of-the-art regularized linear models such as the elastic net.

Regularized linear models attempt to optimize prediction accuracy by searching through a possibly infinite-dimensional hypothesis space of linear models, ranging from a sparse linear model to the full, complex OLS solution. All simple models considered here are special cases of the linear regression model. Even though we tested only four of them, these simple models could sometimes outperform the carefully-optimized elastic net. These results indicate that it may be possible to substantially reduce the size of the hypothesis space of linear models needed to make good inferences. In other words, it is possible to make good inferences based on simple models if one only knows which simple model to choose.

Future work could examine models that adaptively choose between a few but maximally different simple models. For example, a model that chooses between single cue, equal weights, and correlation ranks using only information in the training data could be a fast and robust alternative to current state-of-the-art models, while being computationally less challenging. The main question will be whether this algorithm can choose an appropriate simple model on the basis of only a small number of training examples.

An important research direction is to examine whether people can intuitively pick an appropriate simple model for a given problem. Such a finding may explain how people often make good decisions despite their bounded cognitive capacities.

Acknowledgments

We would like to thank Marcus Buckmann and three anonymous reviewers for comments on earlier drafts.

References

F. Hutton Barron and Bruce E. Barrett. Decision Quality Using Ranked Attribute Weights. *Management Science*, 42(11):1515–1523, 1996.

Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

Henry Brighton. Robust Inference with Simple Cognitive Models. In *AAAI Spring Symposium: Between a Rock and a Hard Place: Cognitive Science Principles Meet AI-Hard Problems*, pages 17–22, 2006.

Jean Czerlinski, Gerd Gigerenzer, and Daniel G. Goldstein. How good are simple heuristics? In *Simple heuristics that make us smart*, pages 97–118. Oxford University Press, 1999.

Jason Dana and Robyn M. Dawes. The superiority of simple alternatives to regression for social science predictions. *Journal of Educational and Behavioral Statistics*, 29(3): 317–331, 2004.

Clintin P. Davis-Stober, Jason Dana, and David V. Budescu. A Constrained Linear Estimator for Multiple Regression. *Psychometrika*, 75(3):521–541, 2010.

Robyn M. Dawes. The robust beauty of improper linear models in decision making. *American psychologist*, 34(7):571, 1979.

Robyn M. Dawes and Bernard Corrigan. Linear models in decision making. *Psychological Bulletin*, 81(2):95–106, 1974.

Victor DeMiguel, Lorenzo Garlappi, and Raman Uppal. Optimal versus naive diversification: How inefficient is the 1/N portfolio strategy? *Review of Financial Studies*, 22(5): 1915–1953, 2009.

Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and others. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

Andrew S. C. Ehrenberg. How Good Is Best? *Journal of the Royal Statistical Society. Series A (General)*, 145(3):364, 1982.

Hillel J. Einhorn and Robin M. Hogarth. Unit weighting schemes for decision making. *Organizational Behavior and Human Performance*, 13(2):171–192, 1975.

Jerome Friedman, Trevor Hastie, Noah Simon, and Rob Tibshirani. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models, April 2015.

Gerd Gigerenzer, Peter M. Todd, and the ABC Research Group. *Simple heuristics that make us smart*. Oxford University Press, USA, 1999.

Andreas Graefe. Improving forecasts using equally weighted predictors. *Journal of Business Research*, 68(8):1792–1799, 2015.

Arthur E. Hoerl and Robert W. Kennard. Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55, 1970.

Robert C. Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.

James E. Laughlin. Comment on" Estimating coefficients in linear models: It don't make no nevermind.". *Psychological Bulletin*, 85(2):247–253, 1978.

Andy Liaw and Matthew Wiener. Classification and regression by randomForest. *R news*, 2(3):18–22, 2002.

Alexander D. Lovie and Patricia Lovie. The flat maximum effect and linear scoring models for prediction. *Journal of Forecasting*, 5(3):159–168, July 1986.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

Frank L. Schmidt. The Relative Efficiency of Regression and Simple Unit Predictor Weights in Applied Differential Psychology. *Educational and Psychological Measurement*, 31(3): 699–714, 1971.

Özgür Şimşek. Linear decision rule as aspiration for simple decision heuristics. In *Advances in Neural Information Processing Systems*, pages 2904–2912, 2013.

Özgür Şimşek and Marcus Buckmann. Learning From Small Samples: An Analysis of Simple Decision Heuristics. In *Advances in Neural Information Processing Systems*, pages 3141–3149, 2015.

Robert Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Detlof von Winterfeldt and Ward Edwards. Costs and payoffs in perceptual research. *Psychological Bulletin*, 91(3):609, 1982.

Howard Wainer. Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, 83(2):213–217, 1976.

Howard Wainer. On the sensitivity of regression and regressors. *Psychological Bulletin*, 85 (2):267–273, 1978.

Niels G. Waller and Jeff A. Jones. Correlation Weights in Multiple Regression. *Psychometrika*, 75(1):58–69, 2009.

Alexander G. Wesman and George K. Bennett. Multiple regression vs. simple addition of scores in prediction of college grades. *Educational and Psychological Measurement*, 1959.

Samuel S. Wilks. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 3(1):23–40, 1938.

Hui Zou and Trevor Hastie. Regularization and variable selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.