Would you prefer your coefficients with a little bias, or rather with a lot of variance?

Marjolein Fokkema and Samuel Greiff

Accepted author manuscript

Introduction

One of the core tasks of psychological assessment is the prediction of human behavior. In turn, one of the core tasks of empirical research in psychological assessment is to assess the extent to which psychological test scores predict human behavior. This is also reflected in publications in our journal: A Google Scholar search indicates that 119 of 485 EJPA publications in the last ten years included the word 'prediction'.

Regression analysis is the weapon of choice for evaluating the predictive power of psychological test scores. Meehl (1954) already argued that regression models optimally combine information for making predictions and consistently outperform humans in this task. Evidence in favor of this argument has been found in many studies since (e.g., Dawes, Faust & Meehl, 1989; Ægisdóttir et al., 2006).

Researchers in psychology tradionally prefer to use unbiased estimators for regression, like ordinary least squares (OLS) or maximum likelihood (ML). The unbiasedness of these estimators may sound like a favorable property, perhaps also due to the connotation of the word 'bias' with prejudice or unreasoned judgment. In statistics, however, bias does not have this normative connotation and merely refers to a systematic deviation in the expected value of an estimate from the quantity it estimates (also referred to as the 'true value'). In this editorial, we would like to focus on the question whether bias in regression estimates should indeed be avoided, or whether it may actually yield desirable consequences, especially in the context of psychological assessment and predicting human behavior.

We assume here that most prediction problems in psychological assessment share some similarities: They involve multiple potential predictor variables (i.e., psychological test scores) and one or more criterion variables. Furthermore, we assume that the main aim is to evaluate the extent to which the potential predictor variables contribute to predicting the criterion. The interest may be, for example, in estimating the optimal weights (regression coefficients) for predicting the criterion, or selecting only those psychological test scores that substantially contribute to prediction of the criterion. In this editorial, we focus on the specific case where there is a single continuous criterion variable, but our conclusions can be generalized to other types of responses within the generalized linear model (e.g., multivariate, categorical or count responses).

Through a simulation experiment, we will investigate the effects of bias in regression coefficients. The experiment can be replicated using the code provided in the supplementary material. As noted above, researchers in psychology tend to employ unbiased regression estimators. However, biased regression estimators are also available, like for example lasso, ridge and elastic net penalized regression (e.g., Tibshirani, 1996; Friedman, Hastie & Tibshirani, 2009). Such methods have been available for some time, but up to date have rarely - if ever - been applied in submissions to EJPA. In contrast, penalized regression methods are already widely applied in fields like genetics, machine learning, or neuroimaging, partly because they allow for analyzing datasets where the number of variables (p) exceeds the number of observations (N). Such $p > N$ datasets are more rarely encountered in the field of psychological assessment, but penalized regression may be beneficial for the analysis of $N > p$ datasets as well.

An in-depth discussion of penalized regression is out of the scope of this editorial, but an introduction aimed at researchers in psychology can be found in Chapman, Weiss and Duberstein (2016). In short: In OLS estimation, regression coefficients are estimated so as to minimize the loss function - the residual sum of squares (RSS). In lasso regression, a penalty term, which is a function of the sum of the absolute values of the regression coefficients, is added to the loss function. The higher the absolute values of the regression coefficients, the higher the penalty term. Thus, whereas OLS minimizes only the RSS, lasso regression minimizes both the RSS and the sum of the absolute values of the regression coefficients. The latter yields estimated coefficients which are biased towards zero.

Simulation experiment

Method

Our experiment is partly inspired by a study from Aluja and Blanch (2004), who assessed the extent to which socialized personality, scholastic aptitude, and study habits are predictive of academic achievement. They found scholastic aptitude to be the strongest predictor of academic achievement. We set up our simulation experiment in a similar vein: We generated three potential predictor variables (socialized personality, scholastic aptitude and study habits), each generated from a normal distribution with $\mu = 10$ and $\sigma = 1$. We calculated the response (academic achievement) as the sum of scholastic aptitude and a random error term. The random error term was generated from a normal distribution with $\mu = 0$ and $\sigma = 1$. Thus, there is one true predictor, scholastic aptitude, with a true regression coefficient of 1. The other two potential predictor variables, socialized personality and study habits, are noise variables, with true regression coefficients of 0. Furthermore, all predictor variables and the error term are independent.

We generated 1000 samples with $N = 150$ observations each. Even though this sample size may seem not very large, note that we created a rather easy regression problem: All variables are normally distributed, there is no multicollinearity, the one true predictor has a strong effect (i.e., the true correlation between scholastic aptitude and academic acievement is 0.71) and the noise variables are pure noise. In the real world, prediction

problems are likely less clear-cut, while sample sizes may be larger, and we can therefore expect the conclusions of our experiment to hold for many real-world studies.

In each of the samples generated thus, we fitted lasso and OLS regression models, each estimating the (linear) effect of the three potential predictors on academic achievement.

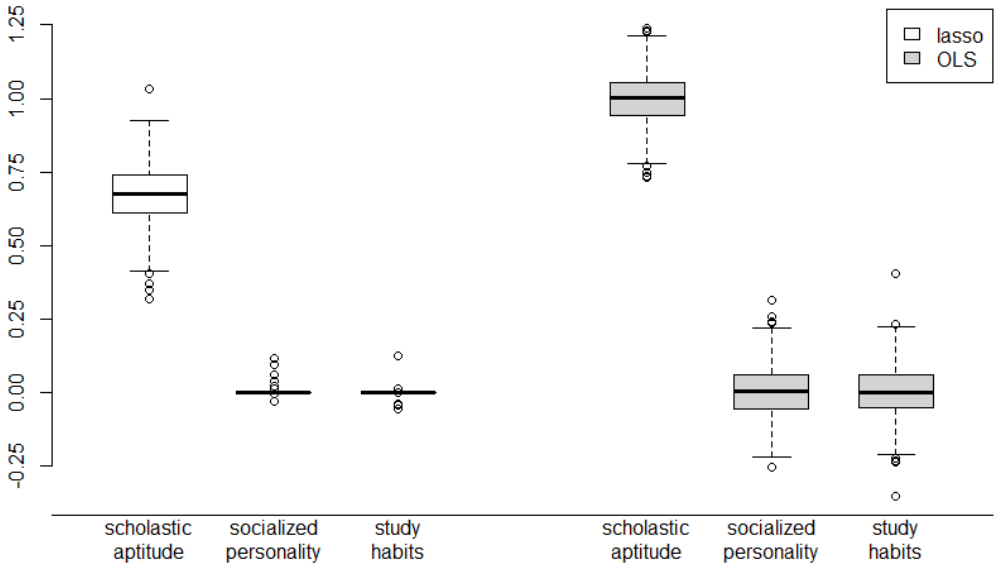

Figure 1. Boxplots of OLS and lasso estimated regression coefficients. Circles represent outliers, defined as values that deviate more than 1.5 times the interquartile range from the box.

Results

Figure 1 depicts the distribution of the estimated coefficients for each of the estimators. We see that, with exception of a few outliers, all lasso estimates for the noise variables are equal to 0. All lasso estimates for the true predictor variable are substantially higher than 0, but systematically lower than the true value of 1. Thus, the lasso estimates are biased towards 0.

The OLS estimates appear unbiased: For the noise variables, the mean of the estimates is equal to the true value of 0. For the true predictor, the mean of the estimates is also equal to the true value of 1. However, we observe considerable variance of the OLS estimates for both predictor and noise variables. Although on average, the OLS estimates are equal to the true values, in individual samples the estimates can substantially deviate from the true value (Figure 1).

To obtain effect sizes, we calculated standardized regression coefficients. For the noise variables (socialized personality and study habits), the standardized OLS coefficients show absolute values $>$ .10 in 9.75% of the cases, while the standardized lasso coefficients show

absolute values $> .10$ in 0% of the cases. Thus, unlike the lasso estimates, the OLS estimates may indicate a small effect, when in fact there is no effect.

These results reveal a beneficial effect of bias in regression estimates: The lasso coefficients perform well in estimating the coefficients of noise variables. If our aim is to select predictor variables, we may thus prefer lasso regression. If we want to obtain unbiased estimates of the regression coefficients, we may prefer OLS regression, but we should be aware that the unbiasedness comes at the cost of a substantial increase in variance of the estimates. With OLS we may approximate the 'true' parameter on average, but estimates from any single sample may strongly deviate from the true values.

We will now look at the performance of the lasso and OLS estimated models in terms of predictive accuracy. We used the models estimated above to generate predictions for new observations (i.e., observations which were not part of the training samples), based on their predictor variable values. In the real world, we would not know the response variable values for these new observations, but here we generated the data from a known distribution, so we also know the response variable values for the new observations. This allows us to assess how well the fitted models can predict the criterion (academic achievement) on new observations. We quantify this by the proportion of variance explained, or R^2, with higher values indicating higher predictive accuracy.

We applied the models fitted above to 1000 samples of 1000 new observations. Note that we generated a large number of test observations, so as to measure the association between predicted and true values as precisely as possible. Figure 2 depicts the resulting distribution of R^2 values.

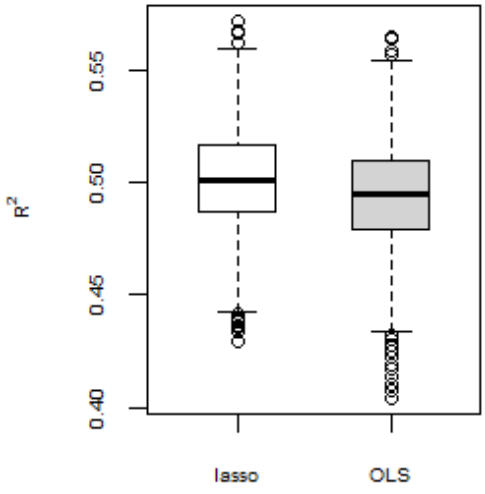

Figure 2. Boxplots of variance explained among new observations by the fitted OLS and lasso models. Circles represent outliers, defined as values that deviate more than 1.5 times the interquartile range from the box.

We see that the lasso slightly outperforms the unbiased OLS in terms of predictive accuracy. The average correlation between predicted and true criterion values was 0.501 (SD = 0.022) for the lasso models and 0.494 (SD = 0.024) for the OLS models. The difference may seem small, but almost equals a third of the standard deviations of the two methods, indicating a medium effect size.

Of course, the correlation between true and predicted values of the response cannot reveal possible bias in the predictions. We therefore calculated the average of the true and predicted criterion values over all test datasets. The mean of the true criterion variable values was 10, for the lasso predictions it was 10.003 and for the OLS predictions it was 10.003. Thus, the lasso predictions do not have more bias than the OLS predictions.

Discussion

If we take the cost of increased variance into account, the unbiasedness of estimators such as OLS may not be a favorable property. Biased regression models like the lasso may provide better predictions on new observations. For one of the core tasks of psychological assessment, the prediction of human behavior, biased regression estimates could therefore be advantegeous. Furthermore, in psychological assessment, we may often want to predict outcomes using an as-small-as-possible set of predictors (e.g., psychological tests), as their assessment is often costly. Our experiment illustrates how lasso may outperform OLS regression in selecting the relevant predictor(s).

We would like to note that the topic of optimal coefficient estimation is not new within the field of psychological assessment. Some 50 years ago, authors in the field already argued that OLS estimated coefficients may too strongly adapt to the idiosyncrasies of the sample, yielding lower predictive accuracy when the fitted regression model is applied to new observations (e.g., Schmidt, 1971; Wainer, 1976). These authors have advocated a unit-weighing scheme, where all predictors receive equal weight in the prediction equation. Recent and insightful discussions on this topic can be found in Grove (2003) and Waller (2008). In the unit-weighing line of thinking, it could be argued that in the prediction of human behavior one of the most important task is the selection of relevant predictor variables; once we know the set of relevant predictors, assigning equal weights may yield almost equally accurate predictions. However, the results of our experiment indicate that lasso estimated coefficients perform better than those of OLS in terms of predictive accuracy on new observations, but one could also argue that the difference in predictive accuracies are small from a practical point of view. From the viewpoint of variable selection, however, the possible benefits of lasso seem more substantial.

In conclusion, with this editorial we have reminded ourselves and hopefully also our readers how psychological assessment often has a very practical aim: predicting human behavior in the real world. We aimed to illustrate how bias in regression estimates - which may theoretically sound undesirable - may be practically beneficial. When the aim is to predict human behavior, we may prefer an algorithm that aims to optimize predictive accuracy and variable selection instead of traditional unbiased methods, which aim to recover the ‘true’ values of the parameters of a stochastic model that generated the data.

For an insightful and engaging discussion of this issue, the difference between prediction and explanation, readers are referred to Breiman (2001). For further reading on penalized regression methods, readers are referred to the intoductions by Chapman & Weiss (2016) and James, Witten, Hastie, & Tibshirani (2013), or to the more comprehensive and in-depth discussion by Friedman, Hastie & Tibshirani (2009).

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., … & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*(3), 341-382.

Aluja, A., & Blanch, A. (2004). Socialized personality, scholastic aptitudes, study habits, and academic achievement: Exploring the link. *European Journal of Psychological Assessment, 20*(3), 157-165.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science, 16*(3), 199-231.

Chapman, B. P., Weiss, A., & Duberstein, P. R. (2016). Statistical learning theory for high dimensional prediction: Application to criterion-keyed scale development. *Psychological Methods, 21*(4), 603-620.

Dawes, R., Faust, D., & Meehl, P. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Friedman, J., Hastie, T., & Tibshirani, R. (2009). *The elements of statistical learning*. New York, NY, USA: Springer Verlag.

Grove, W. (2003). *Correction and Extension of Wainer's "Estimating Coefficients in Linear Models: It Don't Make No Nevermind"*. Unpublished manuscript. url: https://pdfs.semanticscholar.org/cb7d/904302bf4fb95ccf9aa8e1557e4cf177371d.pdf

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York, NY< USA: Springer Verlag.

Meehl, P. (1954). *Clinical vs. statistical prediction: A theoretical analysis and a review of the evidence.* Minneapolis: University of Minnesota Press.

Schmidt, F. L. (1971). The relative efficiency of regression and simple unit predictor weights in applied differential psychology. *Educational and Psychological Measurement, 31*(3), 699-714.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological), 58*(1), 267-288.

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

Waller, N. G. (2008). Fungible weights in multiple regression. *Psychometrika, 73*(4), 691.

Supplementary material

```
##
## R replication code for EJPA editorial 'Would you prefer your coefficients
##   with a little bias, or rather with a lot of variance?'
##
## Code author: Marjolein Fokkema
##
## Code version date: 13 november 2018
##


## Load library for fitting lasso models:
library("glmnet")

## Set number of simulation replications and samples size:
nreps <- 1000
N <- 150

## Set up lists for saving results:
lasso.mods <- OLS.mods <- list()
x <- rep(NA, times = nreps)
lasso.coefs <- OLS.coefs <- lasso.coefs_std <- OLS.coefs_std <-
  data.frame(schol_apt=x, soc_pers=x, stud_hab=x)


## Simulate data and fit models:
set.seed(42)
for (i in 1:nreps) {
  x <- rnorm(N, mean = 10)
  y <- x + rnorm(N)
  x <- cbind(1, schol_apt = x,
             soc_pers = rnorm(N, mean = 10, sd = 1),
             stud_hab = rnorm(N, mean = 10, sd = 1))
  sds_x <- apply(x[,-1], 2, sd)
  sd_y <- sd(y)
  lasso.mods[[i]] <- cv.glmnet(x = as.matrix(x), y = y, alpha = 1)
  data <- data.frame(x[,-1], y = y)
  OLS.mods[[i]] <- lm(y ~ ., data = data)
  lasso.coefs[i,] <- as.matrix(coef(lasso.mods[[i]]))[3:5,]
  lasso.coefs_std[i,] <- (lasso.coefs[i,] * sds_x) / sd_y
  OLS.coefs[i,] <- coef(OLS.mods[[i]])[2:4]
  OLS.coefs_std[i,] <- (OLS.coefs[i,] * sds_x) / sd_y
}


## Evaluate and plot coefficient estimates:
results1 <- data.frame(estimator = rep(c("OLS", "lasso"), each = 3*nreps))
results1$variable <- rep(c("schol_apt", "soc_pers", "stud_hab"), each = nreps,
                         times = 2)
results1$coef <- c(unlist(OLS.coefs), unlist(lasso.coefs))
cols <- c("white", "lightgrey")
```

```
boxplot(coef ~ variable + estimator, data = results1,
        col = rep(cols, each = 3), boxwex = .6,
        axes = FALSE, at = c(1:3, 5:7))
vars <- c("\nscholastic\naptitude", "\nsocialized\npersonality",
          "\nstudy\nhabits")
axis(side = 1, lwd.ticks = 0, at = 0:8, labels = c("", vars, "", vars, ""))
axis(side = 2, at = seq(-.25, 1.25, by = .25))
legend("topright", fill = cols, legend = c("lasso", "OLS"))
lasso_noise <- c(lasso.coefs_std$soc_pers, lasso.coefs_std$stud_hab)
OLS_noise <- c(OLS.coefs_std$soc_pers, OLS.coefs_std$stud_hab)
# standardized OLS coefficients for noise variables with absolute
# values > .10:
mean(OLS_noise > .10  | OLS_noise < -.10) * 100
# standardized lasso coefficients for noise variables with absolute
# values > .10:
mean(lasso_noise > .10  | lasso_noise < -.10) * 100


## Generate test data and assess predictive accuracy:
Ntest <- 1000
lasso.preds <- OLS.preds <- ys <- list()
lasso.cors <- OLS.cors <- rep(NA, times = nreps)
set.seed(43)
for (i in 1:nreps) {
  x <- rnorm(Ntest, mean = 10)
  ys[[i]] <- x + rnorm(Ntest, sd = 1)
  x <- cbind(1, schol_apt = x,
             soc_pers = rnorm(Ntest, mean = 10),
             stud_hab = rnorm(Ntest, mean = 10))
  lasso.preds[[i]] <- predict(lasso.mods[[i]], newx = as.matrix(x))
  lasso.cors[i] <- cor(ys[[i]], lasso.preds[[i]])^2
  newdata <- data.frame(x[,-1])
  OLS.preds[[i]] <- predict(OLS.mods[[i]], newdata = newdata)
  OLS.cors[i] <- cor(ys[[i]], OLS.preds[[i]])^2
}


## Assess and plot R2 values:
results2 <- data.frame(lasso = lasso.cors, OLS = OLS.cors)
boxplot(results2, ylab = expression(R^{2}), boxwex = .5, cex.axis = .6,
        cex.lab = .6, col = cols)
# mean and SD of lasso R2s:
round(mean(lasso.cors), digits = 3)
round(sd(lasso.cors), digits = 3)
# mean and SD of OLS R2s:
round(mean(OLS.cors), digits = 3)
round(sd(OLS.cors), digits = 3)


## Assess bias in predictions:
# mean of the true criterion variable values:
round(mean(sapply(ys, mean)), digits = 3)
# mean of the lasso predictions:
round(mean(sapply(lasso.preds, mean)), digits = 3)
## mean of the OLS predictions:
round(mean(sapply(OLS.preds, mean)), digits = 3)
```