Cross Validation: A method every psychologist should know

Mark de Rooij & Wouter Weeda

Leiden University

Institute of Psychology

Methodology & Statistics Unit

October 11, 2019

Please address correspondence to Mark de Rooij, Methodology and Statistics department, Institute of Psychology, Leiden University. PO Box 9555, 2300 RB Leiden, The Netherlands. Email: rooijm at fsw.leidenuniv.nl

Abstract

Cross validation is a statistical procedure that every psychologist should know. Most of us are possibly familiar with the procedure in a global way, but haven't used it for the analysis of our own data. We introduce cross-validation for model selection in a general sense, provide an R-package for analysis, and show a set of tutorial examples for oftentimes encountered research problems in the social sciences. Cross validation can be used as an easy-to-use alternative to null-hypothesis testing, that does not make as many assumptions.

KEY-WORDS: Predictive accuracy; reproducibility; p-values; cross-validation; prediction

The simple idea of splitting a sample in two and developing the hypothesis on one part and testing it on the remainder may perhaps be said to be one of the most serious neglected ideas in statistics

G.A. Barnard (1974)

1 Introduction

Null-hypothesis significance testing is still the dominant paradigm in psychological research despite numerous rounds of debate (Rozeboom, 1960; Chow, 1998; Cohen, 1994; Hagen, 1997; Krueger, 2001; Nickerson, 2000). Many alternatives have been proposed such as a focus on estimation and confidence intervals (Cumming, 2014) or a Bayesian approach (see Wagenmakers, 2007). A specifically intriguing sentence in the paper by Wagenmakers reads as follows: "The universal yardstick for selecting between competing models is predictive performance". In this manuscript we follow this lead of predictive performance in a cross-validation framework.

Since early this century there is an increased interest in prediction, contrasted with explanation (Breiman, 2001; Shmueli, 2010). Explanatory data analysis starts with a theory about an empirical phenomenon. The statistical model is a translation of the theory into mathematical form and statistical inference (tests, standard errors, p-values) is used to test the theory. The parameters of the model, for example the regression weights in a multiple regression model, are the key elements of this explanatory model, because they provide the test of the theory.

In contrast, in a predictive model, the model itself is not of great interest, but the predictions the model generates are. In other words, in line with our previous example, the regression weights of a multiple regression model are *not* of interest, but only the predictions that the regression model makes are. Breiman (2001) in his seminal paper contrasted the inferential and prediction approaches. Further material can be found in Shmueli (2010) and a recent discussion in psychology can be found in Yarkoni and Westfall (2017).

The theory about predictive models rests on the bias variance trade-off. Simply stated, the more complex a model is (i.e. the more parameters we use), the better it will fit (less bias), but the more variable it's predictions will be (more variance). At the other hand, if we have a simple model, it will fit worse (more bias), but our model's predictions will be less variable (less variance). In a formal sense, when we fit a statistical model to a sample of data and investigate its predictive performance, it can be shown that the expected prediction error decomposes in (squared) bias and variance of the fitted model (Hastie et al., 2009). The bias represents how far the average estimated model is from the true population model, whereas the variance represents the variability of the estimated models from sample to sample. A detailed mathematical treatment can be found in Hastie et al. (2009) and in Matloff (2017); A more narrative treatment in psychology can be found in Yarkoni and Westfall (2017), Chapman et al. (2016), or McNeish (2015).

In the early psychometric literature there was already quite some work trading off bias and variance, most notably in the use of simple weighting schemes for regression (Lawshe and Schucker, 1959; Schmidt, 1971; Wainer, 1976; Pruzek and Frederick, 1978), but also in estimating shrunken regression weights (Rozeboom, 1979; Darlington, 1978). In both cases we allow for bias in the regression equation (simple unit weights or shrunken weights will, on average, not be equal to the population weights) to reduce the variance (i.e. unit weights, not depending on the data, have zero variance).

It is often acclaimed that ordinary least squares regression produces unbiased estimates of the population model parameters. This sounds reassuring but the statement is generally false. The claim should be that if the assumptions of regression are true, ordinary least squares regression produces unbiased estimates. This is an important difference which has lead to a great deal of confusion. Because in general the assumptions of a linear regression model are not strictly true, the regression coefficients might be biased. For example, if the population relationship between two variables is nonlinear, OLS linear regression cannot be on average

correct.

To return to the contrast of explanatory and predictive models: explanatory modelling thus seeks the true, unbiased model, so that a theory can be tested, whereas predictive models actively seek false (biased) but stable models. If a researcher therefore wants to test whether a coefficient in the true underlying probability model equals zero, (s)he better uses a statistical test resulting in a p-value, because cross validation does not consistently find the true underlying model. So, whereas p values test the hypothesis that an effect is zero in the population, cross validation asks the question whether the predictions become better if this effect is added to a model. Hagerty and Srinivasan (1991) showed that including true but small effects in a statistical model do not make predictions always better.

The early days of psychometrics showed quite s strong interest in cross validation (Mosier, 1951) in order to validate a regression equation estimated using one sample in another sample. The motivation was that the explained variance found in a single sample was (and is) often overly optimistic for the performance in a new sample. To investigate this issue cross validation was used. The study lead to several adjustments to the multiple correlation coefficient, see for example Darlington (1968), Rozeboom (1968), and Claudy (1978). Other researchers actually used two samples where a regression model was fitted to the first and validated using the second sample. Mosier (1951) was the first to propose something like what is nowadays often referred to as cross validation, i.e. fitting a regression model in the first sample and validating it in the second and fitting the model in the second sample and validating it in the first. The final cross validatory coefficient was then the average of the two obtained correlations. The advantage of the two sample model over adjustments is that no distributional assumptions have to be made (Browne, 2000).

The two sample cross validation is inefficient in the sense that both the calibration (or training) set and validation (or test) set are much smaller than the complete data set. Therefore the estimation was not as efficient as it could be and the validation set is also small. Stone (1974)

and Geisser (1975) tried to alleviate both by introducing leave-one-out cross validation, where every time $N-1$ persons are used to estimate a model and 1 person to validate; the procedure is repeated till every observation has been used as validation sample. Because the training samples are almost approximately as large as the complete data set, much more efficient use is made of the data.

All these attempts above use cross validation to assess one final regression equation, i.e. a model that has been obtained from the data in some way, say by using p-values. In machine learning cross validation is most often used to evaluate different modeling procedures and for variable or model selection. In the latter case every model under consideration is cross validated and the one with the smallest prediction error (i.e., loss) is selected. Such a procedure might be useful for psychological research, but is till today not often used by psychologists. One of the reasons might be that researchers do not know how to perform cross validation as there is no standard statistical software routine. SPSS, for example, does not have a cross validation tool implemented, nor has it the capability of fitting a regression model on part of the data, and using the fitted model to make predictions on another part of the data.

The goal of this paper is to provide a rough outline of cross validation for model selection, to provide easily useable software, and to show a set of empirical examples. We conclude the paper with a discussion on the use of the R-package, the difficulties of model selection in general and a relationship of our cross validation procedure with other forms of cross validation.

2 Cross Validation: Theory and an R-package

The oldest and simplest way to perform cross validation is *independent verification* in which we have two independent data sets: a calibration (sometimes called training) set ($\mathcal{C}$) and a validation (sometimes called test) set ($\mathcal{V}$). We will index the observations in the calibration set with $i = 1, \ldots, N_{\mathcal{C}}$ and those in the validation set with $j = 1, \ldots, N_{\mathcal{V}}$.

In the calibration set we fit the statistical model, for example $y_i = a + \mathbf{b}^{\mathsf{T}}\mathbf{x}_i + e_i$ to obtain

estimated parameters $\hat{a}$ and $\hat{\mathbf{b}}$. In the validation set, with the estimated parameters and the values on the predictor variables ($\mathbf{x}_j$) we compute predictions $\hat{y}_j = \hat{a} + \hat{\mathbf{b}}^{\mathsf{T}}\mathbf{x}_j$. With the predicted values $\hat{y}_j$ we compute the root mean squared error (RMSE) of prediction

$$RMSE_P = \sqrt{\frac{1}{N_{\mathcal{V}}}\sum_{j=1}^{N_{\mathcal{V}}}(y_j - \hat{y}_j)^2}.$$

In R (R Core Team, 2018) we can easily fit a model on the calibration set using

```
output = glm(y ~ x, data = calibrationdata)
```

to obtain the estimated parameters. With the output of the model we can make predictions for new data, i.e. our validation data, by using

```
predictions = predict(output, newdata = validationdata,
type = "response")
```

Finally, the root mean squared error of predictions can be computed using

```
sqrt(mean((predictions - validationdata$y)^2)).
```

We deliberately focus on an unstandardized measure on the scale of the original response variable. Within a sample there is a direct link between the mean squared error and the explained variance. Out of sample, however, this link is broken because the mean is not calibrated. Therefore, it might occur that model 1 predicts $\hat{y}_j = \{1, 2, 3, 4, 5\}$ and model 2 predicts $\hat{y}_j = \{7, 8, 9, 10, 11\}$, where the actual observations in the validation set are $y_j = \{6, 7, 8, 9, 10\}$. It is clear that for both sets of predictions the correlation with the observed outcome equals 1, whereas the $RMSE_P$ equals 5 for the first set of predictions and 1 for the second set. For more details see Alexander et al. (2015).

We will use cross validation for model selection. In this scenario we have two or more models of interest and are interested which model is best. In cross validation "best" is operationalised as having the smallest root mean squared error of prediction. In this paper, and the software we show, we are focussing on models from the family of generalized linear models. Many standard analysis tools often used in psychology fall in this family, such as the one and two sample t-test, one or multi-way analysis of variance, (multiple) regression, analysis of covariance, and logistic regression. Furthermore, this family is easily enriched for nonlinear regression models using polynomials (as we will show) or splines. See, for example, Fox (2016) for the general framework that links these analyses techniques.

An important disadvantage of independent verification is that we need two data sets. Instead of collecting two data sets, a rather simple idea is to only collect a single data set and then divide it in two independent sets. James et al. (2013) called this the validation set approach to cross validation. One part takes the role of the calibration set, whereas the other takes the role of validation set. With this strategy we can use cross validation having only one data set. This strategy is not very efficient, because only half of the data is used to fit the model. Mosier (1951) already pointed to the loss of information in such an approach. However, the idea of splitting a data set in a calibration and validation set can be recycled, that is we split the data into K independent parts and in turn every part takes the role of validation set for which the other $K - 1$ sets are the calibration set. This is called K-fold cross validation and leads to predictions for every observation in the data set. We thus utilize the data more efficiently. K is often chosen to be equal to 5 or 10. More about this choice in the next Section. Partitioning the data into K sets can be done in many ways and would lead to a slightly different estimate of the root mean squared error of prediction every time. To deal with these different estimates we need to repeat the cross validation several times (Harrell, 2015). Using a large number of replications (also termed repeats) solves the problem that one model is favoured due to a particular partitioning in sets. Repeating the K-fold cross validation a number of times also has

the advantage that we can count the number of times a particular model wins, that is, we can count the number of times a model has a smaller prediction error than the other models under investigation.

We implemented these ideas in the R-package **xvalglms**. The main function in this package is `xval.glm` that does repeated K-fold cross validation on a set of models. To use the function we need to define a list of models and then call the `xval.glm` function. Suppose we have a single predictor variable `x` and a response variable `y` and we would like to know whether predictions become more accurate if we use the predictor variable. Therefore we compare the prediction errors of a model with and without the predictor variable. This can be done as follows:

```
models <- vector(mode = "list", length = 2)
models[[1]] <- y ~ 1
models[[2]] <- y ~ 1 + x
output <- xval.glm(data = mydata, models)
```

In this example we specify two regression models, model one with only an intercept (i.e., without the predictor variable), and model 2 with an intercept and the predictor.

The function outputs the following information:

```
Results for (10-fold, 200 repeats)
  Model:          |   Wins |     2.5% |     mean |    97.5% |
  [1] y ~ 1       |     0% |    0.409 |    0.412 |    0.418 |
  [2] y ~ 1 + x   |   100% |    0.394 |    0.401 |    0.410 |
```

and a graph. In the above information we see for each model 1) the percentage of times the prediction error is the smallest from the models in competition; 2) the average prediction error (in the column 'mean') and the 95 percent confidence bounds for the prediction error (in the

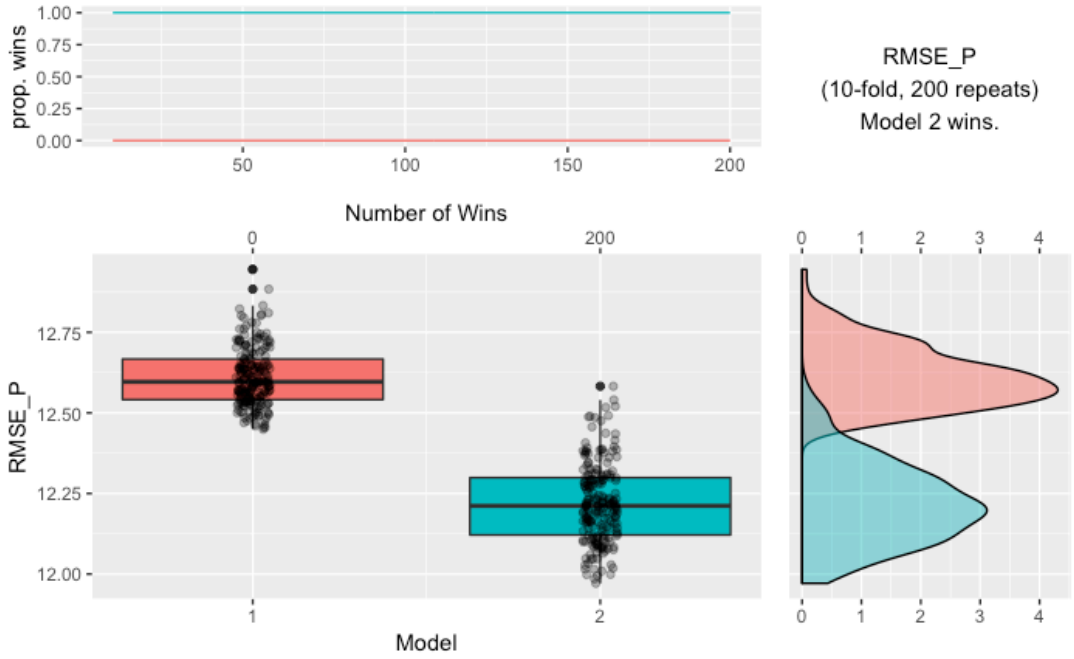

Figure 1: Example output of the xvalglms package.

columns '2.5%' and '97.5%').

The default graph shows three frames(see Figure 1). The upper frame shows for each model the proportion of wins of each model during the repeated cycles of 10 fold cross validation and can be used to verify whether the cross validation results stabilized. If the lines are not flat at the end, the researcher should ask for more repeats of the K-fold cross validation. The main frame shows boxplots of the 200 repetitions of prediction error (which we will also show in the remainder of this paper). The right hand side frame shows a density estimate for the boxplots. All this information is returned in an output object which can be saved.

2.1 Assumptions

Many introductory statistical textbooks emphasize the assumptions of statistical techniques. For linear regression for example, the most important assumptions are 1) a linear relationship between the explanatory variable and the response variable in the population, 2) normally distributed errors with constant variance (homoscedasticity), and 3) independence of the observations (Fox, 2016).

Once we are in the predictive mode, many of these standard assumptions are not needed anymore. By focussing on the bias variance trade-off we actively seek a bit of bias while diminishing variance. In that sense, we do not have to assume a linear relationship between the predictor and response variable, we can compare the predictive performance of several regression models, linear and nonlinear, and select one. The selected model is not necessarily equal to a true model (if that exists), it is the model that provides the best predictions as evaluated using the current data set. If the true population model is nonlinear but the optimal predictive model is linear, this means that the usual assumption of normally distributed residuals with constant variance is false. Therefore, we can conclude that such a distributional assumption for the residuals is not needed when we use cross validation for model selection.

We do, however, assume that the observations are independent. If the observations are not independent, we need to adapt the cross validation procedure to take into account the dependency. For clustered data, where we have for example repeated measures within a subject or participants clustered in teams, a clustered variant of cross-validation might be employed (Roberts et al., 2016).

What is important in cross validation is the loss-function employed to compute prediction error. In our examples we used the square root of the averaged squared difference between the predictions and the actual observations. Instead, we might focus on other loss functions such as the average absolute loss, which will probably lead to other 'optimal' models. We illustrate in the next section (and in the supplementary material) how to do this.

2.2 Choices to be made by the researcher

A user of the cross validation procedure has to make several choices: the choice of the number of folds, the choice of the number of repetitions, the choice of a loss function, and the choice of which models to compare. In the R-package we set some default values, which were chosen wisely and which we will motivate:

- The choice of **the number of folds** is itself a bias variance trade-off. With K equal to the number of observations in the sample an almost unbiased estimate is obtained of the prediction error, because the size of the training sample in each of the folds is almost equal to the sample size. This unbiasedness sounds good, but the variance from sample to sample is large. On the other hand, with $K = 2$ there is much less sample to sample variation, but the bias may be much larger because the training sample in each of the folds is only half of the sample size. Usually $K = 10$ is thought to be a good compromise and we use it as default. In small sample sizes, say smaller than 40, we advice to lower K to for example 5.

- Repetitions are important because they take away the randomness of results due to splitting the sample in K parts. Harrell (2015) advertises such repetitions. For the choice of **number of repeats** the default setting we choose is a large number, 200. Because nowadays we have very much computational power, this is not as issue. To evaluate whether 200 was enough a researcher may look at the proportion of wins of each model over the repetitions, which is visualised in the top figure of Figure 1. If the proportion of wins stabilized at the right hand side of this plot, the researcher can be confident that the number was large enough. If there is still large variation, we suggest to increase the number of repetitions. We like to note here also the interplay between the number of repetitions and the number of folds. If one uses leave one out cross validation (K equal to sample size), there is no need to repeat the cross validation because in every repetition exactly the same prediction error is estimated.

- The default for the **loss function** is the root mean squared error. This is a loss function that can be used for different type of distributions in the generalized linear models framework. However, in certain circumstances a researcher might want to change the loss function. If, for example, we need to make a decision for individual subjects, like selecting them

for treatment or a job, it is sensible to change the loss function for the logistic regression model to the misclassification rate. As another example, we might want to have a more robust version of a loss function and could choose absolute error loss instead of squared error loss for linear regression models. In the next section (Section 3.5) we show how to change the loss function and discuss the sensitivity of the choice of loss function on the model selection results.

- The last choice a researcher needs to make is which models to compare. Ideally one would like to include all possible models, but especially in cases with many variables comparing all possible models is not an option. The question of model selection can be answered from a more data-driven or theory-driven perspective. From a data-driven perspective a researcher could let the data decide which models to include, however these approaches can easily lead to under-estimation of prediction errors (Hastie et al., 2001). The current implementation of our cross-validation package does not include this form of estimation. From a theoretical perspective a researcher has to choose which models to include. This can be done based on, for example, theory, previous literature, but also based on information of how likely certain models are. This latter option is akin to prior selection in a Bayesian framework, where certain models are assigned less weight if they are highly unlikely. Note that these issues arise mainly in cases with a very large set of possible models. The number of variables in most psychological experiments is small enough to allow all models to be tested with our current framework.

3 Applications of Cross Validation

In this Section we show six applications of the cross validation methodology. The first corresponds to a two sample t-test, where the corresponding predictive question is whether either the overall mean or two separate means for the groups provide better predictions. In other

words: do the predictions become better if we use group information? The second application refers to a univariate regression where we would like to see whether a prediction of the response variable based on a predictor variable provides more accurate predictions than predicting based on the overall mean alone (i.e. an intercept only model). Furthermore, we might be interested in nonlinear relationships between the predictor variable and the response, where we ask the question whether a second or third order polynomial predicts better than a linear regression. Thirdly, we show a complex regression situation where we expect higher order interaction effects. Thereafter, in example four we look at a univariate regression with a dichotomous outcome. In this logistic regression we also compare polynomial models. When we choose another loss function in the cross validation procedure this might lead to other results, which we show in the fifth example. In the sixth example we compare two theories using cross validation, where we finally choose the theory that results in the most accurate predictions.

Before we can use the cross validation function we have to download and load the package. This can be done as follows:

```
library(devtools)
install_github("Github-MS/xvalglms")
library(xvalglms)
```

3.1 Two sample t-test situation

The data set we use for this example is described in the textbook by Howell (2015) but originally from Adams et al. (1996). The authors were interested in the theory that homophobia may be unconsiously related to anxiety of being or becoming homosexual. The Index of Homophobia was administered to 64 heterosexual males. Based on their score the participants were classified as either homophobic or nonhomophobic. The men than saw sexually explicit videos portraying homosexual and heterosexual behavior, and their sexual arousal was recorded. Adams

et al. (1996) reasoned that if homophobia was unconsciously related to anxiety about ones own sexuality, homophobic individuals would show greater arousal to homosexual videos than nonhomophobic individuals.

The data can be loaded from the internet using

```
library(foreign)
mfile = "http://www.uvm.edu/~dhowell/methods8/DataFiles/Tab7-5.sav"
Arousal = read.spss(mfile, to.data.frame = TRUE)
```

which puts the data in a data frame called `Arousal`.

In this case we are interested in the predictions based on the overall mean or the group means. We can define these two models as follows and then run our cross validation function.

```
models = vector(mode = "list", length = 2)
models[[1]] = Arousal ~ 1
models[[2]] = Arousal ~ 1 + Group
output = xval.glm(data = Arousal, models)
```

We thus specified a model to predict Arousal level based on the intercept alone (model 1) and a model to predict Arousal based on whether someone is homophobic or nonhomophobic (Group) as a predictor (model 2). Theoretically the question that cross validation will answer is whether adding the Group variable leads to more accurate predictions of Arousal than predicting it from the overall mean Arousal score only.

The output of this function is shown in Figure 2 where it can be seen that model 2 wins in all 200 cases and returns a much lower prediction error. That is, the prediction error is around 12.20 for the best model.

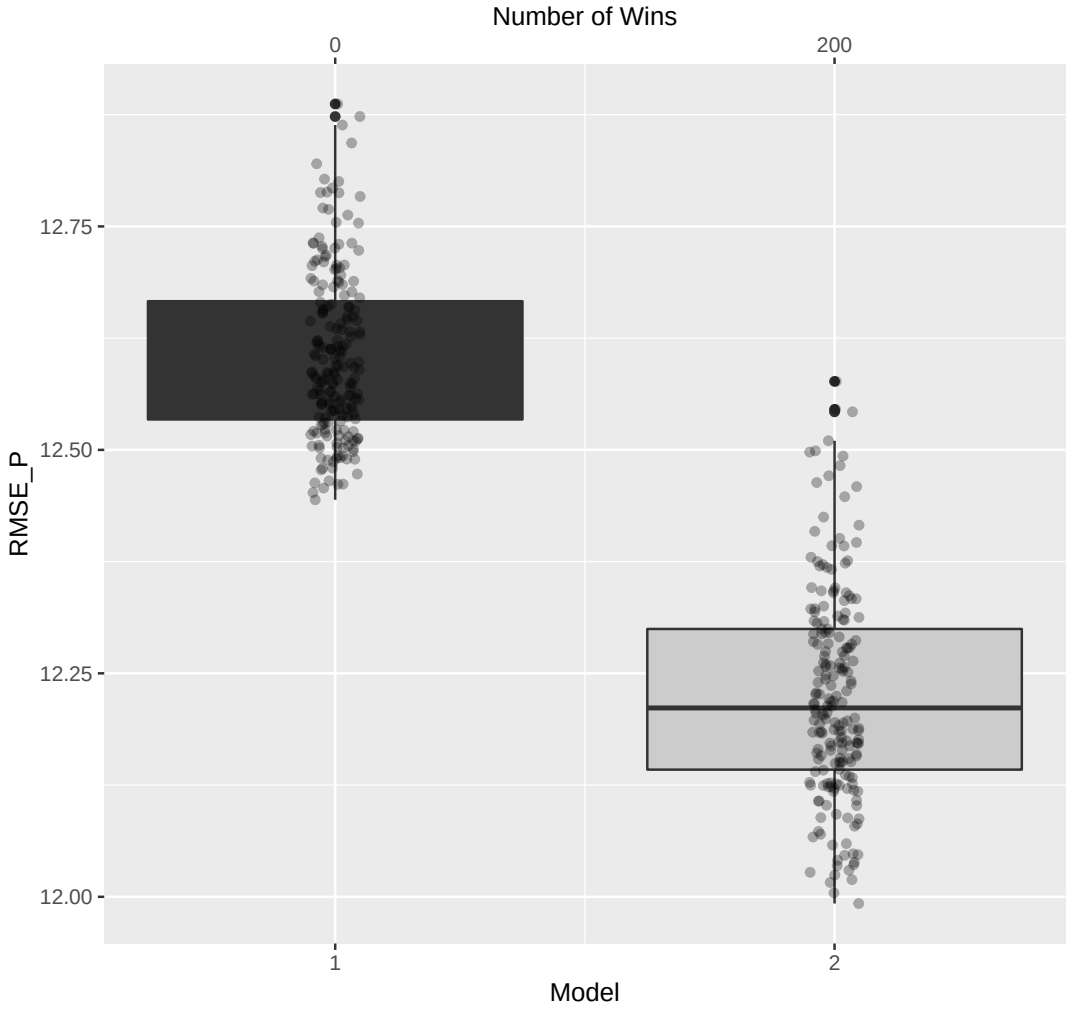

Figure 2: Cross validation results for the Arousal data. Model 1 makes predictions on the overall mean; Model 2 makes predictions based on the group means. The vertical axis represents root mean squared error of prediction. The number of wins are represented at the top of the window.

The two means that we need to use for future predictions are 24.00 and 16.50 for group 1 (Homophobic) and 2 (Nonhomophobic), respectively. Using these means to predict, we will be on average 12.20 units ($RMSE_P$) off from the true value in our prediction. This is quite a lot, but less than using only the overall mean to make predictions.

3.2 Linear regression situation

A general goal of the study conducted by Margolin and Medina as described in Wilcox (2017) was to examine how children's information processing is related to a history of exposure to marital aggression. Data were collected from 47 children. The variable `Aggression` is a measure of marital aggression that reflects physical, verbal, and emotional aggression during the last year and the variable `test` is a child's score on a recall test. We first read in the data in the data frame `agdat`.

```
mfile = "https://dornsife.usc.edu/assets/sites/239/docs/
marital_agg_dat.txt"
agdat = read.table(mfile, header = TRUE)
```

A scatterplot of aggression against the recall test score is given in the left hand side of Figure 3. It can be seen that there is a decreasing trend. That is, with increasing values of aggression the test scores go down, although some upward trend for higher values of aggression is visible.

Next we will specify the theoretically relevant models:

```
models = vector(mode = "list", length = 4)
models[[1]] = test ~ 1
models[[2]] = test ~ Aggression
models[[3]] = test ~ poly(Aggression,2)
```

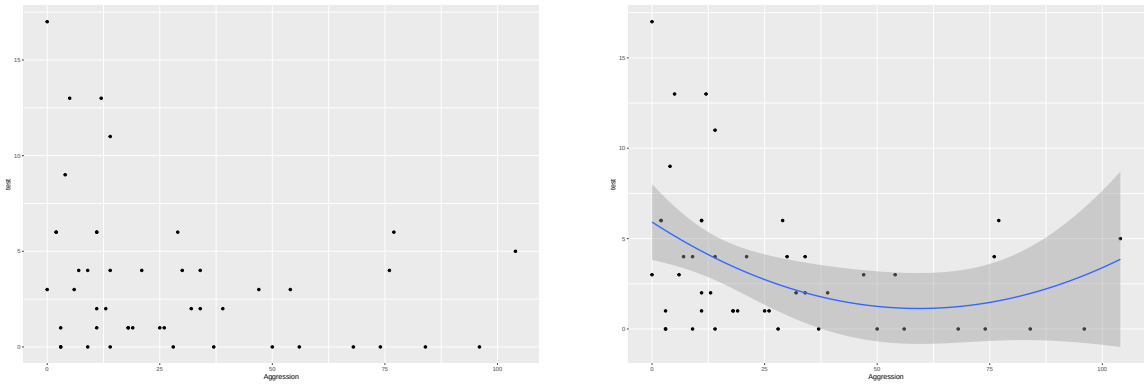

Figure 3: Left: Agression data. Right: Optimal fitted regression model for the Agression Data.

```
models[[4]] = test ~ poly(Aggression,3)

output = xval.glm(data = agdat, models)
```

The first model in this case only specifies the mean (an intercept). The second includes `Aggression` as a predictor. The question that cross validation answers is whether the inclusion of this predictor leads to more accurate out-of-sample predictions. We also include nonlinear models, more specifically a quadratic and cubic polynomial as model 3 and model 4. With the last line of code we call the cross validation function to compare the predictive power of the four models. The results are shown in Figure 4. The quadratic model predicts best, i.e., the prediction error for the quadratic model is lower than for the other models. We fitted the quadratic model to the complete data again and show the results in the right hand sided plot of Figure 3, which might subsequently be used to interpret the relationship between aggression and the recall test.

The standard approach to model selection would use change in explained variance and incremental F-tests. The F-tests and resulting p-values require assumptions such as normally distributed error terms with mean zero and constant variance. In the online supplementary material we show this analysis but also show that the assumptions are not tenable. The conditional mean of the residuals is not always zero and the distribution of the residuals is positively skewed. These might affect the test statistics. The cross validation procedure does not make

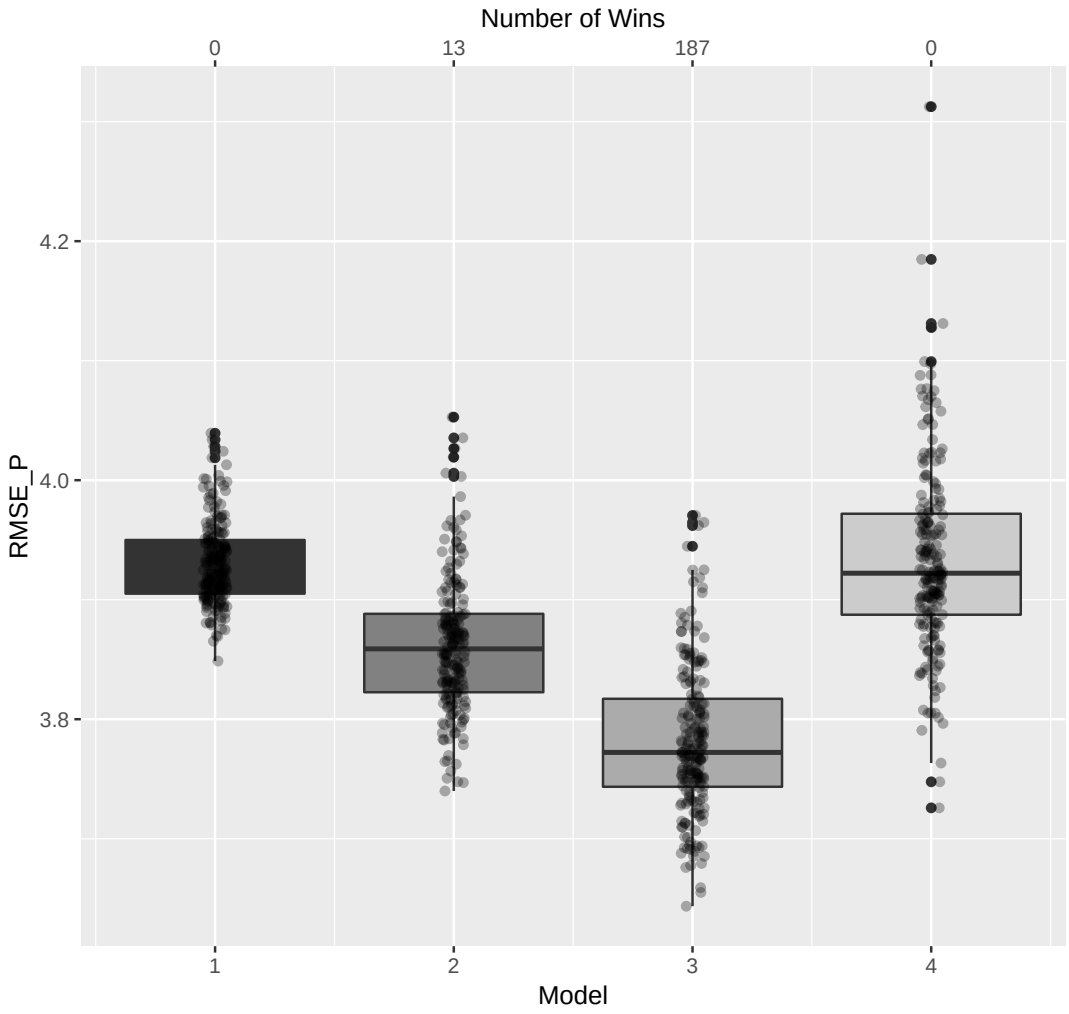

Figure 4: Cross validation results for Aggression data. Model 1 makes predictions based on the overall mean; Model 2 uses a first order polynomial of age, model 3 a second order, and model 4 a third order polynomial.

these assumptions.

3.3 Moderated Regression situation

In this section we use publicly available data from a study that examined the effects of mortality salience (M) among native Dutch people with varying levels of national identification (N) and self-esteem (S) on Attitudes about Muslims and multiculturalism (A) (Tjew-A-Sin and Koole, 2018a,b). In their original study the authors hypothesized and found a three-way interaction between mortality salience, national identification, and self-esteem. They also show a main effect of national identification, a main effect of self esteem, a marginal interaction between

self-esteem and mortality salience.

Exploring the data we found 5 subjects with extreme scores on the national identification and self esteem scales. In the supplementary material we show analyses with and without these five persons removed. Here, for illustrative purposes, we only report on the data with the five persons removed. We like to note that the three way interaction as observed by Tjew-A-Sin and Koole (2018b) is still present in this cleaned data set.

Whereas the authors used one complicated model and checked which effects were statistically significant, we approach this problem as a model selection issue where 16 different models are formulated. The simplest is the intercept only model (model 1), the most complex the one including all main effects, two-way interactions, and the three way interaction (model 16).

```
models = vector(mode = "list", length = 16)
models[[1]] = A ~ 1
models[[2]] = A ~ M
models[[3]] = A ~ N
models[[4]] = A ~ S
models[[5]] = A ~ M + S
models[[6]] = A ~ M + N
models[[7]] = A ~ S + N
models[[8]] = A ~ M + S + N
models[[9]] = A ~ M * S + N
models[[10]] = A ~ M + S * N
models[[11]] = A ~ M  * N + S
models[[12]] = A ~ M  * N + M * S
models[[13]] = A ~ M  * N + N * S
```

```
models[[14]] = A ~ M * S + N * S

models[[15]] = A ~ M * S + N * S + M * N

models[[16]] = A ~ M * N * S
```

The results of this analysis are shown in Figure 5. There are four competing models: 1) the model (numbered 3) with only a main effect from national identification (33% of the wins); 2) model number 7 with main effects of self esteem and national identification (10 % of the wins); 3) model number 9 with a two way interaction between self-esteem and mortality salience and a main effect of national identification (16% of the wins); and 4) the model including the three way interaction (42% of the wins). Although the latter model wins in 84 of the 200 repetitions the average prediction error for this model is 0.667, whereas that of the model with only a main effect is 0.665, a little bit smaller. The other two competing models have a root mean squared error of prediction of 0.667. Therefore, in terms of prediction all these model do equally well. The most complicated model has the widest prediction interval, running from 0.653 (2.5%) till 0.685 (97.5%) (see supplementary material), which is also clearly visible in Figure 5.

In this case we see that cross validation provides the researcher with quite some information. Although the model with the three-way interaction most often wins the gain in prediction accuracy is very small. A model with only a main effect of national identification performs equally well and might, just because of its simplicity, be the preferred model.

In this case researchers could also use a standard analysis approach fitting a series of regression models and looking at the change in explained variance and corresponding test statistics. In the online supplementary material this analysis is shown. One problematic aspect of this procedure is that not all models are nested and therefore cannot be compared with statistical tests. Furthermore, many p-values are computed which raises the question how to correct for multiple comparisons. The same issues arise as in stepwise procedures. Finally, this analysis requires distributional assumptions for the residuals. The diagnostic plots question the validity

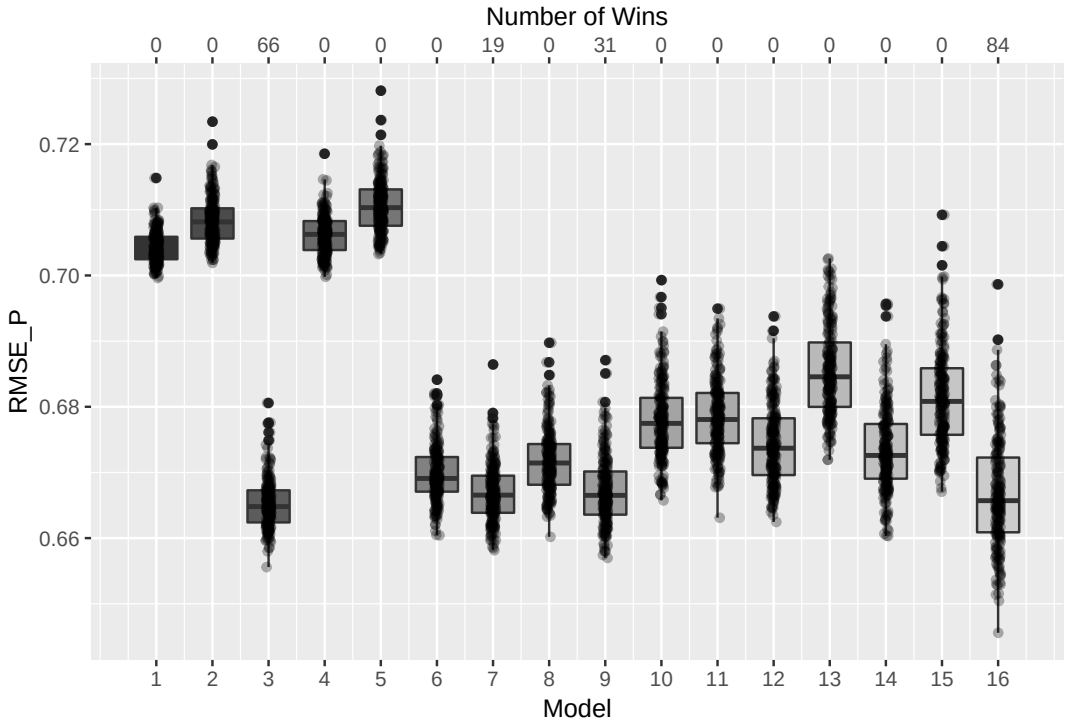

Figure 5: Cross validation results for Attitude towards Muslim and Multiculturalism data. Fifteen different models were formulated.

of these assumptions.

3.4 Logistic Regression Situation

Hastie and Tibshirani (1990) report data on the presence or absence of kyphosis, a postoperative spinal deformity. The predictor variable is the age of the patient in months. For these data we model the relationship between kyphosis and age using a logistic regression. Therefore, let $\pi(x_i) = P(Y_i = 1|x_i)$ denote the conditional probability of having kyphosis given age (X). The logistic regression model with a linear effect of age on the log-odds of kyphosis can be written as

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = a + x_i b.$$

We compare the predictive power of this model to a model including only the intercept, i.e. a

model in which age does not predict kyphosis, and a model with also a quadratic term

$$\log\left(\frac{\pi(x_i)}{1-\pi(x_i)}\right) = a + x_i b_1 + x_i^2 b_2.$$

We repeatedly estimate the model parameters in the calibration sets and then make predictions in the validation sets. These predictions are in the form of probabilities $\hat{\pi}(x_j)$. Using this prediction in our root mean squared error of prediction loss function gives

$$RMSE_P = \sqrt{\frac{1}{N_{\mathcal{V}}}\sum_{j=1}^{N_{\mathcal{V}}}(y_j - \hat{\pi}(x_j))^2},$$

with $y_j \in \{0, 1\}$, which is known as the Brier score.

The data are available in the `gam`-package, where the response variable is a string variable. We first recode the response variable to a 0,1 variable, where 1 indicates presence of kyphosis. Then we fit three logistic regression models, an intercept only model, a logistic regression with Age as predictor, and a logistic regression with Age and Age-squared as predictors. Calling the cross validation function for a logistic regression can be done as follows

```
library(gam)
data(kyphosis)
kyphosis[,1] = as.numeric(kyphosis[,1] == "present")
models = vector(mode = "list", length = 3)
models[[1]] = Kyphosis ~ 1
models[[2]] = Kyphosis ~ Age
models[[3]] = Kyphosis ~ poly(Age,2)
output = xval.glm(data = kyphosis, models, glm.family = binomial)
```

The results of the cross validation are shown in the left hand side of Figure 6, where it can

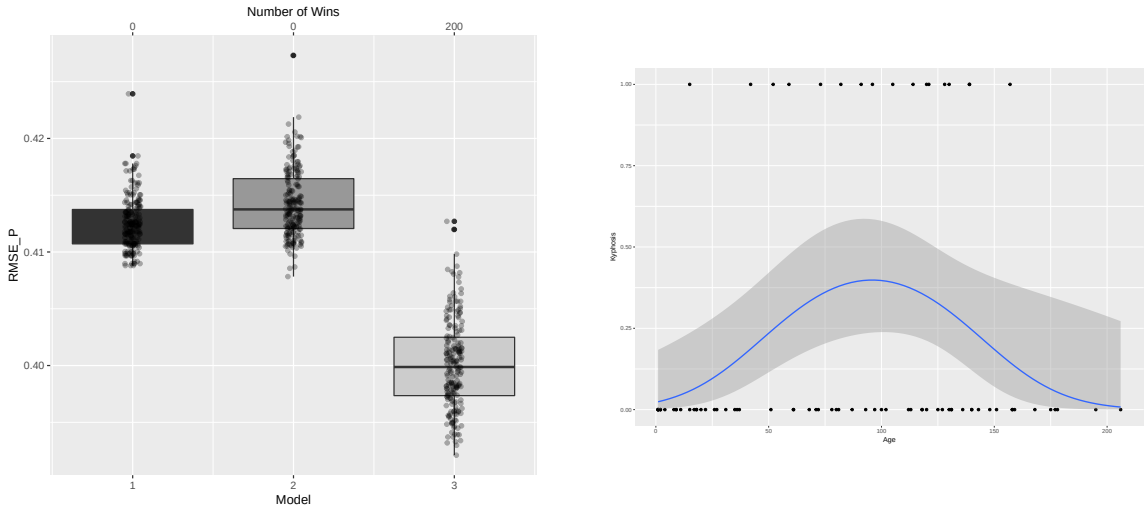

Figure 6: The left hand side shows the cross validation results of three logistic models predicting kyphosis from age: Model 1 makes predictions based on an intercept only model; Model 2 includes age as a predictor; Model 3 is based on a quadratic function of age. On the right hand side the best predicting model is shown, fitted on all data.

be seen that the quadratic polynomial of age gives the best predictions. The right hand side of Figure 6 shows the quadratic fitted model on all data. We see that this is a single peaked curve, first the probability goes up, later it goes down. Nowhere the probability becomes larger than 0.5, so for every person we would predict the absence of kyphosis. Around the age of 100 months there is a probability of about 40 percent that a patient has kyphosis.

3.5 Changing the Loss function

The default way to compute prediction error in our function is by the root mean squared error of prediction. As seen in the previous subsection, for logistic regression this equals the Brier score. In some situations one might be interested in another loss function. For linear models, for example, we might be interested in the average absolute error. For logistic regression one might be interested in the misclassification rate, or the cross validated deviance.

If another loss function should be employed, we first have to define it. Therefore a function is needed with two arguments, the observed responses (`y`) and the predictions (`preds`). The function for the average absolute loss is defined in R as

```
absloss = function(y,preds){mean(abs(y - preds))}
```

With this function we can run the four models from Section 3.2 again with the adapted call

```
output = xval.glm(data = agdat, models, loss = absloss)
```

so that now prediction error is defined as the average absolute error. The results of these four models is shown in Figure 7, where it can be seen that the quadratic model still provides the best results, although in this case the intercept only and linear models are closer to the winner.

In the online appendix we also show an example where we change the loss function in logistic regression from the Brier score to the cross validated deviance and the misclassification rate. The former gives similar results as the Brier score, whereas the misclassification rate completely changes the conclusions, showing the importance of the loss function for our results. The misclassification rate is insensitive to differences in probabilities, whereas the Brier score and the deviance are sensitive to differences. From the analysis using the misclassification rate we would conclude that age does not have an influence on the presence of kyphosis and we predict for every child that kyphosis is not present. From the analysis using the Brier score as loss function we can still conclude that for every child the probability of kyphosis is smaller than 0.5, but we can also conclude that for children around the age of 100 40% of the treated children will develop kyphosis. The main question for the surgeon then is whether this is an acceptable risk or not.

Another consideration in the choice of loss function is the goal of the analysis. Is the goal to classify new patients or is it to obtain better insight (theory) about the relationships between the variables. In the first case, we should use the misclassification rate as measure of predictive performance; in the second case we better use a more sensitive measure.

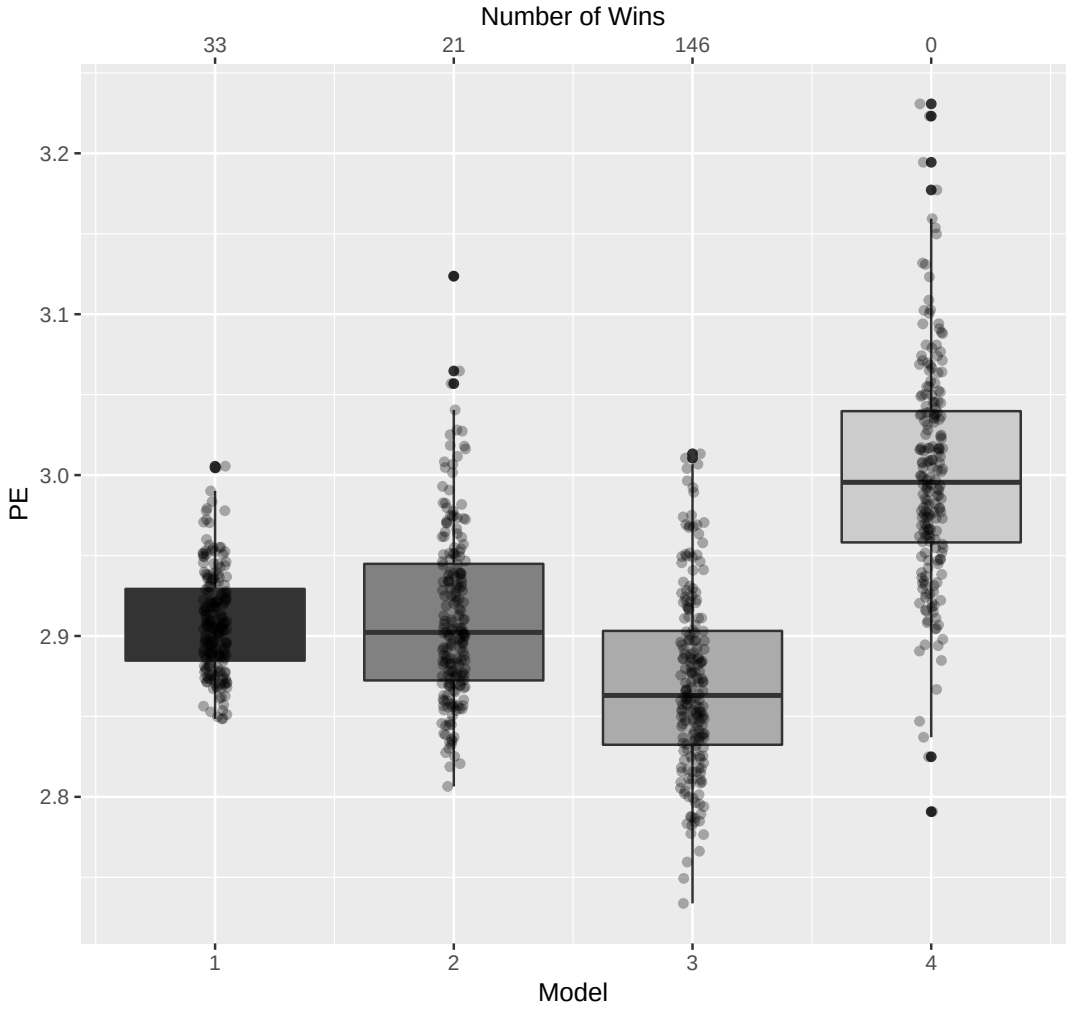

Figure 7: Cross validation results for Aggression data with mean absolute error as loss function. Model 1 makes predictions on the overall mean; Model 2 makes predictions based on the group means. The vertical axis represents root mean squared error of prediction. The number of wins are represented at the top of the window.

3.6 Comparing the predictive power of two theories

In Pollack et al. (2012) the authors investigate the effect of economic stress on intentions to disengage from entrepreneurial activities. The participants in this study were 262 entrepreneurs who were members of a networking group for small-business owners, who responded to an online survey about recent performance of their business, and their emotional and cognitive responses to the economic climate.

The participants were asked a series of questions about how they felt their business was doing. Their responses were used to create an index of economic stress (`estress`, higher scores reflecting greater stress). They were also asked the extent to which they had various feelings related to their business, such as feeling discouraged, hopeless, worthless, and the like, an aggregation of which was used to quantify business-related depressed affect (`affect`, higher scores reflecting more depressed affect). Another measure is entrepreneurial self-efficacy. This measure indexes a person's confidence in his or her ability to successfully engage in various entrepreneurship-related tasks such as setting and meeting goals, creating new products, managing risk, and making decisions (`ese`). Finally, they were also asked a set of questions to quantify their intentions to withdraw from entrepreneurship in the next year (`withdraw`, higher scores are indicative of greater withdrawal intentions). Moreover, we have a set of covariates: `sex` (0 = female; 1 = male), `age` in years, and `tenure` (length of time in business).

For these data there are two theories:

1. The first theory says that economic stress has an influence on withdrawal intentions but that this effect is mediated by business-related depressed affect. Taking the covariates into account this leads to a regression model with `withdraw` as response and `estress`, `affect`, `sex`, `age`, and `tenure` as predictors.

2. A second theory is that economic stress is not at all related to withdrawal intentions and that withdrawal intentions are just an effect of individual differences. That is, more confi-

dent persons have less depressed affect and therefore less intentions to withdraw. Taking the covariates into account this leads to a regression model with `withdraw` as response and `ese`, `affect`, `sex`, `age`, and `tenure` as predictors.

Important to note is that the two theories lead to two regression models that are not nested. That is, these models are hard to compare using statistical tests like, for example, the likelihood ratio test. It is, however, quite easy to compare the two theories using cross validation. This can be done using the following code:

```
library(foreign)
ecdata = read.spss("estress.sav", to.data.frame = TRUE)
models = vector(mode = "list", length = 2)
models[[1]] = withdraw ~ tenure + estress + affect + sex + age
models[[2]] = withdraw ~ tenure + affect + sex + age + ese
output = xval.glm(data = ecdata, models)
```

The results are shown in Figure 8 which shows that the second theory leads to more accurate predictions (i.e. the prediction error is lower).

4 Discussion

Lately, there has been an increased interest in an old methodology: cross validation. The importance of cross validation has already been recognised for psychological research quite early (Mosier, 1951). Nevertheless, cross validation is not often used in psychology possibly because no simple software tools are available. We developed such a tool in the open software R. The package can be downloaded from the internet. The function uses K-fold cross validation and uses many (200) repetitions in order to compare a set of models. Most often the models differ in one variable (parameter) and the question that is answered is than: does this extra variable

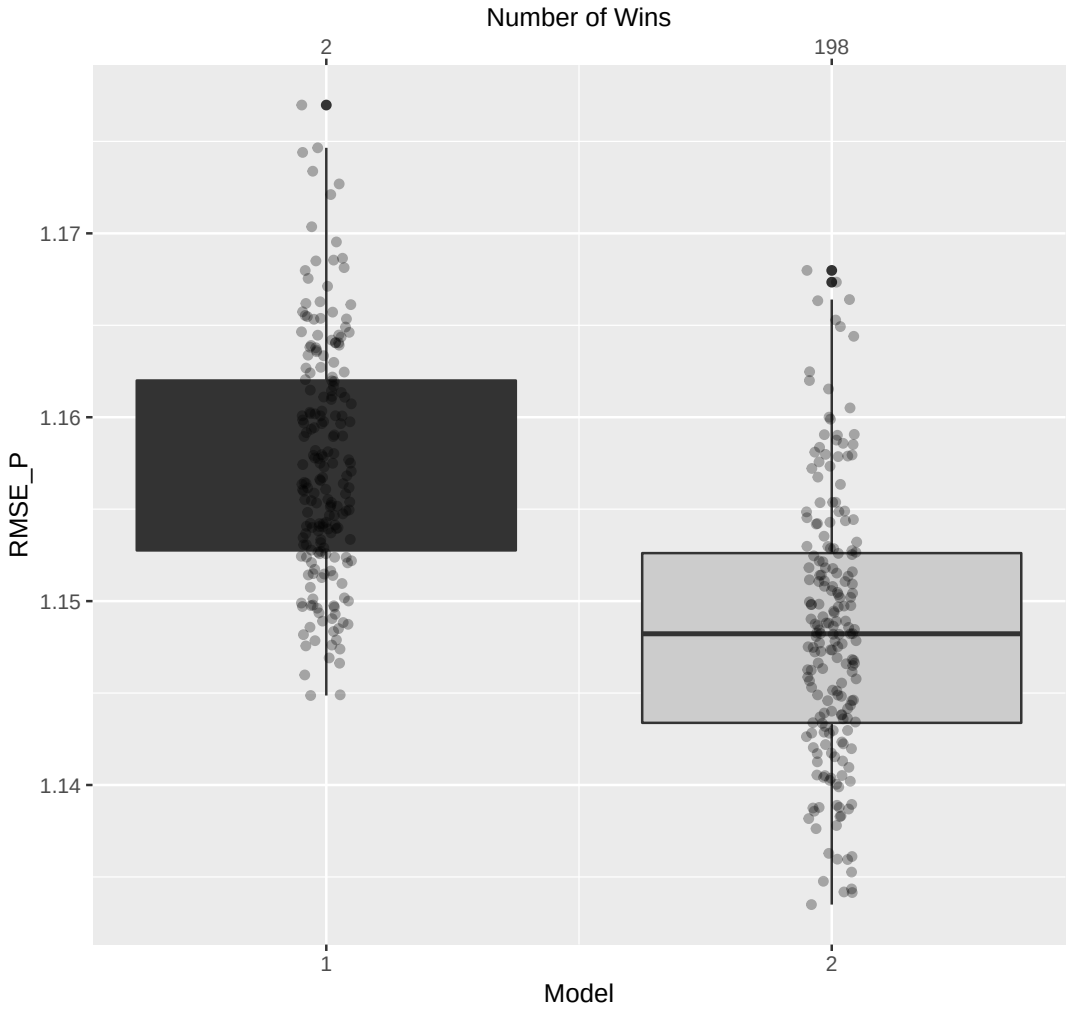

Figure 8: Cross validation results to compare the predictive power of two theories regarding economic stress. Theory 1 corresponds to model 1, theory 2 to model 2. Theory 2 is more accurate.

lead to better predictions? The software is built around the family of generalized linear models, which encompass many different analysis methods often used in psychology. The main function (`xval.glm`) is easily adapted to any other statistical method implemented in R that has an estimation and a prediction function.

We showed several tutorial examples of cross validation for types of analyses often encountered in psychology. We hope that researchers can adapt the code for their own analyses. All analyses focussed on model selection. Therefore, several models of interest are defined and compared in a repeated K-fold cross validation regime. The best model is selected and fitted to all data. The result might be used as a prediction tool for the future. Due to the repetitions we obtain a distribution of the prediction errors. Note that because we select the optimal model there is a risk that these future predictions are a bit worse compared to the distribution that is obtained in the analysis. This is simply due to regression to the mean, a phenomenon often observed. If an honest measure of prediction error is needed, besides our procedure we need another data set or independent part of the data. The model selection is then performed as described in the examples, the best model is fitted to the complete data and afterwards a good measure of prediction error is obtained from this independent data set. This would amount to doing repeated K-fold cross validation and testing the selected model on an independent holdout set.

The function returns 1) boxplots of the prediction error of the different models under consideration; 2) The number of wins of each of the models in the 200 repetitions; 3) A density curve of the prediction errors for the different models. In most of our examples we had clear winners, i.e. the distribution of prediction errors does not show much overlap and the percentage of wins is largely in favor of one of the models. In our moderated regression example results are not that clear, and four competing models can be identified. Experience shows that even if one model wins all the time, the distribution of prediction errors may have large overlap, suggesting that for every repeat one model performs a little better than the other ones. In such

a case the percentage of wins might be a misleading indicator and a researcher has to take into account the gain in prediction accuracy. To do so, it is important to relate the measure of prediction error to knowledge about the dependent variable. Is a gain of, for example, 0.5 meaningful in relation to the distribution of the response variable and would such a gain be noticeable in practice? Sometimes there is quite some overlap in the distributions and the winning percentages might be, for example, 55% and 45%. This means that the data cannot really distinguish between the two models. One choice in such a case is to favor the most parsimonious model, i.e. concluding there is not enough evidence that the extra effect provides incremental validity. Another choice is to conclude that we do not have enough information to make a choice between the two models. In choosing between models we might take into account how easy it is to obtain the extra information: If it is just a matter of asking a single question it might be worthwhile to use this information; on the other hand, if it is a really expensive test in terms of time or money the researcher may consider not to obtain this information in order to get a little better predictions.

Cross validation is a resampling technique. Another resampling technique is the bootstrap (Efron and Tibshirani, 1993). The bootstrap is often used to obtain confidence intervals of the parameters of a statistical model, but can also be used to assess predictive performance. Two approaches can be distinguished. The first approach uses bootstrap samples to fit the statistical models, i.e. the bootstrap sample is used as the calibration sample, and uses the observations not in the bootstrap sample (often called the out-of-bag observations) as validation set. The advantage of this procedure is that the calibration samples are as large as the original sample. A disadvantage is that only 63.2% of the observations of the original sample are in the bootstrap sample therefore creating bias. The second way was developed to deal with this bias and was called the .632 bootstrap (Efron, 1983; Efron and Tibshirani, 1997). Like before it uses bootstrap samples to calibrate the statistical models. The validation is subsequently done on the bootstrap sample as well as on the out-of-bag observations. This gives two estimates of prediction error;

the difference between these estimates indicates the optimism of the in-sample approach. The final prediction error is computed as a weighted average of the two measures of prediction error.

A topic we did not go into is models with tuning parameters, as they occur in modern regression models like the lasso (Tibshirani, 1996). If we have such a model we should find 1) an optimal value for the penalty parameter and 2) for the regression model with the optimal penalty parameter we would like to know the prediction error. In such a case we need *nested cross validation* (Varma and Simon, 2006), in which we do K-fold cross validation within K-fold cross validation. In more detail, we split the data in K parts, select one to be the test set and the others as training set. In the training set, we again do K-fold cross validation in which we fit the whole series of models for every possible value of the penalty parameter and select the value that gives the smallest prediction error. With this value we make predictions in the test set of the outer loop.

There are many ways to do cross validation. We focussed on repeated K-fold cross validation that includes leave one out cross validation. Above we briefly discussed relationships with the bootstrap and nested cross validation. There are other related forms of cross validation; brief descriptions can be found in Steyerberg (2009), Arlot and Celisse (2010), Kuhn and Johnson (2013), and Krstajic et al. (2014). Krstajic et al. (2014) provide an overview of pitfalls and benefits of different cross validation strategies. Kohavi (2014) compares cross validation with the bootstrap.

As concluding remark we like to mention that cross validation should be a standard procedure in data analysis, either as model selection procedure as discussed in this paper or as validation of an otherwise selected model. By focussing on out of sample performance the chances that an obtained results is replicable increase (Yarkoni and Westfall, 2017), or as Bokhari and Hubert (2018) point out: "the lack of cross validation can lead to inflated results and spurious conclusions".

Acknowledgement

The authors would like to thank two anonymous reviewers for their constructive reviews.

Author Contributions

MdR initiated an R-function and developed several examples. WW further extended the R-function and developed the code into an R-package. MdR wrote the first draft of the paper after which it iterated between WW and MdR. MdR finalized the manuscript. MdR and WW discussed the reviews and adapted parts of the manuscript and the supplementary material.

References

Adams, H., Wright, L., and Lohr, B. (1996). Is homophobia associated with homosexual arousal? *Journal of Abnormal Psychology*, 105:440 – 445.

Alexander, D. L. J., Tropsha, A., and Winkler, D. A. (2015). Beware of r^2: simple, unambiguous assessment of the prediction accuracy of qsar and qspr models. *Journal of Chemical Information and Modeling*, 55:1316–1322.

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4:40–79.

Bokhari, E. and Hubert, L. (2018). The lack of cross validation can lead to inflated results and spurious conclusions: A re-analysis of the macarthur violence risk assessment study. *Journal of Classification*, 35:147 – 171.

Breiman, L. (2001). Statistical modeling: the two cultures. *Statistical Science*, 16:199 – 231.

Browne, M. W. (2000). Cross validation methods. *Journal of Mathematical Psychology*, 44:108–132.

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

Chow, S. L. (1998). Precis of statistical significance: rationale, validity, and utility. *The behavioral and brain sciences*, 21:169–239.

Claudy, J. (1978). Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 2:595– 607.

Cohen, J. (1994). The earth is round ($p < .05$). *The American psychologist*, 49:997–1003.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25:7 – 29.

Darlington, R. (1968). Multiple regression in psychological research and practice. *Psychological Bulletin*, 69:161– 182.

Darlington, R. (1978). Reduced-variance regression. *Psychological Bulletin*, 85:1238 – 1255.

Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331.

Efron, B. and Tibshirani, R. (1993). *An introduction to the bootstrap*. Chapman and Hall, New York.

Efron, B. and Tibshirani, R. (1997). Improvements on cross-validation: The 632+ bootstrap method. *Journal of the American Statistical Association*, 92(438):548–560.

Fox, J. (2016). *Applied regression analysis & generalized linear models*. Sage Publications, Inc., Thousand Oaks, California.

Geisser, S. (1975). The predictive sample reuse method with applications. *Journal of the American Statistical Association*, 70:320–328.

Hagen, R. L. (1997). In praise of the null-hypothesis statistical test. *The American psychologist*, 52:15–42.

Hagerty, M. R. and Srinivasan, S. (1991). Comparing the predictive powers of alternative multiple regression models. *Psychometrika*, 56:77 – 85.

Harrell, F. (2015). *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*. Springer, New York.

Hastie, T., Friedman, J., and Tibshirani, R. (2001). The elements of statistical learning. *Springer Series in Statistics*.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, New York.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning, 2nd edition*. Springer, New York.

Howell, D. C. (2015). *Statistical Methods for Psychology*. Cengage, UK.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer-Verlag, New York.

Kohavi, R. (2014). A study of cross-validation and bootstrap for accuracy estimation and model selection. *International Joint Conference on Artificial Intelligence*, 6:14.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Chemometrics*, 6:10.

Krueger, J. (2001). Null-hypothesis significance testing: On the survival of a flawed method. *The American psychologist*, 56:16–26.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer, New York.

Lawshe, C. and Schucker, R. (1959). The relative efficiency of four test weighting methods in multiple prediction. *Educational and Psychological Measurement*, 19:103 – 114.

Matloff, N. (2017). *Statistical Regression and Classification: From Linear Models to Machine Learning*. CRC Press, Taylor & Francis Group, Boca Raton, FL.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioural sciences. *Multivariate Behavioural Research*, 50:471 – 484.

Mosier, M. W. (1951). I. problems and design of cross-validation. *Educational and Psychological Measurement*, 11:5–11.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods*, 5:241–301.

Pollack, J. M., VanEpps, E. M., and Hayes, A. F. (2012). The moderating effects of social ties on entrepreneurs' depressed affect and withdrawal intentions in response to economic stress. *Journal of Organizational Behavior*, 33:789 – 810.

Pruzek, R. and Frederick, B. (1978). Weighting predictors in linear models: Alternatives to least squares and limitations of equal weights. *Psychological Bulletin*, 85:254 – 266.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Hartig, F., and Dormann, C. F. (2016). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, 40(8):913–929.

Rozeboom, W. (1968). Estimation of cross-validated multiple correlation: a clarification. *Psychological Bulletin*, 69:1348– 1351.

Rozeboom, W. (1979). Ridge regression: Bonanza or beguilement. *Psychological Bulletin*, 86:242 – 249.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin*, 57:416–428.

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

Shmueli, G. (2010). To explain or to predict. *Statistical Science*, 25:289 – 310.

Steyerberg, E. W. (2009). *Clinical Prediction Models: A practical approach to development, validation, and updating*. Springer, New York.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147.

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

Tjew-A-Sin, M. and Koole, S. L. (2018a). Data from paper 'terror management in a multicultural society: Effects of mortality salience on attitudes to multiculturalism are moderated by national identification and self-esteem among native dutch people'. *Journal of Open Psychology Data*, 6:5.

Tjew-A-Sin, M. and Koole, S. L. (2018b). Terror management in a multicultural society: Effects of mortality salience on attitudes to multiculturalism are moderated by national identification and self-esteem among native dutch people. *Frontiers in Psychology*, 9:721.

Varma, S. and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91.

Wagenmakers, E. (2007). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin & Review*, 14:779–804.

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

Wilcox, R. R. (2017). *Understanding and Applying basic statistical methods using R*. John Wiley & Sons, Hoboken, New Jersey.

Yarkoni, T. and Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12:1100 – 1122.