Working Paper

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Workshop on Statistical Data Editing
(31st August - 4th September 2020)

Two-Phase Learning

Prepared by Tatsiana Pekarskaya and Li Chun Zhang, Statistics Norway

I. INTRODUCTION

1. Machine learning (ML) models are increasingly more often used in statistical production for different purposes, one of which is editing and imputation of data. To solve these type of problems one usually applies methods of supervised ML. The main goal is to get accurate predictions; less attention is given to uncertainty assessment.

2. Getting a *valid* model means to get statistically correct predictions $E(Y - \hat{Y}|x) = 0$, i.e. errors average to zero when the model covariates are held fixed. However the errors of a statistically correct prediction can have *heterogeneous variance* across the different individuals, or have *heterogeneous mean* that invalidates prediction at more disaggregated levels. Both types of heterogeneity are largely inevitable for any statistical modelling, when the fitted model is applied at a lower lever than that is accounted for by the chosen model covariates. Since prediction of any kind is not of interest in the sample available for ML, but for the rest of the population, standard summary performance measures obtained from the hold-out sample generally do not provide adequate uncertainty assessment.

3. Uncertainty concerns prediction errors. Residuals are estimates of errors for the observations used to fit the model. Insofar as any model tends to over-fit the data used, the residuals tend to understate the uncertainty of the errors. The split-data approach that is standard in ML has an advantage in this respect, compared to model-fitting based on the whole sample.

4. *Two-Phase Learning (TPL)* is a combination of supervised ML applied in two phases. In the 1st-phase, one aims to learn to predict the dependent variable itself; in the 2nd-phase, one aims to learn the predict errors of the 1st-phase model, in a manner that hopefully allows one to capture both types of heterogeneity. When the results of TPL is applied to the rest of population, one aims:

a) to provide appropriate individual-level description of the prediction uncertainty,
b) to possibly adjust the 1st-phase model prediction on-the-flight.

In principle TPL is relevant and applicable to any ML models.

5. One can notice a similarity of TPL to gradient boosting, in terms of adjusting an initial prediction. However, gradient boosting does not deal with the goals TPL in terms of heterogeneous uncertainty assessment. Moreover, it does not use additional covariates as generally required for TPL, and it learns from the residuals not errors.

6. Further will be described the TPL procedure and provided an example of its application, to a problem of imputation of working time in agriculture survey.

II. TPL procedure

1. The procedure of TPL is similar to supervised ML with hold-out set. The difference is that the procedure should be applied two times, following the algorithm below.

i) Make data-splitting.
ii) Run ML in the 1st-phase, to obtain a prediction model of the variable of interest:

$$Y = \hat{Y} + \epsilon \qquad \text{and} \qquad \hat{Y} = f(y|x)$$

f - 1st-phase model, Y - variable to be predicted, x - explanatory variables, ϵ - 1st-phase error.
iii) Run ML in the 2nd-phase, to obtain prediction models of functions of the 1st-phase error:

$$h(\epsilon) = \hat{h}(\epsilon) + \xi \qquad \text{and} \qquad \hat{h}(\epsilon) = g\big(h(\epsilon)|x, z\big)$$

g - 2nd-phase model, z - additional features, ξ - 2nd-phase error.
Here, $h(\epsilon)$ can be $h(\epsilon) = \epsilon$, or functions of ϵ such as sign(ϵ), ϵ^2.
iv) Uncertainty description of the 1st-phase model based on $\hat{h}(\epsilon)$, or on-the-flight adjustment of the 1st-phase prediction given as $\hat{\hat{Y}} = \hat{Y} + \hat{\epsilon}$ using the 2nd-phase model for $h(\epsilon) = \epsilon$.

2. The starting point is data-splitting. For each phase of the learning one needs one training set and one hold-out set. Thus, for the whole TPL one splits the data in three sets:

a) *Training set*, which is used for training the 1st-phase model and gives us *1s-phase residuals*.
b) *Testing set*, which is a hold-out set for the 1st-phase model and gives us the *1st-phase errors*. The errors are used to train the 2nd-phase model and give us the *2nd-phase residuals*.
c) *Final testing set* is used to choose the best 2nd-phase ML model, yielding the *2nd-phase errors*.

3. Depending on the ML model, one may need to further separate out an additional *validation set* both in the training set and the testing set, or alternatively apply k-fold cross-validation, for model tuning, such as in the case of random forest though not in the case of linear regression.

III. An example

1. We use the data of agriculture census conducted in 2010 and an agriculture sample from 2009. As the Y variable we use *total working time* on each farm, related to all agriculture activity in year 2010. The extreme outliers of working time are removed. In reality all of them are surveyed and do not need prediction anyway. We shall assume that working time is only observed in the sample, and we would like to impute it for the rest of the population.

2. Auxiliary data are available from a relevant administrative register, including type of farming activities, size of area used for cultivation, number of different kinds of animals, turnovers after farming activities with animals and plants, area used for cultivation of some kinds of crops, county and organisation form. The last two are kept out of the 1st-phase as additional features (z) for the 2nd-phase for illustrative purpose, the rest (x) are available to both phases.

3. Scatter plots of Y and some continuous x can be found in Figure 1. These show that the data are noisy and follow mixed distributions, e.g. contrasting those near $x = 0$ and the rest. We shall not apply any special methods here, as our primary aim is to illustrate how the 2nd-phase learning works, rather than obtaining a 1st-phase predictor that is as good as possible.

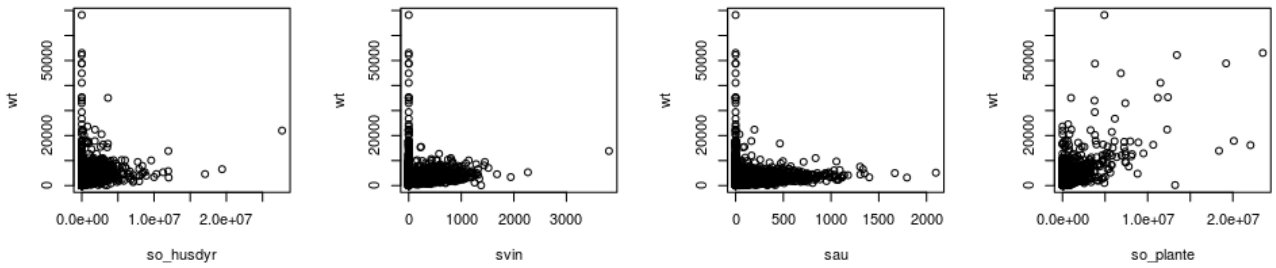

FIGURE 1. Scatter plots of Y and some exploratory continuous x. In the plot svin, sau - number of different kinds of animals; so_husdyr and so_plante - turnover from farming activities with animals and plants; wt - working time

A. **Data-splitting**

1. In our case here we can simply use for final testing all the units out of the sample, i.e. the rest of population, the size of which is 35604, whereas the sample size is 8866. Training and testing sets are each 50% of the sample by random selection. For tuning of model parameters we apply bootstrap cross validation with 30% split for validation set.

2. As can be seen from Table 1, the sample Y-average is higher than the rest of population, due to higher sample inclusion probabilities of larger units. It means that we do not exactly have a balanced split between the sample and the rest of population, which potentially may be a concern if the heterogeneity of errors are not appropriately taken into account.

TABLE 1. Averages of Y after data-splitting

Sample	2671.780	Training	2675.818
		Testing	2667.743
Rest of population	1885.602	Final testing	1885.602

3. The training and testing sets appear to be reasonably balanced, judging from the averages and histograms of Y in Table 1 and Figure 2. Of course, if needed, it is possible to exercise a greater degree of control for balance when splitting the data in the sample.

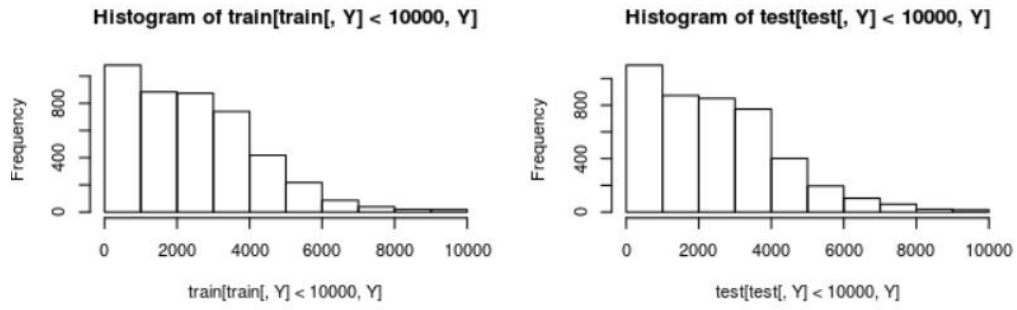

FIGURE 2. Histogram of Y for training and testing sets. Max Y-value shown is 10000

B. First-phase learning

1. As 1st-phase models we consider: sample mean, linear regression, decision tree, random forest and gradient boosting. We use the R package `caret` and its auto-tuning of model parameters wherever it is needed. Table 2 gives the individual-level root mean squared error (RMSE) obtained in the testing set. Random forest gives the best results, and its errors will be used for the 2nd-phase learning.

TABLE 2. Summary of errors of 1st-phase models in testing set

Model	Sample Mean	Linear regression	Tree	Random Forest	Gradient Boosting
Mean	8.075	-30.224	-29.760	-2.073	-21.462
RMSE	2611.612	1798.742	2141.117	1786.287	1871.054

2. Table 3 gives mean, RMSE and probability of positive sign (PPS) of the 1st-phase random forest prediction error in the testing set (left) and rest of population (right). The metrics are calculate overall (the first column) and in groups by organisation form DA, ENK and OTHER.

a) For comparisons of these results, one can easily gauge the statistical significance. For instance, the mean -2.703 overall in the testing set has a standard error 26.825, which is significantly different from -157.587 in the rest of population. Note that -2.703 is not significantly different from 0 (in the training set), which is not surprising given the balance between the two.
b) Without labouring the details, it is e.g. clear that the results for group OTHER is quite different from the other two groups, both in the training set and the rest of population.
c) In either the testing set or the rest of population: the results of mean across different groups illustrate mean heterogeneity of prediction error, and those of RMSE illustrate variance heterogeneity, whereas the deviation of PPS from 0.5 indicates both types of heterogeneity within each group.
d) The differences of results between the testing set and the rest of population appear somewhat smaller than those between the different groups within either set. Nevertheless, direct extrapolations of the summary metrics from the testing set to the rest of population can be invalid in many instances.

TABLE 3. Random forest prediction errors.

	Testing set				*Rest of Population*			
	Overall	DA	ENK	OTHER	Overall	DA	ENK	OTHER
Mean	-2.073	15.194	-25.805	717.738	-157.587	-125.796	-173.411	701.6334
RMSE	1786.287	1667.939	1563.843	5406.214	1309.380	1256.090	1188.181	4407.440
PPS	0.400	0.406	0.399	0.424	0.326	0.390	0.321	0.465

3. With these illustrative results in mind, one can now appreciate the basic shortcoming of an undiscerning approach to uncertainty assessment by reflecting on the following question.

> *Suppose one applies the 1st-phase random forest predictor to obtain $\hat{Y}$ for a given unit in the rest of population. Suppose one reports the overall mean, RMSE and PPS in the testing set, i.e. (-2.073, 1786.287, 0.400), as the associated uncertainty regardlessly. Do these metrics seem appropriate in light of the other results in Table 3?*

This is what we mean that, by using the 2nd-phase learning, one aims to provide appropriate individual-level description of the prediction uncertainty. Finally, without the split between training and testing sets, the error metrics based on fitting a model to the whole sample would be 0 for mean, a number estimated from the sample residuals for RMSE, and 0.5 for PPS by assumed normality of errors, which obviously are even worse than the overall metrics obtained in the testing set.

C. Second-phase learning

1. As the outcome variable $h(\epsilon)$ for 2nd-phase learning, we shall consider ϵ, $\text{sign}(\epsilon)$ and ϵ^2. Denote by x the features selected for the 1st-phase model and by z the unselected features. For illustration purpose, we have kept organisation form and county away from the 1st-phase learning, so that we easily have these as part of z for the 2nd-phase. Even in situations where there are no hold-out features like this at the 1st-phase, additional features for 2nd-phase learning can generally include the unselected features at the 1st-phase, and transformations of the features that are selected at the 1st-phase. Effects of feature selection for 2nd-phase learning will be illustrative below.

C.1. *Prediction of ϵ.*

1. As usual for model building one should start with *exploratory data analysis*. In Figure 3 one can find two plots of 1st-phase model errors in the testing set. The left plot illustrates that a feature used in 1st-phase can still be useful for predicting errors at the 2nd-phase, e.g. by categorical split: $< 4.0e+6$ - more positive errors, $> 4.0e+6$ - more negative errors with greater spread. This is because the 1st-phase model can never be exactly correct. In the right plot, one can see that the maximum predicted value is much lower than the true maximum value. Meanwhile, the error plot of the 1st-phase *linear regression model* in Figure 4 does not have this problem. This suggests generally (a) different models can be considered for the 2nd-phase learning, and (b) the predicted values by different 1st-phase models, or functions of them, can be considered as additional features for the 2nd-phase learning.

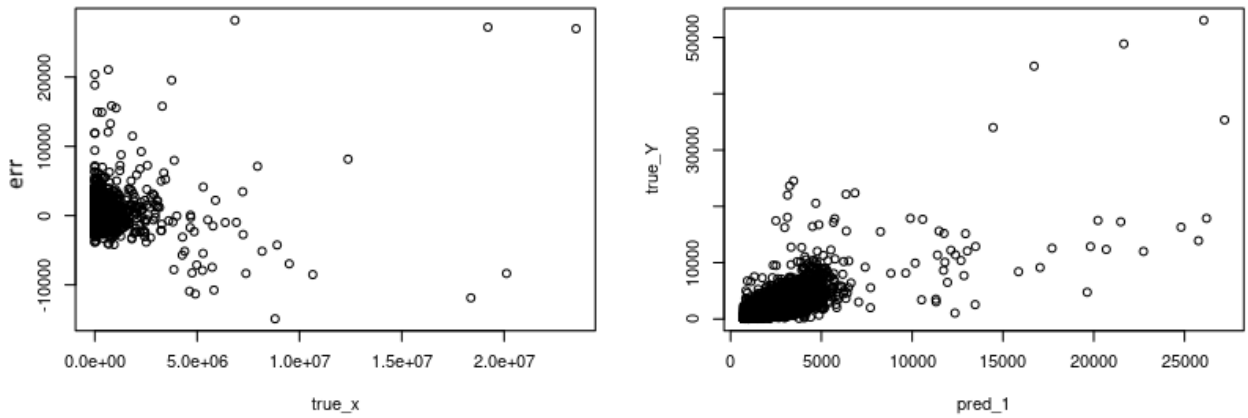

FIGURE 3. Plots for random forest prediction errors in testing set: true_x - turnover of farming activities with plants, err - errors, pred_1 - predicted Y, true_Y - true Y.

2. We illustrate feature selection for 2nd-phase learning with three settings.

A) Use only 1st-phase features x.
B) Use only the hold-out features (from 1st-phase) as additional features z.
C) Use hold-out features and transformations of the 1st-phase features (incl. $\hat{Y}$) as z.

For each setting, several models are considered for the 2nd-phase learning. For simplicity, we present in Table 4 only the results of the random forest model. The left block shows the overall mean of the predicted errors $\hat{\epsilon}$ and in the groups by organisation form, under different settings of feature selection. These can be compared to the mean of the true errors ϵ in Table 3 earlier. The right clock shows the

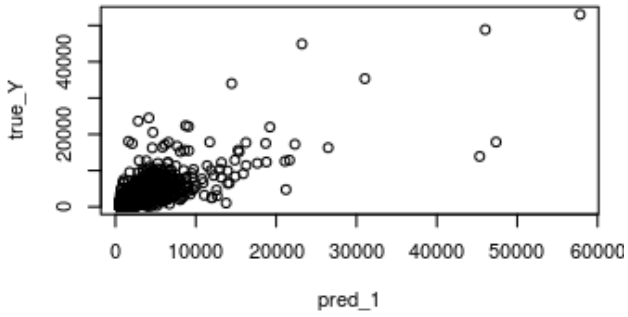

FIGURE 4. Random forest prediction errors in testing set: pred_1 - predicted Y, true_Y - true Y.

model diagnostics, where R^2 and root mean squared residual ($\hat{\xi}^2$) are calculated in the testing set, and RMSE (ξ^2) is calculated on the final testing set (i.e. rest of population here).

TABLE 4. Random forest 2nd-phase error prediction for rest of population

Feature	Mean($\hat{\epsilon}$)				Model Diagnostics		
Setting	Overall	DA	ENK	OTHER	R^2	RMSR	RMSE
A	-115.860	23.944	-118.323	-212.950	0.096	656.419	1255.236
B	-141.038	-25.570	-147.495	34.204	0.108	1293.086	807.916
C.1	-138.603	-11.210	-142.628	-124.247	0.591	469.837	1365.111
C.2	-139.147	-71.477	-148.551	288.031	0.134	1300.835	1282.672

3. In terms of the overall mean of the prediction error, all the 2nd-phase error prediction results are better than direct extrapolation of the corresponding metric -2.073 in the testing set, where the true mean of the 1st-phase model error for the rest of population is -157.587 (Table 3). For grouped means of the prediction errors, the improvement is clear for ENK, which comprises of about 88% of the population units, where the true 1st-phase model error in the rest of population is -173.411 (Table 3). The grouped means are not close to the truth for the other two smaller groups DA and OTHER. This is largely caused by the imbalance of these groups, so that the model is automatically tuned towards better-fitting for ENK at the cost of the other two groups. Of course, implementing separate models in each group would have resolved this problem. But since the same issue would then arise for other disaggregations that are not controlled for in the model, it is a general issue that does not have a simple fix. In practice, therefore, one would need to prioritise *disaggregated* 2nd-phase modelling according to the needs of dissemination, and strive for a sensible compromise.

4. When it comes to the different settings of feature selection, incorporating additional features in different ways generally improves the results based only on the same 1st-phase features (setting A). The difference between the two alternatives of setting C is worth noting.

a) In setting C.1, we use many additional features, including organisation form, county, Y predictions of the 1st-phase models, differences in predictions of the 1st-phase models, together with their absolute values, err predictions of the 2nd-phase models without additional features, categorical variables for turnover after farming activities with plants. The results in a much better fitting in terms of the diagnostics R^2 and RMSR in the testing set, from which the 2nd-phase model is built. However, judging from the error prediction is somewhat worse than the model build in setting C.2, indicating the issue of over-fitting.

b) In setting C.2, we use fewer additional features, including organisation form, county, ratio of $\hat{Y}$ by linear regression, tree, gradient boosting and $\hat{Y}$ by random forest. The results are better than those of setting C.1, although the fitting is much worse according to the model diagnostics in the testing set. However, if one use the closeness between RMSR and RMSE as the criterion, obtained from the testing set and the final testing set, respectively, it is possible to select setting C.2 instead of C.1. This illustrates again the potential advantage of data-splitting, which allows one to choose models based on errors instead of residuals.

Both model building and selection are obviously topics that need further investigation for TPL.

C.2. *Prediction of sign(ϵ) and ϵ^2.*

1. Various error functions, such as sign(ϵ) and ϵ^2, concern different aspects of the prediction uncertainty due to mean and variance heterogeneity. Prediction of sign(ϵ) and ϵ^2 require separate models than ϵ, since a model of $E(\epsilon|x, z)$ cannot be used to infer sign(ϵ) or ϵ^2.

2. Table 5 gives some results of predicting PPS (left) and RMSE (right) based on the 2nd-phase models for sign(ϵ) and ϵ^2, respectively, together with the corresponding true values in the rest of population for easy comparison. In the case of PPS, comparing the results of setting A and B for feature selection, one can see that using organisation form as an additional feature improves the prediction in the groups by organisation form, which is not surprising. It is used to illustrate the point that additional features may be included to serve the needs of dissemination, even if these features are not very effective for prediction of Y and are therefore not selected in the 1st-phase model. For RMSE, we simply present here the results under setting A, which uses the same features (x) for 2nd-phase model as for the 1st-phase model, which are already greatly improved compared to direct extrapolation of the testing-set metric 1786.287 (Table 3) for all the units in the rest of population.

TABLE 5. PPS and RMSE for rest of population: true vs. predicted.

		PPS Predicted			RMSE Predicted
	PPS	Setting A	Setting B	RMSE	Setting A
Overall	0,326	0.345	0.341	1309.380	1372.608
DA	0.390	0.419	0.392	1256.090	1517.335
ENK	0.321	0.343	0.339	1188.181	1298.856
OTHER	0.465	0.356	0.391	4407.440	3541.732

C.3. *On-the-flight correction.*

1. It is possible to use the 2nd-phase prediction model of ϵ to adjust the 1st-phase prediction $\hat{Y}$. Since the predictor is built at the individual level, the adjustment can be made on-the-flight when responding to ad hoc queries, whichever the subpopulation that is of interest on the occasion. Since the performance of the 1st-phase model will always be relatively poor for some parts of the population, on-the-flight adjustment based on 2nd-phase learning can provide valuable improvements.

2. In the current example, on-the-flight adjustment of the 1st-phase random forest predictions yields an overall mean error -18.441 (with RMSE 1282.672) for the rest of population. The improvement is statistically significant, compared to -157.587 by the 1st-phase model itself (Table 3). There is clearly a connection between the different 2nd-phase models of prediction uncertainty and the choice of when (i.e. for which units) to apply on-the-flight adjustments. It is an intriguing issue for future research.

IV. **Conclusion**

1. Performance metrics of the 1st-phase model in the testing set are generally inadequate for assessing the individual-level prediction errors for the out-of-sample units. The proposed Two-Phase Learning (TPL) provides a framework for capturing both mean and variance heterogeneity of the prediction errors. It can improve the assessment of prediction error at the individual level, as well as offering the possibility of on-the-flight adjustment of the chosen 1st-phase model.

2. Models for the prediction error and its functions should be built separately. Incorporating additional features for the 2nd-phase learning is generally needed and beneficial to the results. Effective approaches to the 2nd-phase model building and selection should be further investigated. How to make on-the-flight adjustment of the 1st-phase model prediction, for the instances where it may perform poorly, is another useful research topic in the context of TPL.