EFFECT OF OMITTING OUTLIERS ON THE UNCERTAINTY OF DATA-BASED PREDICTION

Author's name(s): Ziller, M.

Affiliation(s): Federal Research Centre for Virus Diseases of Animals, Institute of Epidemiology, Seestr. 55, D-16868 Wusterhausen/Dosse, Germany

E-mail: mario.ziller@wus.bfav.de
Telephone: +49 33979 80182
Fax: +49 33979 80200

Corresponding author: Dr. Mario Ziller

Keywords: Prediction, Outliers, Uncertainty

Topic area of the submission: Risk Assessment

Abstract

Objectives

Statistical results can strongly be affected by outliers. Therefore, such outliers are sometimes omitted in order to perform a more robust analysis. For a variety of problems, related consequences have been widely discussed. This paper presents a theoretical approach for the situation of data-based prediction which nowadays is omnipresent in many applications of risk analysis in politics, economics, public health, veterinary epidemiology, biosecurity, pollution control etc.

Data-based prediction means forecasting based on extrapolation using regression models. Assessing uncertainty is one item within the established principles of forecasting [1]. The main question within this paper will therefore be: How does the omission of outliers affect the uncertainty of forecasting?

In general, variances will be underestimated from the ‘cleaned’ data. It has apparently the effect of more confidential results. But in prediction, uncertainty has at least two aspects: Uncertainty of the estimation and uncertainty of the validity of the prediction. The question above will be examined considering prediction intervals, and in addition measures of the
relative accuracy of prediction [2], describing the relation between the variances of the model residuals and of the predictive residuals.

Methodology

Consequence assessments in risk analysis frequently require the prediction of extreme values, too [3]. Worst-case scenarios are essential for cash-flow evaluations, epidemiological emergency control measures and in many other cases. The variance of the estimator influences the accuracy of the forecasting of extremes in a particular manner. Therefore, it is necessary to estimate the predictive values with a sufficient precision. In general, this is only possible in a close region around the given data-points. But what is acceptable? An objective criterion would be desirable.

Although a model may sufficiently describe the underlying dependence within the data-region, no evidence for the validity of the model is given outside this region. For this reason, considering two measures of the uncertainty of forecasting is necessary: the absolute and the relative accuracy of prediction.

The prediction interval represents an absolute measure of the accuracy of prediction, describing the uncertainty of estimation. It is based on the variance of the predictive residuals, the sum of the variances of the model residuals and of the estimator. The relative accuracy of prediction describes the uncertainty about the validity of the prediction, i.e. the validity of the model. It can be quantified by the proportion $\Psi_{pr}^2$ of the variances $\sigma^2$ of the model residuals and $\omega^2$ of the predictive residuals [2].

Analytical solutions for the estimation of predictive residuals are known for several models. For others, simulation is needed. Usual techniques like split-half, predictive sample re-use, one-at-a-time-omission, etc. may be applied [4]. In what follows, we consider the general case of least square regression. In an analytical approach, the effect of outliers on the variance of the model residuals will be evaluated. For specific calculations concerning the absolute and relative accuracy of prediction, the simple linear regression model will be assumed. Analytical solutions are known for this case [5].

Results

The following statements will be proved for the case of least square regression.
1. Omitting an outlier leads to a decrease of the variance of the model residuals. A simple sufficient outlier-criterion and a more complicated but necessary and sufficient criterion will be presented.

2. If the linear model is assumed, omitting an outlier leads to a decrease of the variance of the estimator and therefore to an increase of the absolute accuracy of prediction. The corresponding outlier-criteria will be presented.

3. Omitting any data point in a linear regression analysis, and in particular an outlier, leads to a decrease of the relative accuracy of prediction.

![Linear regression results for a dataset with and without outliers (filled dark grey).](image)

The presented results are illustrated in Figure 1 on principle. The regression lines for a complete dataset (dashed line) and for the same dataset after omitting the outliers (dotted line) are plotted together with the corresponding simultaneous prediction intervals ($\alpha=0.05$). Regions of the same acceptable level of the relative accuracy of prediction ($\Psi_{pr}^2 > 0.8$) are shaded dark grey for the case with outliers, and light grey for the case without outliers.

**Discussion**

Forecasting has an additional aspect of uncertainty than only estimating parameters. The concept of goodness of fit will not suffice in data-based prediction. An extension of the considerations is needed for the interpretation of any forecasting result.
In this paper, a measure of the relative accuracy of prediction, i.e. the relation between the variances of the model residuals and of the predictive residuals, will be applied in addition. It reflects the uncertainty about the validity of the model. It is scale-free and therefore suitable for study-planning and model choice in prediction-tasks as well as for the comparison between different forecasting approaches.

The consequences of omitting an outlier (or generally of omitting information from the data) on the uncertainty of forecasting have been investigated. For the regression problems under consideration, omitting an outlier results in an apparently more precise prediction, but valid at the same level only for a smaller time period. The uncertainty of the estimation decreases, but the uncertainty of the validity of the model increases.

This corresponds with the general experience that a problem may appear clearer, more explicit, and more obvious after omitting unspecific information. But in fact, this works like a filter resulting in simplification and less universal conclusions.

References