MINIMUM SAMPLE SIZE CALCULATION FOR MULTIVARIABLE REGRESSION OF CONTINUOUS OUTCOMES: IMPLICATIONS FOR PLANETARY EXPLORATION. M. Konstantinidis[1,2], M. Veneranda[3], E. A. Cloutis[4], G. Lopez-Reyes[3], M. G. Daly[2], and E. A. Lalla[2]. [1]Division of Biostatistics, Dalla Lana School of Public Health, University of Toronto (155 College St. Toronto ON, Canada; menelaos.konstantinidis@mail.utoronto.ca), [2]Centre for Research in Earth and Space Science, York University, [3]Unidad Asociada UVA-CSIC-CAB. C/ Francisco Valles 8, 47151, Boecillo, Spain [4]Department of Geography, University of Winnipeg.

Introduction: Multivariable regression models are increasingly developed for automatic characterization of geological samples in the planetary context. This is particularly true in contexts such as elemental abundance on Mars, where chemometric models have been developed for the ChemCam and SuperCam instruments [1- 2].

In the simplest case, a prediction model seeks to approximate the expected value of an outcome as a function of one or more explanatory variables. Such a model for continuous outcomes may be as "simple" as a multivariable linear model, or as complicated as a neural network. Regardless of the model, however, the final objective is to use the developed model for the prediction of the outcome in samples for which the value is unknown.

While the ease of data acquisition for model calibration is clear in many fields, generating data for planetary exploration remains a laborious process and is often resource intensive. Therefore, "convenience samples" (i.e., a dataset that has not been systematically developed) are often used to develop a calibration model. While useful in principle, such a sample may give rise to methodological challenges in model fitting such as overfitting and lack of parameter precision.

Overfitting refers to a situation in which there are too many degrees of freedom [3]. In other words, it is a situation in which a model lacks the necessary sample size, relative to the number of parameters. This may then lead to a false sense of security that the model in question will provide reasonable estimates of external data. It is therefore necessary that when calibrating a model, the necessary sample size be obtained. The rest of this abstract is focused on how to determine what this minimum is.

Methods: Consider, in the simplest case, a multivariable linear model of the form

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + e_i \quad (1)$$

for $i = 1, \ldots, n$ observations, such that $e_i \sim N(0, \sigma^2)$.

Given such a model, we wish to find the minimum sample size n_{min} for a given number of predictor parameters p such as the risk of overfitting is minimized.

Coefficient of determination. Previous work has suggested that to minimize the risk of overfitting, the difference between the apparent coefficient of determination R^2 and the adjusted coefficient of determination R^2_{adj} should be small (e.g., at most 0.05) [4]. We define this objective function as

$$\Delta = R^2 - R^2_{adj} \quad (2)$$

We note that

$$R^2_{adj} = (1 - R^2)\frac{(n-1)}{(n-p-1)} \quad (3)$$

is an approximately unbiased estimator of R^2_{adj} [5]. Then,

$$R^2 = \frac{R^2_{adj}(n-p-1)+p}{n-1} \quad (4)$$

Subsequently, plugging (4) into (2), we find that

$$\Delta = \frac{p(1-R^2_{adj})}{n-1} \quad (5)$$

and that

$$n_{min} \geq 1 + \frac{p(1-R^2_{adj})}{\Delta} \quad (6)$$

Global shrinkage factor. As an alternative approach, we consider the so-called Global Shrinkage Factor. In this approach, following the fitting of eq. 1, a shrinkage factor S is applied to the estimated parameters. This is effectively a penalization method used to minimize overfitting. Subsequently, we have a revised version of eq. 1,

$$Y_i = \beta_0' + S(\beta_1 X_1 + \beta_2 X_2 + \cdots) \quad (7)$$

where β_0' is a revised intercept and S. In this work, we proceed with the Copas estimator which is defined as

$$S_c = 1 - \frac{p-2}{LR} \quad (8)$$

where LR is the likelihood ratio between the null model and the full model. Moreover, we note that

$$LR = -n\log(1 - R^2) \quad (9)$$

from which

$$S_c = 1 + \frac{p-2}{n\log(1-R^2)} \quad (10)$$

Thus, applying eq. 4,

$$S_c = 1 + \frac{p-2}{n\log\left(1-\left(\frac{R^2_{adj}(n-p-1)+p}{n-1}\right)\right)} \quad (11)$$

It therefore follows that we can find the minimum sample size by finding the value of n that satisfies a given threshold of S_c. That is,

$$\min_n \left\{ 1 + \frac{p-2}{n\log\left(\frac{\left(1-R^2_{adj}\right)(n-p-1)}{n-1}\right)} \geq S_c \right\} \quad (12)$$

which can be easily solved numerically.

Results: For the purpose of illustration, we assume that we are interested in developing a dataset consisting of spectra of geological samples, for example from Laser-induced Breakdown Spectroscopy (LIBS). As is common in chemometric methods, such a dataset might first be transformed into a latent space, by Principal Component Analysis, for example, such that the number of dimensions is substantially reduced. Therefore, let $p = 25$, a not unreasonable assumption.

In previous work, a value of $\Delta \leq 0.05$, and $S_c \geq 0.9$ has been suggested [4]. What then remains is to determine R^2_{adj}. The present approach requires that R^2_{adj} be assumed a priori, since in general, the dataset has not yet been developed. How to select R^2_{adj} is a multifactorial decision; however, it has been suggested that when the explanatory variables are mechanistic in their effect on the outcome, $R^2_{adj} = 0.5$ is reasonable [4]. Such is the case in LIBS for example, where the emissions are correlates of elemental abundance. Therefore, we can now apply eq. 6 and eq. 12 to calculate the minimum sample size required for the calibration set, based on the two criteria discussed above. Specifically, we find that based on Δ,

$$n_{min} = 1 + \frac{25(0.5)}{0.05} = 251$$

and based on S_c,

$$S_c = 1 + \frac{25-2}{n\log\left(1-\left(\frac{0.5(n-25-1)+25}{n-1}\right)\right)}$$

such that $n_{min} = 269$. Thus, to err on the side of caution, with 25 predictors, we would choose the greater of the two sample sizes, that is, we need a sample size of 269 to minimize the risk of overfitting.

Discussion: Despite the fact that minimum sample size calculations are commonplace in many fields such as medicine and biostatistics, such calculations remain scare in chemometrics and space science. However, we have demonstrated that such calculations are both essential and relatively straightforward to determine.

Nevertheless, it may be the case that the calculated sample size is infeasible (e.g., due to cost of sample acquisition). In such a case, we suggest that the number of dimensions be reduced, while keeping the thresholds (Δ and S_c) constant. Moreover, such calculations often require some a priori assumptions (e.g., what R^2_{adj} might be). While we have suggested a value of 0.5 for mechanistic relationships are frequent as is the case in spectroscopy, it is advisable that when establishing minimum sample sizes, researchers refer to past literature (e.g., previous relevant calibration models).

Lastly, the present work has been presented in the context of linear models; however, the principles are easily extendible to more complex models as might be observed in Partial Least Square Regression, for example. In more complicated cases, even if a closed form solution is not attainable, sample size simulations are easily conducted to the same end.

Acknowledgments: The authors would like to acknowledge the financial support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References: [1] Clegg S. M. et al. (2017) *Spectrochim. Acta Part B 129: 64-85.* [2] Anderson R. B. et al. (2021) *Spectrochim. Acta Part B: 106347.* [3] Steyerber E. (2019) *Clinical Prediction Models: 95-112.* [4] Riley R. D. et al. (2018) *Stat Med 38(7): 1262-1275.* [5] Copas J. B (1983) *J R Stat Soc Series B 45(3):311-354.*