Sample size effects when using R^2 to measure model input importance

Michael D. McKay
Mark A. Fitzgerald
Richard J. Beckman
Los Alamos National Laboratory, Los Alamos, NM, USA

ABSTRACT: Sample sizes affect identification of important inputs for computer models. For illustrative purposes, a partial differential equations model with 84 input variables is used to investigate the behavior of R^2 as an importance indicator for various sample sizes and designs.

1 INTRODUCTION

Sample sizes affect identification of "important" inputs for computer models when one is using statistical measures, like correlation. In this regard, it is difficult to make useful general statements about sample size—like, one needs n samples—because effects depend on the particular model being studied. In this paper, we present illustrative results for an environmental pathways model with 84 inputs. Our purpose is to gain insight into possible sample size effects for a realistic model with a moderate number of inputs, for which the number of computer runs is not a limitation. In the sections that follow, we define in mathematical terms what we mean by "importance," present an experimental design for computer experiments, show typical results for a single case study and, finally, summarize results from simulation experiments using different sample sizes.

2 SITUATION

Suppose that the prediction y from a model $m(\cdot)$ is determined by a vector x of input variables. The input variables can define initial conditions of a system being modeled as well as parameter values in the rules determining y from the initial conditions. We think about the importance of individual inputs and, more generally, the importance of subsets x^s of inputs in the sense that they "drive" or "control" the calculation of y. Let us assume that a probability distribution describes variation in the model input vector x. That is, x is a random variable with probability (density) function f_x. One can think of f_x as describing incomplete knowledge of values of x, or imprecisely specified conditions being modeled, or the range of values for which the model $m(\cdot)$ is designed to operate. It is significant that we will be considering global importance over a domain of values of x and not local importance or the effect of an incremental change in x about a nominal or fixed value.

The prediction $y(x)$ depends on the entire input vector x. We introduce the restricted prediction $\widetilde{y}(x^s)$ that depends only on a subset x^s of inputs as follows. Let f_s be the (marginal) probability density of the subset x^s. The distribution of the complementary subset $x^{\overline{s}}$ conditioned on the value of the subset x^s is denoted by

$$f\left(x^{\overline{s}} \mid x^s\right) = \frac{f_x(x)}{f_s(x^s)} \,. \tag{1}$$

The components of x are not required to be statistically independent. We define $\widetilde{y}(x^s)$ to be the conditional expectation of y, conditional on the value of x^s,

$$\widetilde{y}(x^s) = \int y(x) f\left(x^{\overline{s}} \mid x^s\right) dx^{\overline{s}} \,. \tag{2}$$

The predictor $\widetilde{y}(x^s)$ is the value of y at x^s averaged over the values of the remaining variables $x^{\overline{s}}$. We arrive at a measure of importance through consideration of the squared error loss function

$$\mathcal{L}(x) = [y(x) - \widetilde{y}(x^s)]^2 \tag{3}$$

and its expected value, the mean squared error of prediction,

$$E(\mathcal{L}) = \int [y(x) - \widetilde{y}(x^s)]^2 f(x) dx \,. \tag{4}$$

Because the variance of y can be written

$$\mathrm{Var}[y] = \mathrm{Var}[\widetilde{y}] + E(\mathcal{L}) \,, \tag{5}$$

the importance of the subset x^s can be measured by the Pearson correlation ratio (Kendall & Stuart 1979)

$$\eta^2 = \frac{\text{Var}[\widetilde{y}]}{\text{Var}[y]}\,, \tag{6}$$

which is a value between zero and one. The Pearson correlation ratio has been used also by Krzykacz (1990) and by Iman & Hora (1990).

The sample "multiple correlation coefficient" R^2 from a random effects model is a biased estimator of the correlation ratio η^2 (McKay 1995, 1997). The estimator takes the form

$$R^2 = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}\left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2} \tag{7}$$

where

$$\begin{aligned} \overline{y}_{i\cdot} &= \frac{1}{J}\sum_{j=1}^{J} y_{ij} \\ \overline{y}_{\cdot\cdot} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} y_{ij}\,. \end{aligned} \tag{8}$$

The $\{y_{ij}\}$ are model calculated values from an independent random sample $\left\{x_1^s, \cdots, x_I^s\right\}$ of x^s, and conditionally independent samples $\left\{x_{i1}^{\overline{s}}, \cdots, x_{iJ}^{\overline{s}} \mid x^s = x_i^s\right\}$ of $x^{\overline{s}}$.

The objective to find important subsets x^s is closely related to the familiar process of subset selection in multiple-regression analysis. In the following sections, we examine the quality of R^2 for this task. Evaluation is based on three criteria. One criterion is power. The power of R^2 to identify truly important inputs should be high. Next, we look at the probability that truly unimportant inputs are incorrectly identified as being important. This probability should be low. The final criterion is consistency. That is, the inputs identified as important by R^2 should remain fairly constant across independent samples of the same size. The manner in which R^2 performs with respect to these criterion depends on the specific model $m(\cdot)$ and the sample size and design $N = I \times J$.

3 SUBSETS OF SIZE ONE: INDIVIDUAL INPUTS

We present results of the investigation of R^2 for the simplest case—subsets of size one. To see how sample size might affect the performance of R^2, we performed a simulation experiment.

3.1 *Experimental design of the sample*

When the inputs are independent, a single set of computer runs can be reused to calculate R^2 for each input with an experimental design based on Latin hypercube sampling (McKay, Conover & Beckman 1979). In an LHS of size I, the range of each input variable is stratified into I intervals of equal probability. Values are sampled from the conditional probability distribution in each interval and combined completely at random across inputs. Obviously, LHS is appropriate only for statistically independent random variables. For a "replicated" LHS (rLHS) of size $N = I \times J$, J independent random combinations are constructed. In the resulting sample design, I distinct values of each input are paired with J independent combinations of values of the other inputs. Therefore, for each input, the N calculated output values can be labeled $\{y_{ij} \mid j = 1, \cdots, J;\ I = 1, \cdots, I\}$ and used in Equation 7 to calculate an R^2 for the input.

3.2 *Case study*

In the case study used for the simulation experiment, the model is an environmental pathways model having 84 inputs (see McKay et al. 1992). The model is a set of differential equations describing the flow of material within a network of 8 interconnected compartments. The output is the equilibrium concentration in one of the compartments. The inputs are the transfer rates in the differential equations. As part of the investigation of properties of R^2, we added 16 dummy inputs not used in any calculation. Values for these fictitious inputs were randomly assigned. We treat all 100 inputs as statistically independent and use the experimental design of Section 3.1. Results for a typical analysis using the largest sample size in the simulation experiment are presented in the next section. The sample size of $N = 5000$ is allocated to $I = 100$ and $J = 50$. While there is still some sample variability even for this large design, we think that the example is a good and fair illustration of a well designed experiment. The objective of the analysis is to identify those inputs that individually (singly) have distinguishably larger effects on the output than do the remaining inputs. Although individually important, it may not be the case that collectively they form important subsets of sizes greater than one. We leave that discussion to another time.

3.3 *Typical results from case study*

The R^2 values from an $I = 100$ and $J = 50$ design are plotted in Figure 1, largest to smallest. The graph shows 4 distinct sets of inputs. Inputs with the largest R^2 values are those with labels $\{1, 68, 69\}$. Those with the next largest are $\{24, 63, 84\}$. Those with the third largest value of R^2 are $\{35, 48, 67, 83\}$. The final set is all the other inputs. Within each set, the

inputs are not distinguishable as to importance. We omit theoretical discussion of the graph.

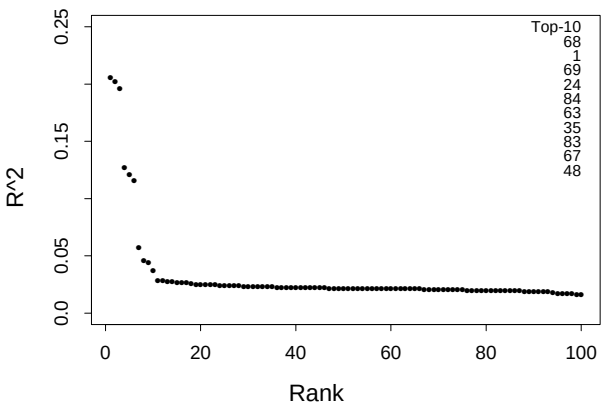

Figure 1. R^2 values for an $I = 100$ and $J = 50$ design with the top 10 inputs indicated.

Figure 2 gives graphical displays of the effect of the top 9 inputs. The 100 values of each input are on the horizontal axis. The 5000 values of y are ranked and scaled by 5000. There are 50 transformed values of y corresponding to each value of an input. It is the average of these values that is plotted in Figure 2. The variables in the top row in the figure account for low values (around 0.2) of y. The effect of the 9th input in the lower right corner is slight, causing only a small trend about the overall average value of 0.5.

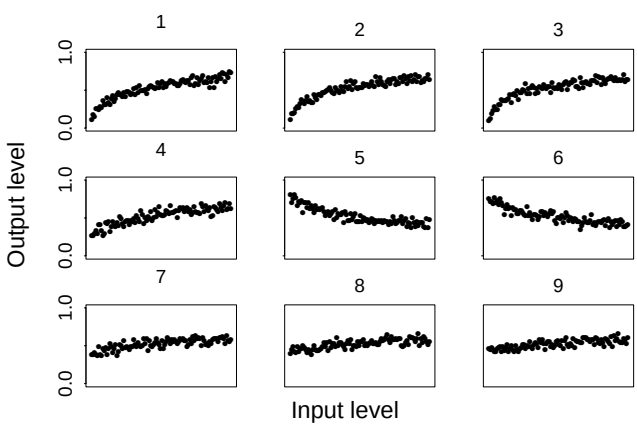

Figure 2. Patterns of average rank-transformed y for top 9 inputs.

4 SIMULATION STUDY OF R^2

In order to evaluate power to identify, false identification and consistency, we performed a simulation experiment composed of 25 different case studies like the one in the last section. Values of I from the set $\{5, 10, 20, 50, 100\}$ were each matched with values of J from the set $\{2, 5, 10, 20, 50\}$ to form the 25 different cases. Each case study was repeated 9 times to allow for visual examination of sample-to-sample variability and consistency.

4.1 *Power to identify*

The term "power" is used in a technical sense to refer to the probability that R^2 can distinguish a true model input from a dummy input variable, as described in the next section. The plots in Figure 3 show how the values observed for R^2 in our simulation study change as a function of I and J. For the $I = 5$ and $J = 2$ case in the upper left, the relatively straight line indicates the inability of R^2 to identify important inputs. Moving down from that plot, we see some improvement. However, moving horizontally with J increasing to 50, we see at least 6 inputs beginning to stand out. The graph in the lower right shows that 10 inputs stand out. As before, we omit theoretical discussion of the figure.

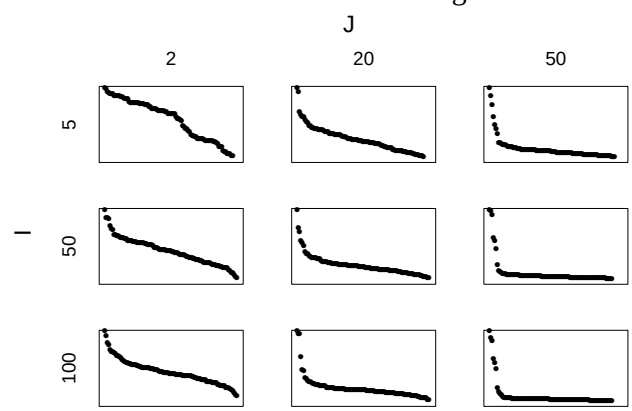

Figure 3. Patterns of sorted R^2 for 100 inputs for different sample designs I and J.

4.2 *Incorrect identification*

Sixteen dummy inputs were included with the 84 real inputs in the analyses. Their frequency of occurrence in the top 10, for example, would be an indication of the inability of R^2 to identify important inputs. If R^2 were unable to distinguish the dummy inputs from the real ones, one would expect to see, on average, 1.6 dummy inputs among the top 10 selected. If R^2 were selecting inputs at random, then at least one of the dummy inputs would appear in the top 10 in 85% of samples. Therefore, we look for I and J values for which few dummy inputs appear. Table 1 suggests the region towards the lower right is acceptable.

Table 1. Average number of dummy inputs in top 10 from 9 simulations of 25 case studies.

Value of I	Value of J				
	2	5	10	20	50
5	1.2	1.2	0.9	1.1	0.7
10	1.7	1.0	0.7	0.8	0.3
20	2.0	0.8	0.4	0.7	0.3
50	1.9	0.4	0.8	0.1	0.1
100	1.3	1.3	0.1	0.1	0.0

4.3 *Consistency*

We examine the consistency of variable selection using R^2 by looking at the composition of the sets of the top k inputs, for $k = 1, 2, \cdots, 100$. Let $s(k)$ be a subset selection vector of length $n = 100$. The elements of $s(k)$ are 0–1 indicators of the k inputs with the largest R^2 values. We measure the consistency of selection between any two cases, as a function of k, by the squared distance between their normalized subset selection vectors. For this simulation study of 9 cases, we compute the mean squared distance (MSD) for the 36 distinct pairs and plot MSD as a function of k in Figure 4. The MSD is calculated as follows.

$$d_{ij}(k) = \frac{1}{\sqrt{k}}[s_i(k) - s_j(k)]$$
$$\text{MSD}(k) = \frac{1}{36}\sum_{i=1}^{8}\sum_{j=i+1}^{9} d_{ij}^T(k)d_{ij}(k)\,, \tag{9}$$

where T indicates transpose of the vector. For comparison, when the k out of n ones of $s(k)$ are assigned at random, the expected value of MSD is linear in k and given by

$$E[\text{MSD}(k)] = 2\left(1 - \frac{k}{n}\right)\,. \tag{10}$$

The graph in the upper left corner of Figure 4 indicates that for the $I = 5$ and $J = 2$, the selection of subsets appears to be random, by comparison to Equation 10. That is, for this design, R^2 cannot distinguish among the different inputs. Therefore, there is no consistency across samples. Sample size effects are inferred by examination of the figure.

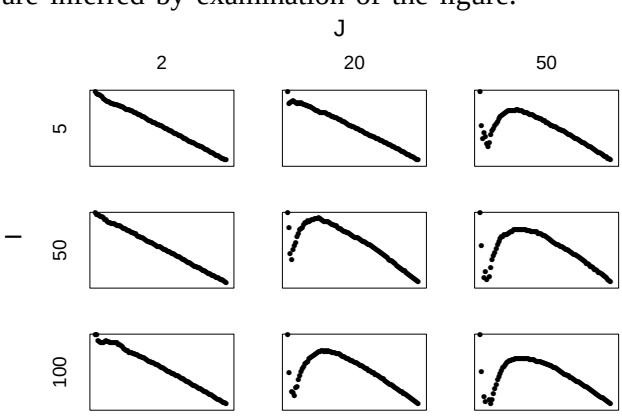

Figure 4. Patterns of MSD for k from 1 to 100 for different sample design values of I and J.

When the top 10 inputs in the $I = 100$ and $J = 50$ designs comprising the graph in the lower right corner of Figure 4 are replaced by dummy inputs, the graph changes to that in Figure 5. The relatively straight line is the one predicted by Equation 10 for 100 random inputs, and supports the hypothesis that there are only 10 distinguishably important inputs.

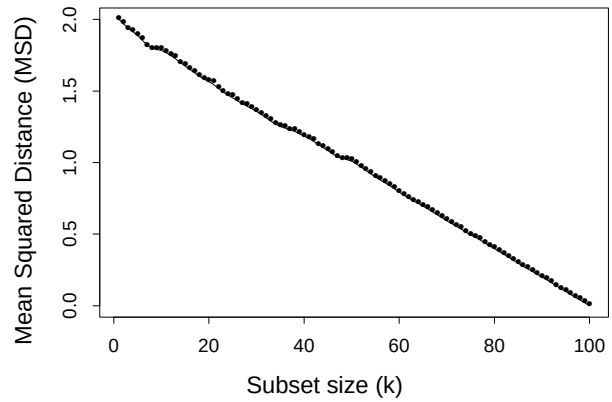

Figure 5. MSD for $I = 100$ and $J = 50$ with the top 10 inputs replaced by dummies.

5 CONCLUDING REMARKS

The simulation study suggests directions for important simulation and mathematical exploration. The study points out weaknesses of R^2 as a measure of importance for smaller sample sizes. It also shows that caution is advisable until sample size effects are known for the model under investigation. The simulation study also suggests interesting and powerful properties of R^2 as an importance measure.

6 REFERENCES

Iman, R.L. & S.C. Hora 1990. A robust measure of uncertainty importance for use in fault free system analysis. *Risk Analysis* 10(3):401–406.

Kendall, M. & A. Stuart 1979. *The advanced theory of statistics*, 4th ed., vol. 2, ch. 26. New York: Macmillan.

Krzykacz, B. 1990. SAMOS: a computer program for the derivation of empirical sensitivity measure of results from large computer models. Tech. Rep. GRS-A-1700. Gesellschaft fur Reaktorsicherheit (GRS) mbH. Garching, Republic of Germany

McKay, M.D. 1995. Evaluating prediction uncertainty. Tech. Rep. NUREG/CR-6311. U.S. Nuclear Regulatory Commission and Los Alamos National Laboratory.

McKay, M.D. 1997. Nonparametric variance-based methods of assessing uncertainty importance. *Reliability Engineering and System Safety* 57:267–279.

McKay, M.D., R.J. Beckman, L.M. Moore & R. R. Picard 1992. An alternative view of sensitivity in the analysis of computer codes. *Proc. of the American Statistical Association, Section on Physical and Engineering Sciences.* Boston.

McKay, M.D. & W. J. Conover, & R. J. Beckman 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* 21(2):239–245.