DIMENSIONAL ANALYSIS IN DATA MODELLING [*]

G A Vignaux [†]

June 1991

Abstract

Dimensional Analysis can make a contribution to model formation when some of the measurements in the problem are of physical factors. The analysis constructs the set of independent dimensionless factors that should be used as the major variables of the regression in place of the original measurements. There are fewer of these than the originals and they may have a more appropriate interpretation. The technique is described briefly and its proposed role in regression illustrated with an example.

1 Introduction

When setting up a problem we should, to paraphrase Einstein, "use all the information we have but no more." To use all our information suggests a Bayesian approach; to use no more is in the spirit of Maximum Entropy. Despite this sage advice we often ignore information carried by the physical characteristics of the variables involved. Dimensional Analysis, written as DA throughout the rest of this paper, ensures that this physical information is always used.

[*] Published in C R Smith, G J Erickson and P O Neudorfer(eds), *Maximum Entropy and Bayesian Methods, Seattle, 1991*, Kluwer Academic Publishers, 1992

[†] Institute of Statistics and Operations Research, Victoria University, Wellington, New Zealand

Experienced statisticians have long agreed that there is more to a statistical analysis than throwing all the data into a computer program and reporting the results of the regression. They will go through a phase of exploratory analysis of the data inspecting graphs and examining residuals. Based on the observed curvatures they will manipulate and transform the data until approximately linear relationships are obtained.

DA can be used in this preliminary analysis and assists, and may replace, the stage of data transformation. It was, and is, a standard working tool for many of the greatest mathematical physicists, such as Rayleigh[7]. It should also be quite familiar to readers but I have discovered that it is new to many statisticians and operations research analysts. Or perhaps it is not new, for many will have learned it at school or college. What is novel is that it might actually be useful.

I propose to demonstrate its application to a regression problem where at least some of the variables involved are physical quantities. For those to whom it is unfamiliar, DA is introduced in the next section.

2 What is Dimensional Analysis?

DA is a technique for restructuring the original dimensional variables of a problem into a set of dimensionless products using the constraints imposed upon them by their dimensions [4, 9, 10]. It is ultimately based on the simple requirement for dimensional homogeneity in any relationship between the variables.

Buckingham[1] showed in his Pi theorem that if the original unknown relationship is represented by $f(x_1, x_2, \ldots, x_n) = 0$, where the x_i are the variables, it can be transformed into a new function $\phi(\pi_1, \pi_2, \ldots, \pi_{n-m})$ of $n - m$ independent dimensionless products π_j of the original x_i variables. m is the number of fundamental dimensions out of which the dimensions of the original physical variables are ultimately composed. In the physical sciences these are length, mass, and time $[L, M, T]$. In my own discipline we would add quantity and cost $[Q, \$]$.

Some problems also involve special dimensional constants such as viscosity, or resistance, each of which correspond to the statement of a relevant physical relationship such as Stokes law or Ohm's law.

The dimensionless πs can be constructed in a pretty mechanical manner. One needs a basis of m of the original variables to satisfy the m constraints in the exponents corresponding to the m fundamental dimensions. Any basis can be used but some bases have advantages over others. The basis should include only variables with linearly independent exponent vectors (e.g. one cannot have two velocities in the basis). This leaves $n - m$ independent πs for the regression.

Immediately we see an advantage in the analysis. The number of variables to be included has been reduced from the original n to $n - m$. This certainly makes the problem simpler and may make it trivial. It may even make the regression unnecessary.

A popular demonstration of DA involves the derivation of the formula for the period of a pendulum[8]. What starts out as a 4-variable problem is reduced to finding the value of a single dimensionless constant. A similar collapse to a single dimensionless constant is seen in the derivation of the optimum lot-size in simple deterministic inventory problems[6, 9, 10].

Reminded again of Maximum entropy methods, it appears at first sight that in using DA we are getting something for nothing. In fact we are only applying to the problem at the start those constraints that we know must be satisfied by the final solution.

3 An Example - Hocking's Automobile Data

This classical data analysis problem is to fit a regression model to predict the gas consumption m (miles per gallon) from other characteristics of the automobile. The collection of gas mileage data from the 1974 *Motor Trend* magazine was used as a test case by Hocking[3] to investigate automatic methods of regression. Henderson and Velleman[2] examined the same data (and also a second set, collected from *Consumer Reports*), and argued for an alternative philosophy of computer assisted data analysis - a collaboration between the analyst and the computer in an interactive mode. This is in accord with the philosophy of 'data analysis' rather than blind regression.

Here we add a preliminary Dimensional Analysis, combined with some physical insight. We introduce some additional dimensional constants and convert the set of variables to a smaller number of dimensionless products. The regression is then carried out on the products.

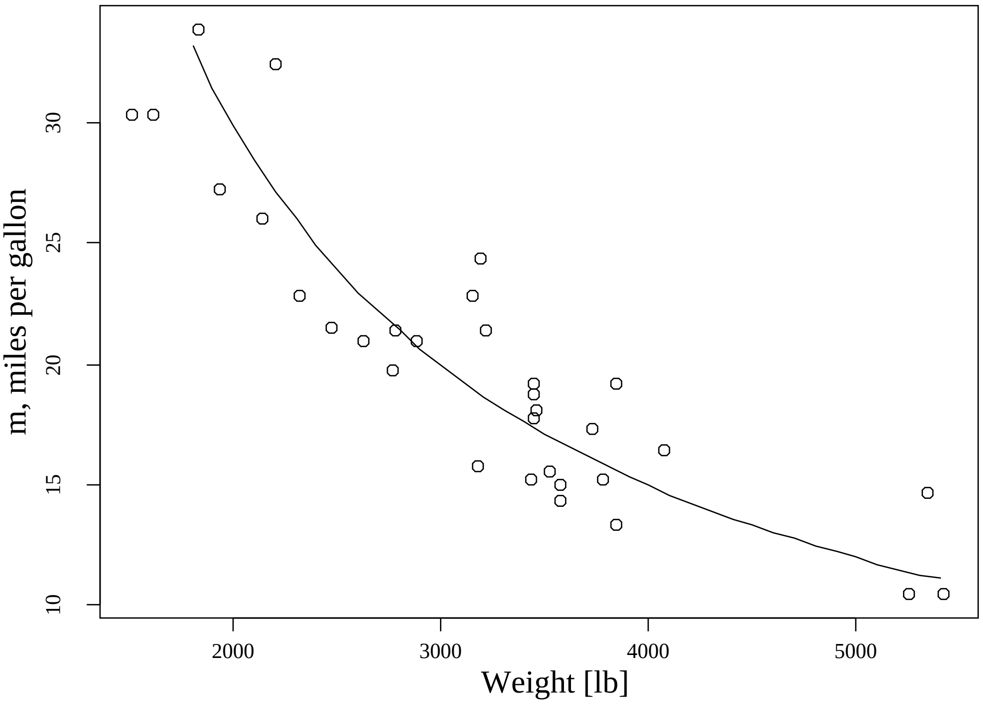

Figure 1: Automobile Miles per Gallon, m, versus Weight, w. The curve is given by the simplest dimensionless model.

3.1 Examining the variables

The data (reproduced in [2]) contains 11 variables, of which 6 are already dimensionless, being, in the previous authors' notation, either ratios (such as $DRAT$, the final drive ratio) or numbers (CYL, the number of cylinders). The four variables that have physical dimensions, m, h, w and q are listed in Table 1 with their dimensions. The remaining variable, $DISP$, the cylinder displacement, has an intermediate nature in that it has dimensions (cubic inches in this case) but is really a surrogate for engine size.

The dimensionless variables and ratios are appropriate candidates for inclusion in the data analysis or regression; they are already of the correct dimensionless product form that we hope to construct from the other variables and may carry information about the behaviour of the system.

Table 1: The variables and their dimensions

variable	*symbol*	*Description*	*Dimensions*
m	MPG	miles per gallon	LG^{-1}
h	HP	horsepower	ML^2T^{-3}
w	WT	weight	M
q	$QSEC$	Quarter mile time	T

3.2 Further Consideration

If one considers how these four variables might be collected together into the form of dimensionless products, one is immediately struck by the fact that m $[LG^{-1}]$ cannot be combined with the others because of the $[G]$ term which does not appear in any of the other three. (DA experts will notice that I have kept dimension $[G]$ separate rather than replacing it with a volume measurement, $[L^3]$)

One notes that gallons are those of a fuel containing energy. This suggests that we introduce a new dimensional constant E $[ML^2T^{-2}G^{-1}]$ giving the energy content in each gallon to reflect the implied physical "law" that 1 gallon contains contains a fixed amount of energy. Energy has conventional dimensions, ML^2T^{-2}.

This has added another dimensional constant E to our set of 4 variables with the gain of a sensible link between m and the others. We expect E to remain constant, though with a more extensive data set this could be included as a variable which might change its value with different types of fuel.

The variable q $[T]$, Quarter mile time, is really a surrogate measure of acceleration. At least we should add to the variables the distance over which this acceleration was measured, d $[L]$.

3.3 Other physical relationships

We may wish to consider further physical effects that could be involved. These might include the possible effect of rolling or frictional resistance, and the effect of air resistance. These will have "laws" that connect the variables to each other and to new dimensional constants. Since rolling or frictional resistance is likely to be important, I have added the gravitational acceleration g $[LT^{-2}]$ to the set of variables.

3.4 Dimensional analysis

Including the energy coefficient of gas E, the acceleration of gravity g, and the distance measurement, d, we are presented with 7 dimensionless variables and constants. 7 variables and 4 fundamental dimensions (M, L, T, G) gives us $7 - 4 = 3$ πs to connect them.

Though problems of this size can be analysed by hand[4, 8], an S program was written that, provided with a basis, carries out the analysis, generates a list of the expressions for the πs and then goes on to complete the regression.

The dimensional analysis yields the following dimensionless products:

1. the first represents the effect of energy use and weight (mass), perhaps due to rolling or frictional resistance:

$$\pi_1 = \frac{mwg}{E} \tag{1}$$

2. the second represents the effect of Horsepower and weight (mass) on acceleration:

$$\pi_2 = \frac{h}{wd^{0.5}g^{1.5}} \tag{2}$$

3. the third shows that q is indeed a surrogate for acceleration. This represents the distance versus time relationship, where acceleration is measured as a ratio to g:

$$\pi_3 = \frac{qg^{0.5}}{d^{0.5}} \tag{3}$$

3.5 Regression with the new dimensionless products

DA by itself cannot get us further. In the absence of further physical or engineering input, we must leave the final determination of the relationship between these factors and the other dimensionless factors of the original problem to the usual methods of regression. In terms of the products we have:

$$\frac{mwg}{E} = \phi\left(\frac{h}{wd^{0.5}g^{1.5}}, \frac{qg^{0.5}}{d^{0.5}}, \cdots\right) \tag{4}$$

If we are interested in a forecast of the m we would expect a relation of the form:

$$m = \frac{E}{wg}\phi\left(\frac{h}{wd^{0.5}g^{1.5}}, \frac{qg^{0.5}}{d^{0.5}}, \cdots\right) \tag{5}$$

where the ... represent the list of dimensionless variables such as $DRAT$ and CYL in the observations.

Based on the data analysis, Henderson and Velleman[2] noted that it might be sensible to use m^{-1} as the dependent variable to get a linear relationship with w. Our analysis confirms this observation. Alternatively we could use mw as a variable if we do not know the dimensional constant g/E.

Regression with the πs will bring in the natural nonlinearities of the problem but it will also bring in an alternative error structure. Minimising the sum of squared deviations about mwg/E is not the same as the (assumed) original problem of deviations about m. Of course, any transformation of the variables would have a similar effect.

4 Conclusion

The DA literature is replete with warnings about the inappropriate use of the technique. First one should not make the mistake of applying it to situations in which even the physical laws involved are unknown. Taylor[8] remarks that such attempts “give the subject the reputation of being a black art, which in these circumstances it is indeed.”

Though the conversion from the original variables to a set of πs is automatic and programs have been written to do this easily, the initial consideration of the problem is important. DA cannot invent variables and can only deal with the variables it is presented with.

How we choose the basis and hence the particular set of πs is also important. Though their number is fixed, the set of πs can be manipulated by multiplying them together or dividing one by another to produce new, but still dimensionless, factors. These operations correspond to changing basis in the exponent space of the original variables.

We therefore have the option of choosing the best set of πs. One criterion is essentially an aesthetic one - good products are those that have a physical interpretation. For example, one factor may represent the balance of two important effects. Many named dimensionless products, such as Reynold's number, are of this kind.

Alternatively we can choose a set that forms the simplest model when the regression is done. This is an appeal to Ockham's razor. We might be lucky in that such a model would have a clear physical meaning.

We are, of course, not limited to linear regression in the πs. We would, however, hope that a good fitted model would be simple in that it is either linear or of simple polynomial form. Even here there is some freedom of choice for the analyst and more sophisticated but still simple models might be tested. Kasprzak[5, p. 28] give an example of a search for such a simple model.

Using another criterion, we might use a non-linear programming algorithm to minimise the residuals not only by fitting the regression parameters but also, by changing the basis, the πs to be used in it. Work is continuing on ways of doing this.

While there is often a tendency to regard the data as mere numbers, to throw them into a computer package and then try to interpret the results, good statisticians recognise the need for careful preliminary investigations. I believe that Dimensional Analysis will prove to be a useful addition to the set of tools available for this purpose.

References

[1] E Buckingham. On physically similar systems; illustrations of the use of dimensional equations. *Phys.Rev.*, 4:345–76, 1914.

[2] Harold V Henderson and P E Velleman. Building multiple regression models interactively. *Biometrics*, 37:391–411, June 1981. (discussion 38, 511-516, June 1982).

[3] R R Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32:1–49, March 1976.

[4] H E Huntley. *Dimensional Analysis.* Dover Publications, New York, 1967.

[5] Waclaw Kasprzak, Bertold Lysik, and Marek Rybaczuk. *Dimensional Analysis in the Identification of Mathematical Models.* World Scientific, 1990.

[6] Eliezer Naddor. Dimensions in operations research. *Operations Research*, 14:508–514, 1966.

[7] J W S Rayleigh. The principle of similitude. *Nature*, 95(66):591 and 644, 1915.

[8] Edward S Taylor. *Dimensional Analysis for Engineers.* Clarendon Press, Oxford, 1974.

[9] G A Vignaux. Dimensional analysis in operations research. *NZ Operations Research*, 14(1):81–92, Jan 1986.

[10] G A Vignaux and Sudha Jain. An approximate inventory model based on dimensional analysis. *Asia-Pacific Journal of Operational Research*, 5(2):117–123, November 1988.