Dimensional Analysis and Its Applications in Statistics

WEIJIE SHEN

The Pennsylvania State University, University Park, PA 16802, USA

TIM DAVIS

We Predict Ltd., Technium 1, Kings Road, Swansea, SA1 8PH, UK

DENNIS K. J. LIN

The Pennsylvania State University, University Park, PA 16802, USA

CHRISTOPHER J. NACHTSHEIM

University of Minnesota, Minneapolis, MN 55455, USA

Dimensional analysis (DA) is a well-developed, widely-employed methodology in the physical and engineering sciences. The application of dimensional analysis in statistics leads to three advantages: (1) the reduction of the number of potential causal factors that we need to consider, (2) the analytical insights into the relations among variables that it generates, and (3) the scalability of results. The formalization of the dimensional-analysis method in statistical design and analysis gives a clear view of its generality and overlooked significance. In this paper, we first provide general procedures for dimensional analysis prior to statistical design and analysis. We illustrate the use of dimensional analysis with three practical examples. In the first example, we demonstrate the basic dimensional-analysis process in connection with a study of factors that affect vehicle stopping distance. The second example integrates dimensional analysis into the regression analysis of the pine tree data. In our third example, we show how dimensional analysis can be used to develop a superior experimental design for the well-known paper helicopter experiment. In the regression example and in the paper helicopter experiment, we compare results obtained via the dimensional-analysis approach to those obtained via conventional approaches. From those, we demonstrate the general properties of dimensional analysis from a statistical perspective and recommend its usage based on its favorable performance.

Key Words: Buckingham's Π Theorem; Design of Experiment; Dimensions; Statistical Analysis.

1. Introduction

DIMENSIONAL ANALYSIS (DA) is a well-established method in physics (see Sonin (2001), Szirtes (2007)). Bridgman (1931) stated that "The principal use of dimensional analysis is to deduce from a study of the dimensions of the variables in any physical system certain limitations on the form of any possible

Mr. Shen is a Doctoral Student in the Department of Statistics, Pennsylvania State University. His email address is weijie.shen@psu.edu.

Dr. Davis is the Director of Timdavis Consulting Ltd., Chief Technical Officer at We Predict Ltd., and Professor at the University of Warwick. His email address is tim@timdavis.co.uk.

Dr. Lin is a Distinguished Professor of Statistics and Supply Chain in the Department of Statistics, Pennsylvania State University. His email address is dkl5@psu.edu.

Dr. Nachtsheim is a Professor, the Frank A. Donaldson Chair of Operations Management, and Chair of the Supply Chain and Operations Department in the Carlson School of Management, University of Minnesota. His email address is nacht001@umn.edu.

relationship between those variables. The method is of great generality and mathematical simplicity". It is mainly used to find the relations among physical quantities in complicated physical systems by their dimensions. A variety of literature has applied dimensional analysis in various fields. See Asmussen and Heebooll-Nielsen (1955), Islam and Lye (2009), and Stahl (1962), for examples. Through these analyses, some simple rules among those quantities can be extracted. As a dimension-reduction and feature-extraction methodology, dimensional analysis could be of great use to the field of statistics. This *a priori* analysis gives us a conceptual and analytical view of the problem we are dealing with, thereby providing guidance in both the design and analysis steps. Furthermore, the physical origin of dimensional analysis improves the interpretability of the final results, which is particularly desirable to the fields of physics and engineering.

Unfortunately, statisticians seem to have overlooked the advantages of dimensional analysis. Finney (1977) commented that "I am surprised by the lack of attention given to dimensions as a check on the theory and practice of statistics. The basic ideas, readily appreciated, should form part of the stock-in-trade of every statistician". In this paper, we focus on building functional relationships between inputs and outputs. We will first introduce the basic concept and general procedure of dimensional analysis. Illustrated by two examples, the pine tree and paper helicopter, we show how to apply dimensional analysis in real problems and also compare the results with classic approaches. We summarize the properties and discuss the advantages of dimensional analysis in the statistical context.

The rest of the paper is organized as follows. In the Section 2, we introduce the definitions and general procedures of dimensional analysis with an illustrative example. In Section 3, we use dimensional analysis for data analysis and show its generality and importance. In Section 4, dimensional analysis is applied to design of experiments. The last two sections summarize general properties, followed by some concluding remarks and prospective research.

2. Definitions of Dimensional Analysis and General Procedure with Illustrative Example

2.1. Physical Dimensions

In mathematics, dimension typically refers to the number of coordinates required to define points in abstract spaces, whereas in statistics, dimension typically refers to the number of variables in a design problem or a data set. However, physical dimensions refer to the measurement systems to characterize certain objects. Each physical dimension has several empirical scales of the measurements and they are called "units". Ignoring nuclear effects such as isospin, charm, and strangeness, there are seven fundamental physical dimensions: namely, mass M, length L, time T, temperature Θ, electric current I (or charge Q), amount of substance mol, and luminous intensity I_v. The corresponding units, defined by SI (International System of Units), are kilogram, meter, second, kelvin, ampere, mole, and candela, respectively. All other physical quantities are combinations of these fundamental quantities and their units are combinations of the units of the corresponding fundamental quantities, combined in the same way. For example, speed has the dimension of length per time, for which the SI unit is meters per second.

2.2. Background of Dimensional Analysis

Physical quantities cannot be constructed unrestrictedly. For example, it makes no sense to add "length" to "mass" due to the natural constraints in the physical quantities. The main constraint is that "a physical law must be independent of the units used to measure the physical quantities". This was first proposed by Joseph Fourier in the 19th century (see Mason (1962)). This principle has been formalized in two important theorems, Buckingham's Π-theorem (Buckingham (1914, 1915a,b)) and Bridgman's principle of absolute significance of relative magnitude (Bridgman, 1931). Buckingham's Π-theorem shows that physical equations must be dimensionally homogeneous. In other words, any meaningful equations (and inequalities) must have the same dimensions in both the left and right sides. Bridgman's principle of absolute significance of relative magnitude shows that such formulae should be in the power-law form. Basically, Bridgman's principle allows us to transform physical quantities properly, especially into dimensionless forms. The method of using dimensionless quantities and Buckingham's Π-theorem to remove such constraints is called dimensional analysis. Next, we introduce Buckingham's Π-theorem and how to use it in practice.

2.3. General Procedure

We recommend applying dimensional analysis before statistical analysis to give a general view of the problems and the variables involved. From the physi-

cal perspective, procedures can be found in Taylor et al. (2007) and others. From the statistical perspective, the general procedure of dimensional analysis can be specified as follows:

Step 1. Determine the input and output variables and their dimensions, respectively.

Step 2. Determine the basis quantities.

Step 3. Transform input and output variables into dimensionless quantities by using basis quantities in step 2.

Step 4. Re-express the model functions via transformed variables in step 3.

Step 1

Determine the input and output variables of the system we consider. Denote input variables as $Q_1, \ldots, Q_p$ and the output variable (response) as Q_0. The conventional model will be $Q_0 = f(Q_1, \ldots, Q_p)$, where f is the model function to be estimated. Note that, in dimensional analysis, Q_i may include relevant physical constants with dimensions, such as gravitational constant $= 6.67300 \times 10^{-11} \mathrm{m}^3 \mathrm{kg}^{-1} \mathrm{s}^{-2}$. The units are often standardized to avoid dimensionless multiplicative constants. But standardization is not always necessary because these constants are combined into unknown functional relationships. After checking the physical meaning of all the variables we consider, we determine the relevant fundamental physical quantities in the system of the seven SI units as shown in the previous section: denote them as $q_1, \ldots, q_7$. Further denote the dimensions of Q_i as $[Q_i]$ and q_j as $[q_j]$ for $i = 0, 1, \ldots, p$, $j = 1, \ldots, 7$. We express the dimensions of $Q_0, Q_1, \ldots, Q_p$ in terms of $[q_1], \ldots, [q_7]$ as $[Q_i] = [q_1]^{r_{i1}} \cdots [q_7]^{r_{i7}}$, for some proper choices of $\{r_{ij}\}$ with $i = 0, 1, \ldots, p$, $j = 1, \ldots, 7$.

Step 2

Determine the basis quantities. The basis quantities constitute a subset of the inputs. We reorder and denote them as $Q_1, \ldots, Q_t$, where $t \leq 7$ as discussed above and $t \leq p$. The basis quantities should satisfy two conditions: (1) "Representativity": the dimensions of any other quantities, $[Q_0], [Q_{t+1}], \ldots, [Q_p]$ can be expressed by the combinations of the dimensions of the basis quantities, $[Q_1], \ldots, [Q_t]$. The combinations take the form of power law. (2) "Independence": the dimension of any basis quantities cannot be expressed by the combinations of the dimensions of other basis quantities. Furthermore, assume that $[Q_0]$ can be expressed by the combinations of $[Q_i]$,

$i = 1, \ldots, p$. If not, dimensional homogeneity is violated. This assumption leads to the existence of the basis quantities but they are not unique. However, the number of basis quantities is a fixed constant. The concept of basis in linear algebra is a very good analogy to the concept of basis quantities.

Step 3

Transform input and output variables into dimensionless quantities by using basis quantities. We mainly transform variables that are not basis quantities, i.e., $Q_0, Q_{t+1}, \ldots, Q_p$, based on Buckingham's Π-theorem. Due to the two properties of basis quantities in step 2, we can have $[Q_i] = [Q_1]^{d_{i1}} \cdots [Q_t]^{d_{it}}$, $i = 0, t+1, t+2, \ldots, p$. Consequently, the transformed dimensionless quantities are $\Pi_i = Q_i \cdot Q_1^{-d_{i1}} \cdots Q_t^{-d_{it}}$, $i = 0, t+1, t+2, \ldots, p$, because

$$
\begin{aligned}
[\Pi_i] &= [Q_i Q_1^{-d_{i1}} \cdots Q_t^{-d_{it}}] \\
&= [Q_1]^{d_{i1}} \cdots [Q_t]^{d_{it}} [Q_1]^{-d_{i1}} \cdots [Q_t]^{-d_{it}} \\
&= 1.
\end{aligned}
$$

Step 4

Re-express the response functions. Before using dimensional analysis, we have $Q_0 = f(Q_1, \ldots, Q_t, Q_{t+1}, \ldots, Q_p)$. Using Π_i instead of Q_i, we have the following expression:

$$
\begin{aligned}
\Pi_0 Q_1^{d_{01}} &\cdots Q_t^{d_{0t}} \\
&= f(Q_1, \ldots, Q_t, \Pi_{t+1} Q_1^{d_{t+1,1}} \cdots Q_t^{d_{t+1,t}}, \ldots, \\
&\qquad \Pi_p Q_1^{d_{p1}} \cdots Q_t^{d_{pt}})
\end{aligned}
$$

and

$$
\begin{aligned}
\Pi_0 &= Q_1^{-d_{01}} \cdots Q_t^{-d_{0t}} \\
&\quad \times f(Q_1, \ldots, Q_t, \Pi_{t+1} Q_1^{d_{t+1,1}} \cdots Q_t^{d_{t+1,t}}, \ldots, \\
&\qquad \Pi_p Q_1^{d_{p1}} \cdots Q_t^{d_{pt}}),
\end{aligned}
$$

where f is the function we hope to estimate. So we can rewrite it as

$$
\Pi_0 = g(Q_1, \ldots, Q_t, \Pi_{t+1}, \ldots, \Pi_p),
$$

where $\Pi_i, i = 0, t+1, \ldots, p$ are dimensionless and $Q_1, \ldots, Q_t$ are "independent". Buckingham's theorem indicates that $Q_1, \ldots, Q_t$ should not be in the formula. This implies $\Pi_0 = g(\Pi_{t+1}, \ldots, \Pi_p)$ to be the final model.

2.4. Example: Vehicle Stopping Distance

Here we use the vehicle stopping distance example to illustrate the general procedure. In the experiment, we estimate the stopping distance for cars, a

key indication of their safety. Assume that the driver requires a certain amount of time for reaction to an emergency and that the wheels are not locked when braking. We show the dimensional analysis below using the procedure described in the previous section.

Step 1. Identify input and output variables and their dimensions as follows:

$Q_0 = D$: vehicle stopping distance $[D] = \mathrm{L}$.

$Q_1 = v$: velocity of the vehicle $[v] = \mathrm{LT}^{-1}$.

$Q_2 = \tau$: thinking time $[\tau] = \mathrm{T}$.

$Q_3 = m$: mass of the car $[m] = \mathrm{M}$.

$Q_4 = F$: braking force on the brake discs $[F] = \mathrm{MLT}^{-2}$.

$Q_5 = \mu$: friction coefficient of brakes $[\mu] = 1$.

The response is $\{D\}$ and the predictors are $\{v, \tau, m, F, \mu\}$. The model function is $D = f(v, \tau, m, F, \mu)$. The dimensions of respective variables are listed above and we summarize corresponding $\{r_{ij}\}$ in Table 1. There are three fundamental quantities in this system. Their dimensions are length L, time T, and mass M. The entries in the table are the power of the fundamental dimensions (denoted by rows) for dimensions of each variable (denoted by columns).

Step 2. Determine the basis quantities. We choose $\{Q_1 = v, Q_2 = \tau, Q_3 = m\}$ in this case.

Step 3. Determine the dimensionless transformation for the remaining three quantities $\{Q_0 = D, Q_4 = F, Q_5 = \mu\}$, and formulate $\{\Pi_0, \Pi_4, \Pi_5\}$ as follows:

$$[v\tau] = \mathrm{L}, \quad [\tau] = \mathrm{T}, \quad [m] = \mathrm{M}, \quad [\mu] = 1$$

and

$$\Pi_0 = \frac{D}{v\tau}, \quad \Pi_4 = \frac{F\tau}{mv}, \quad \Pi_5 = \mu.$$

Step 4. Re-express the model. Here our objective will be to estimate the function g, with Π_0 as the response variable and only two input variables, Π_4 and Π_5:

$$\Pi_0 = g(Q_1, Q_2, Q_3, \Pi_4, \Pi_5)$$
$$= g(\Pi_4, \Pi_5),$$

or equivalently,

$$\frac{D}{v\tau} = g\left(\frac{F\tau}{mv}, \mu\right).$$

TABLE 1. Dimensions of Variables in the Vehicle Stopping Distance

Dimension	D	v	τ	m	F	μ
L (length)	1	1	0	0	1	0
T (time)	0	−1	1	0	−2	0
M (mass)	0	0	0	1	1	0

From the procedure and the example, we can see that, in physical phenomena, we have certain restrictions in the forms of f, satisfying certain dimensional requirements. After dimensional analysis, the potential effects on responses are attributable to the combinations of quantities considered. These quantities act like groups. If we base our estimated function g on the group values, we do not have dimensional restrictions.

3. Dimensional Analysis for Data Analysis

Statistics extracts information from the data of experiments to find or justify properties, laws, and performance. Based on the fact that those experiments are results of physical phenomena and the data are physical quantities measured in experiments, it is often justifiable and beneficial to perform dimensional analysis in the first place. For data analysis, the advantages of using dimensional analysis stand out clearly: it is rather straightforward. Incorporating dimensional analysis only transforms the data in a predetermined fashion; the basic development strategies remain the same. Furthermore, it is supported by physics to make dimensionless transformations. We conjecture that making variables dimensionally independent is helpful in making them statistically independent, although this important issue needs further investigation. After dimensional analysis, the information from inputs and outputs is more concentrated, leading to statistical models with fewer variables and simpler analysis. Below, we compare data-analysis procedures with and without dimensional analysis based on the pine tree data from Bruce and Schumacher (1935, p. 226). From there, we show how to perform dimensional analysis for data analysis and its potential benefits.

3.1. Pine Tree Example

The pine tree data has been used by various authors to illustrate the use of diagnostics and transfor-

mation methods in linear regression (see, e.g., Atkinson (1994)). The data arise from the measurements of 70 shortleaf pine trees. Three measurements of interest here are d, the diameter of the tree in feet taken at "breast height" above the ground; h, the height of the tree in feet; and v, the volume in cubic feet. The objective of the analysis is to establish a relationship between the volume v and the variables d and h. In other words, we hope to predict the volume of a tree from its known diameter and height. The complete data set is given in the Appendix.

3.2. Regression Method Without Dimensional Analysis for Pine Tree Data Set

The conventional linear regression assumes that

$$v_i = \alpha + \beta_1 d_i + \beta_2 h_i + \epsilon_i, \qquad (1)$$

with $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$. It gives us the following estimated function with standard errors of estimations in subscripts:

$$\hat{v} = -45.3_{(5.0)} + 77.2_{(5.9)}d + 0.12_{(0.11)}h, \qquad (2)$$

with $\hat{\sigma} = 9.87$. The coefficient of h is not significant. Although a univariate analysis makes the residuals appear to be reasonably normally distributed with constant variance, we notice that (i) because the data (see Appendix) are ordered by diameter d, Figure 1 shows a distinct trend of residuals relative to the diameters; and (ii) there is a potential outlier of tree #70 (the largest tree). Both of these diagnostics strongly suggest that model (2) is inadequate. We could proceed by applying the log transformation to all variables before the linear model fitting, which leads to the following result:

$$\widehat{\ln(v)} = -1.06_{(0.24)} + 1.943_{(0.038)} \ln(d) \\ + 1.054_{(0.055)} \ln(h),$$

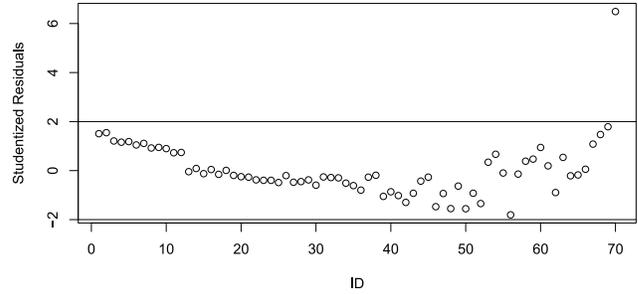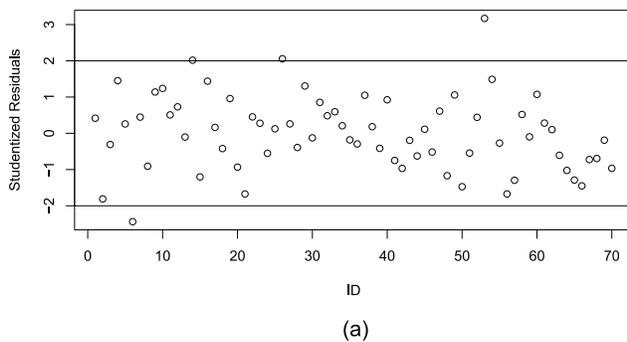with $\hat{\sigma} = 0.0673$. It is appealing to approximate the

FIGURE 1. Studentized Residuals Plot of Model (1).

coefficients of diameter and height effects (1.943 and 1.05) by the integers 2 and 1, respectively. After fixing the coefficients, we obtain a regression model on the original scale without intercept:

$$v_i = \beta d_i^2 h_i \cdot \delta_i, \qquad (3)$$

with $\epsilon_i = \ln \delta_i \sim$ i.i.d. $N(0, \sigma^2)$. The estimated regression function is

$$\hat{v} = 0.4411_{(0.0081)} d^2 h. \qquad (4)$$

On the other hand, the Box–Cox transformation on the response variable recommends a transformation parameter of $\hat{\lambda} = 0.384$. Thus, a cubic-root transformation ($\lambda = 1/3 = 0.333$) seems appropriate. The use of the cubic-root transformation of the response (without intercept) suggests the following linear model:

$$v_i^{1/3} = \beta_1 d_i + \beta_2 h_i + \epsilon_i, \qquad (5)$$

with $\epsilon_i \sim$ i.i.d. $N(0, \sigma^2)$. The estimated regression function turns out to be

$$\widehat{v^{1/3}} = 2.084_{(0.044)}d + 0.01471_{(0.00059)}h. \qquad (6)$$

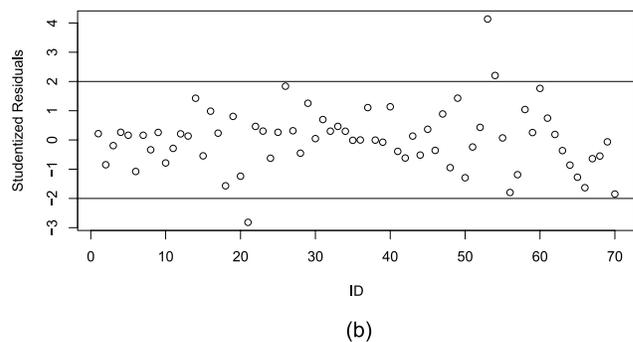The residual plots in Figures 2a and 2b show that

(a)

(b)

FIGURE 2. Studentized Residuals Plot of Models (4) in (a) and (6) in (b).

both log transformation and Box–Cox transformation have fixed the problems highlighted in Figure 1. Note that tree #53 deserves special attention, but will not be further studied in this paper.

The preceding analyses were, of course, conducted without using dimensional analysis. However, a further look at equations (4) and (6) reveals that both methods provide dimensionally homogeneous solutions to the prediction problem. The physical dimensions are coherent in the equations. Model (4) has the dimension of cubic length on both sides while model (6) has the dimension of length on both sides.

3.3. Regression Method with Dimensional Analysis for Pine Tree Data Set

Following the general procedure that we proposed in Section 2.3, the dimensional analysis can be implemented as below. A similar approach was outlined in Vignaux and Scott (1999).

1. Our objective is to predict the output volume v as a function of diameter d and height h: $v = f(d, h)$. We determine the physical dimensions of these quantities in Table 2. The dimension of v is cubic length L^3 with unit feet^3. Both dimensions of d and h are length L with units in feet.

2. Because the only dimension involved is length, let h be the basis quantity.

3. Transform other quantities into dimensionless forms,

$$\Pi_v = \frac{v}{h^3} \quad \text{and} \quad \Pi_d = \frac{d}{h}.$$

4. By Buckingham's Π-theorem, the predicted function should be $\Pi_v = g(\Pi_d)$, or equivalently,

$$\frac{v}{h^3} = g\left(\frac{d}{h}\right).$$

Suppose $\Pi_v = g(\Pi_d)$, and we choose $g(\Pi_d) = k\Pi_d^\gamma$. After taking the logarithm of both sides, we obtain the linear model,

$$\ln(\Pi_{v,i}) = \ln(k) + \gamma \ln(\Pi_{d,i}) + \epsilon_i,$$

with $\epsilon_i \sim N(0, \sigma^2)$. The estimated regression function is

$$\ln(\hat{\Pi}_v) = -1.07_{(0.16)} + 1.942_{(0.036)} \ln(\Pi_d) \qquad (7)$$

Alternatively, we might prefer that $\Pi_v = g(\Pi_d) = k\Pi_d^2$. Figure 3 shows the data and linear fits in terms of Π_v and Π_d^2. In fact, fitting a linear model yields

TABLE 2. Dimensions of Quantities in Pine Tree Data Set

Name	Quantity	Dimension	Unit
Volume	v	L^3	feet^3
Diameter	d	L	feet
Height	h	L	feet

$$\hat{\Pi}_v = 0.4363_{(0.0036)}\Pi_d^2,$$

i.e.,

$$\hat{v} = 0.4363_{(0.0036)}d^2h. \qquad (8)$$

The result is similar to the log-transformed model (4) and it also accommodates the problematic 70th case. The differences in parameter estimates can be attributed to differences in associated error structures.

In summary, certain assumptions regarding the form of the function g will give corresponding parameterizations and results. If $\Pi_v = k\Pi_d^\gamma$, then $v = kd^\gamma h^{3-\gamma}$. If $\Pi_v = (A\Pi_d + B)^3$, then $v^{1/3} = Ad + Bh$. These are exactly the two models previously obtained in equations (4) and (6). The dimensionally homogeneous results we derived from regression analyses are merely special cases of choosing different functions g after dimensional analyses. The procedure ensures that the results are dimensionally homogeneous and intuitively interpretable, while leaving choice of the function g to the investigator, as informed by the data. It gives us a guide of how to model in an efficient and parsimonious way based on the physical laws.
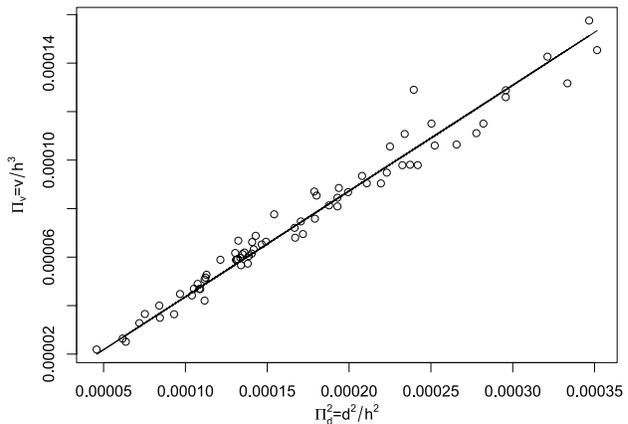

FIGURE 3. Plot of Pine Tree Data Set in Terms of Dimensionless Variables.

In contrast with the analysis and discussion of the same dataset by Atkinson (1994), dimensional analysis points to an "automatic" transformation of the data, without prior assumption of the cone shape, and posterior diagnosis of "transform both sides model" as Atkinson (1994) did. Moreover, because the dimensional analysis of each tree is the same, there is less need to worry about individual trees influencing the choice of transformation, and hence regression methods such as constructed variable plots, which Atkinson (1994) used, are no longer necessary.

3.4. Remarks

From an analytic perspective, dimensional analysis offers several advantages relative to the conventional procedures. First, it decreases the number of variables, which may lead to a simpler model. Second, physical independence may establish a simpler statistical relationship. Third, it gives more sensible interpretations by having dimensionless variables and coefficients. Physicists and engineers often favor dimensionless coefficients as indices for describing systems. Fourth, dimensional analysis produces scalable results, which is necessary for extrapolation, although the scalability depends on a good choice of the model. This is attributed to the ratio form of dimensionless variables that is invariant to scale changes. Often, extrapolating in the original scale results in values interpolating in the transformed scale. Fifth, it captures the inherent nonlinear relationship between physical quantities.

From a practical perspective, dimensional analysis is applicable when modeling physical relationships. It is a straightforward method before collecting data and modeling. It fits all kinds of data structures and modeling requirements. Furthermore, dimensional analysis does not lose generality when transforming the data. As shown above, it does not constrain the forms of estimating functions for dimensionally homogeneous solutions. Due to its analytical nature, the proposed procedure is easy to implement.

4. Dimensional Analysis for the Design of Experiments

Dimensional analysis can also serve as a guidance in the design of experiments. For the design of experiments, incorporating dimensional analysis can significantly improve efficiency by reducing the number of experimental variables. In this section, we demonstrate the use of dimensional analysis in the design and analysis of the popular paper helicopter experiment. We also compare the results to those obtained using the conventional design of experiments approaches.

4.1. The Paper Helicopter Experiment

The paper helicopter experiment is a widely used teaching device for the design of experiments. The objective is to predict the "flight time" performance for a particular configuration of the helicopter dimensions in Figure 4. Upon its launch, twin blades spin
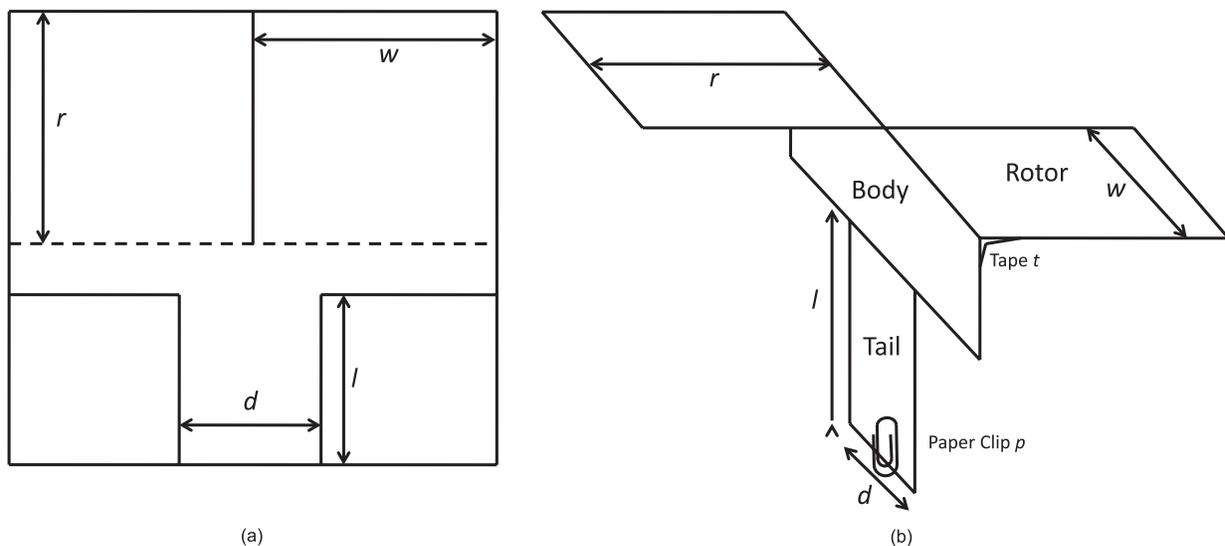

FIGURE 4. Paper Helicopter Illustration.

around the central ballast shaft to provide lift as it descends. It can be easily constructed from a sheet of paper using only scissors and tape. As displayed in Figure 4, the upper part of the model consists of two wings (or rotors). The central part is the body and the lower part is the tail folded into the ballast. The length and width of each part can be varied to achieve differing levels of performance. We wish to predict the flight time for a given configuration. The usual design factors include the rotor radius (r), the rotor width (w), the tail length (l), the tail width (d), paper clip (p), and the tape (t).

4.2. Conventional Design of the Paper Helicopter Experiment Without Using Dimensional Analysis

Johnson et al. (2006a, b) studied experimental design on the paper helicopter as part of a Six Sigma Black Belt project. They considered a Resolution VII design with seven two-level factors in a half fraction, with two replicates. This led to $2^{7-1} \times 2 \ (= 128)$ runs in total. They provided a step-by-step routine to design experiments and maximize the flight time. Box and Liu (1999) and Box (1999) discussed the use of sequential design in the paper helicopter experiment. First, they conducted a two-level Resolution IV fractional factorial design with eight factors in four replicates, i.e., $2^{8-4}_{IV} \times 4$, for a total of 64 runs. In the second

setting, they designed a full-factorial experiment involving four important factors. Two key lessons from their work, among others, are (1) with the help of response surface and steepest ascent, sequential designs search for an optimum point effectively and efficiently and (2) minimum variance or dispersion of flight time is included in addition to longest flight time to enrich the meaning of optimum. Annis (2005) derived the aerodynamics of this flying object in a rigorous physical sense before designing the experiment. He presented a physical model of flight time in terms of the length and width of wing and body. He employed two three-level factors in a single replicate of a 3×3 factorial design for wing length and width and employed response surface methods to identify the optimal operating condition. The use of physics to identify promising factors in advance turned out to be extremely advantageous.

Table 3 summarizes the results of the above three papers, including Johnson et al., (2006a,b), Box and Liu (1999), and Annis (2005). The directions of respective effects are provided in the parentheses. Three variables are included in all of the three experiments; namely, body length, body width, and wing length. Additionally, Johnson et al. (2006a, b) considered (i) paper type, (ii) whether or not taping body and wing, and (iii) whether or not clipping in the bottom. They also examined the two-way in-

TABLE 3. Summary of Literature on Paper Helicopter

	Johnson et al. (2006a,b)	Box and Liu (1999)	Annis (2005)
Input variables	Body length ($-$) Body width ($-$) Wing length ($+$) Paper type ($-$) Taped body ($-$) Taped wing ($-$) Clip ($-$) Interactions	Body length ($-$) Body width ($-$) Wing length ($+$) Paper type ($-$) Taped body ($-$) Taped wing ($-$) Clip ($-$) Fold ($-$)	Body length ($-$)* Body width ($-$)* Wing length ($+$) Wing width (dip)
Design	$2^{7-1}_{VII} \times 2 = 128$	$2^{8-4}_{IV} \times 4 = 64$	$3^2 = 9$
Feature	Interactions	Sequential learning	Physical insight
Significant variables	Body length ($-$) Body width ($-$) Wing length ($+$)	Body length ($-$) Body width ($-$) Wing length ($+$)	Body length ($-$) Body width ($-$) Wing length ($+$)

*Not a design factor.

teractions between the variables of interest. Box and Liu (1999) considered whether to fold the paper and use an additional 50 runs to search for the optimum. Their second design included a full-factorial experiment on both the length and width of both body and wing. Annis (2005) took wing width as an additional variable and chose wing length and width to be experimental variables and considered effects of body length and width as given by the physical formula. The "dip" effect of wing width means the relationship is not monotone. In general, the signs in the parentheses indicate the effects of each variable on the flight time. For example, wing length has a positive effect on the flight time, meaning that the flight time will increase if wing length is increased. All the above studies concluded that the wing length is the most important factor for determining flight time, with factors affecting helicopter mass also being important.

4.3. Design of the Paper Helicopter Experiment Using Dimensional Analysis

The physics of falling objects in a gravitational field follows the following assumptions: (1) the flight time (T) is determined by the launch height (H) and average velocity (v), i.e., $T = H/v$; (2) the falling object reaches terminal velocity quickly after dropped, when the drag force of the air becomes equal to the force of gravity; (3) the drag force depends on the density of dry air ($\rho = 1.20412$ kgm^{-1} at sea level at 20°C), drag coefficient (c_d, dimensionless) and the shape of the helicopter (rotor radius r and rotor width w, or their combinations, such as ratio r/w and area rw); (4) the weight of the helicopter depends on the mass (m) and acceleration due to gravity ($g = 9.8$ ms^{-2}).

Suppose the flight time follows the model below,

$$T = f_1(m, g, r, c_d, \rho, H).$$

Because c_d is dimensionless and $T = H/v$, the expression can be represented as

$$v = f_2(m, g, \rho, r). \quad (9)$$

Next, we apply dimensional analysis step-by-step.

Step 1. Table 4 displays the dimensions and units of the variables. We have three fundamental dimensions, length L, time T, and mass M.

Step 2. Variables r, ρ, and g are chosen as the base quantities.

Step 3. From Buckingham's Π theorem, we can reduce the number of variables from five to two. Following Gearhart (2004), we define the two dimensionless variables as $\Phi_v = vr^a\rho^bg^c$ and $\Psi_m = mr^d\rho^eg^f$. Dimensions of Φ_v and Ψ_m are, thus, as follows:

$$[\Phi_v] = (\text{LT}^{-1})(\text{L})^a(\text{ML}^{-3})^b(\text{LT}^{-2})^c$$
$$= \text{L}^{1+a-3b+c}\text{M}^b\text{T}^{-1-2c}$$
$$[\Psi_m] = (\text{ML})^d(\text{ML}^{-3})^e(\text{LT}^{-2})^f$$
$$= \text{L}^{d-3e+f}\text{M}^{1+e}\text{T}^{-2f}.$$

We enforce nondimensionality and solve the two sets of linear equations,

$$1 + a - 3b + c = 0$$
$$b = 0$$
$$-1 - 2c = 0$$

and

$$d - 3e + f = 0$$
$$1 + e = 0$$
$$-2f = 0.$$

From the first set, we obtain $a = c = -1/2$ and $b = 0$ and, from the second, $d = -3$, $e = -1$, and $f = 0$. The transformed variables are thus

$$\Phi_v = \frac{v}{\sqrt{gr}} = \frac{h}{T\sqrt{gr}}; \quad \Psi_m = \frac{m}{\rho r^3}. \quad (10)$$

Step 4. The final equation is obtained as the following form:

$$\Phi_v = g(\Psi_m). \quad (11)$$

Because there is only one input variable (Ψ_m), we conduct the paper helicopter flying experiment with four runs: $\Psi_m = 0.937$, 2.087, 3.088, and 4.642. The resulting flight times (T) are 5.18, 3.87, 3.48, and 2.98, respectively. This implies the values of Φ_v to be 0.873, 1.264, 1.537, and 1.795. The results

TABLE 4. Dimensions of Quantities of Paper Helicopter

Quantity name	Quantity symbol	Dimension	Unit
Velocity	$v = h/T$	LT^{-1}	ms^{-1}
Mass	m	M	kg
Gravity acceleration	g	LT^{-2}	ms^{-2}
Air density	ρ	ML^{-3}	kgm^{-3}
Wing length	r	L	m

TABLE 5. Table of Design and Results of Paper Helicopter Experiment

No.	Ψ_m $(m/\rho r^3)$	Paper type	Helicopter mass(m)	Rotor radius(r)	Flight time(T)	Φ_v $(h/(T\sqrt{gr}))$
1	0.937	80 g/m^2	3.09 g	140 mm	5.18 s	0.873
2	2.087	120 g/m^2	4.34 g	120 mm	3.87 s	1.264
3	3.088	100 g/m^2	3.72 g	100 mm	3.48 s	1.537
4	4.642	160 g/m^2	5.59 g	100 mm	2.98 s	1.795

Note: The results are the averages of three flights recorded independently twice.

are displayed in Table 5. Figure 5 is the scatter plot of Φ_v and Ψ_m.

The data are modeled using a simple linear regression, giving $\hat{\Phi}_v = 0.70_{0.09} + 0.25_{0.03}\Psi_m$. Converting to the original variables, this becomes

$$\hat{T} = \frac{h}{\sqrt{gr}(0.70 + \frac{0.25}{\rho r^3})}. \qquad (12)$$

Alternatively, we regress data on a log scale. This gives us $\log(\Phi_v) = -0.102_{0.01} + 0.46_{0.01}\log(\Psi_m)$. The power of 0.46 suggests a square-root transformation of Ψ_m. We thus take a square-root transformation on Ψ_m and fit a linear model without intercept, obtaining $\hat{\Phi}_v = 0.859_{0.014}\sqrt{\Psi_m}$. Converting to the original variables, this can be expressed as

$$\hat{T} = \frac{hr}{0.859}\sqrt{\frac{\rho}{mg}}. \qquad (113)$$

In both models (12) and (13), the coefficients are dimensionless and the equations are dimensionally homogeneous. We prefer model (13) because it is able to capture the potential curvature shown in Figure
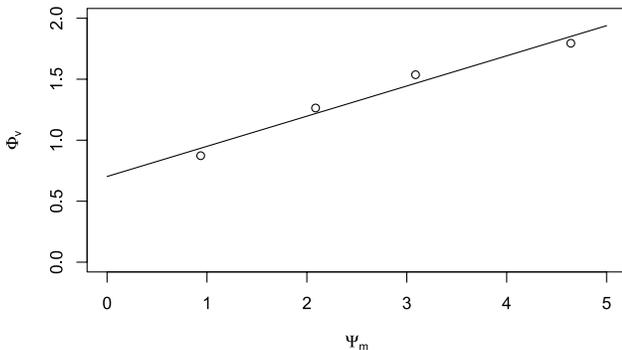
5. For validation of the final model, we conducted eight confirmation runs with various combinations of 80/100/120/160 g/m^2 A4 paper and 100/120/140 mm rotor radius. The actual flight times versus predicted flight times are displayed in Figure 6. The points align closely to the line $y = x$, indicating a good model indeed.

Table 6 provides a comparison of results between designs using dimensional analysis and those without it. It can be seen that the key factors are exactly the same, while the settings for maximal time are close. However, the design we used had only one dimensionless variable; therefore, fewer runs were needed (4 runs vs. 128 runs in Johnson et al. (2006b), 64 runs in Box and Liu (1999), and 9 runs in Annis (2005)). In addition, the prediction model is elegant and easy to interpret: the response is proportional to some
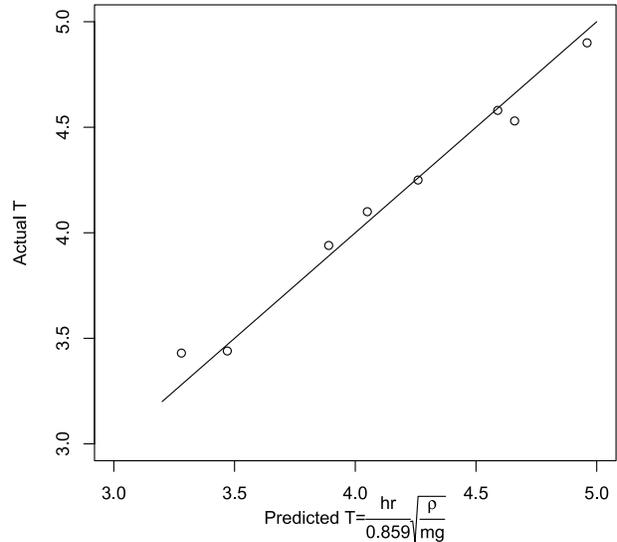

FIGURE 5. Plot of Simple Linear Regression on Paper Helicopter.

FIGURE 6. The Plot of Predicted Flight Time and Actual Flight Time of the Confirmation Runs.

TABLE 6. Comparison Between Designs with and Without DA

	Without DA	With DA
Variables	2 or 3 Levels	Continuous (interpolate and extrapolate)
Design	On variables	On dimensionless quantities
Number of experimental runs	128, 64 and 9	4
Key variables	Wing length (+) Body length (−)	Wing length (+) Body length (−)
Settings for maximal flight time	L = 15.2c m, w = 4.60 cm m = A4 (gsm unstated)	L = 14 cm, w = 5.25 cm m = 3.09 g (80 gsm)
Validation model	Full factorial	Confirmation runs

power of each factor. The ratio form also produces scalable results. Furthermore, after transformation, the dimensionless quantities seem to be linearly related to the response, which avoids the complicated relationships such as "dip" effect found in Box and Liu (1999) and Annis (2005). Last, robustness analysis through transmitted variation (Morrison (1957)) can be directly applied to the fitted model, avoiding costly replications to determine this variation empirically. Further results on a dimensional-analysis approach to the paper helicopter can be found in Davis (2013).

4.4. Remarks

From the design of experiments perspective, the number of experimental runs required tends to increase with the number of experimental factors. Dimensional analysis combines variables according to physical laws and creates new design variables that can be incorporated into the design. It reduces the number of factors and consequently reduces the required number of runs. It potentially allows the inclusion of variables not included in the original experiment, as long as they can be expressed in the dimensionless variables. We also benefit from the scalability and the interpretability of the solutions. See, for example, Albrecht et al. (2013) for a detailed treatment of designing experiments to dimensional-analysis models.

5. General Properties

From the above examples, we summarize the following advantages for dimensional analysis on general cases.

1. Dimensional analysis starts from basic and natural physical assumptions. The resulting factors and their coefficients are dimensionless and easy to interpret for practitioners.

2. Dimensional analysis combines and eliminates unnecessary variables using Buckingham's Π Theorem. This leads to dimension reduction, which is especially helpful for the design of experiments.

3. The power law is used in the combination, revealing that the nature of relationships between variables with different dimensions is often not linear. It is believed that, after transformation, the dimensionless quantities may become more independent (fewer interactions) and their relationships with response are simpler, as shown in our examples.

4. It is compatible with all kinds of methods, as it is a "data-free" method. It transforms the variables according to their dimensions, not values. It can be done even before we get the real data or any pilot experiment. The subsequent statistical procedures are valid without changes.

5. The resulting models are often scalable. For example, extrapolation in linear regression is often misleading, because the assumed form of the model may not be valid beyond the range of the data. The dimensionless models developed using dimensional analysis do not depend on absolute quantities, but are rather defined in terms of relative amounts; thus, scale is not relevant in most cases.

Drawbacks of dimensional analysis may include the requirement of physical knowledge about the experimental environment and the possibility of severe problems if any related variable is excluded (Albrecht et al. (2013)). Piepel (2013) also raises the issue of spurious correlation. A comprehensive discussion of statistical issues of dimensional analysis is presented in Lin and Shen (2013).

Conclusion

Dimensional analysis has been well developed in physics, engineering, and other fields. However, its significance was overlooked for years by statisticians. Little effort was made to incorporate it into statistical practice. In this paper, we describe the use of the dimensional-analysis method for both data analysis and design of experiments. Additional examples and comments can be found in Davis (2011) and Lin and Shen (2013). Our purpose is to promote greater integration of dimensional analysis into statistical design and analysis by statisticians, engineers, and scientists.

The fundamental insight of dimensional analysis is to identify key dimensionless variables from physical considerations, and then use data and statistical analysis to understand them. Engineers provide theory for guidance in the statistical analysis using physical prior knowledge. Statisticians design and analyze experiments for the unknown physical structure, check the validity of physical assumptions and recommend further experiments. Complementing each other in this fashion often leads to solutions that neither could achieve alone.

This paper introduces the basic idea of dimensional analysis and its potential applications in statistics, notably in regression analysis and the design of experiments. There are many more issues to be studied. First, the error structure should be further investigated for dimensional analysis. Latent errors in covariates and the robustness need to be taken into account as well. For example, an orthogonal distance regression can be applied when both sides of the modeling equation have errors. Moreover, errors could propagate very differently for different nonunique DA representations. Second, once designing the transformed dimensionless quantities, the corresponding design on the original quantities is not unique. The various design options on those operating quantities offer a way to test the validity of Buckingham's Π-theorem statistically. Third, a formal sequential or recursive scheme is suggested to

interweave the knowledge of engineers and that of statisticians. Special designs or analyses may be preferred after conducting dimensional analysis. Fourth, we believe dimensional analysis could be generalized into fields outside of physics and engineering. Certain common measure units in economics, biology, or sociology could be candidates to enlarge the application of dimensional analysis. Fifth, it seems promising to generalize the idea of combining variables. PCA is one kind of combining under linear schemes in feature extraction. Dimensional analysis implies that combinations may be done nonlinearly by power law.

References

ALBRECHT, M. C.; NACHTSHEIM, C. J.; ALBRECHT, T. A.; and COOK, R. D. (2013). "Experimental Design for Engineering Dimensional Analysis". *Technometrics* 55(3), pp. 257–270.

ANNIS, D. H. (2005). "Rethinking the Paper Helicopter: Combining Statistical and Engineering Knowledge". *The American Statistician* 59(4), pp. 320–326.

ASMUSSEN, E. and HEEBOOLL-NIELSEN, K. (1955). "A Dimensional Analysis of Physical Performance and Growth in Boys". *Journal of Applied Physiology* 7(6), pp. 593–603.

ATKINSON, A. C. (1994). "Transforming Both Sides of a Tree". *The American Statistician* 48(4), pp. 307–313.

BOX, G. E. P. (1999). "Statistics as a Catalyst to Learning by Scientific Method Part II—A Discussion". *Journal of Quality Technology* 31(1), pp. 16–29.

BOX, G. E. P. and LIU, P. Y. T. (1999). "Statistics as a Catalyst to Learning by Scientific Method Part I—An Example". *Journal of Quality Technology* 31(1), pp. 1–15.

BRIDGMAN, P. (1931). *Dimensional Analysis*, 2nd edition. Yale University Press.

BRUCE, D. and SCHUMACHER, F. X. (1935). *Forest Mensuration*, p. 226. New York, NY: McGraw Hill.

BUCKINGHAM, E. (1914). "On Physically Similar Systems; Illustrations of the Use of Dimensional Equations". *Physical Review* 4(4), pp. 345–376.

BUCKINGHAM, E. (1915a). "Model Experiments and the Forms of Empirical Equations". *Transactions of the American Society of Mechanical Engineers* 37, pp. 263–296.

BUCKINGHAM, E. (1915b). "The Principle of Similitude". *Nature* 96(2406), pp. 396–397.

DAVIS, T. P. (2011). "Dimensional Analysis in Experimental Design". Talk at Isaac Newton Institute. Available online at http://sms.cam.ac.uk/media/1191742/.

DAVIS, T. P. (2013). "Comment: Dimensional Analysis in Statistical Engineering". *Technometrics* 55(3), pp. 271–274.

FINNEY, D. J. (1977). "Dimensions of Statistics". *Applied Statistics* 26(3), pp. 285–289.

GEARHART, C. (2004). "Using Dimensional Analysis to Build a Better Transfer Function". SAE Technical Paper, 2004–01–1129.

ISLAM, M. and LYE, L. M. (2009). "Combined Use of Dimensional Analysis and Statistical Design of Experiment Methodologies in Hydrodynamics Experiments". *Ocean Engineering* 36, pp. 237–247.

JOHNSON, J. A.; WIDENER, S.; GITLOW, H.; and POPVICH, E. (2006a). "A 'Six Sigma' Black Belt Case Study: G.E.P. Box's

Appendix

The Shortleaf Pine Tree Data Set

ID	Diameter (feet)	Height (feet)	Volume (feet^3)	ID	Diameter (feet)	Height (feet)	Volume (feet^3)
1	0.383	33	2.2	36	0.917	71	25.8
2	0.367	38	2.0	37	0.925	81	32.8
3	0.417	40	3.0	38	0.933	91	35.4
4	0.425	49	4.3	39	0.958	66	26.0
5	0.425	37	3.0	40	0.975	65	29.0
6	0.433	41	2.9	41	1.000	72	30.2
7	0.433	41	3.5	42	1.017	66	28.2
8	0.458	39	3.4	43	1.017	72	32.4
9	0.458	50	5.0	44	1.042	90	41.3
10	0.467	69	7.2	45	1.075	88	45.2
11	0.492	58	6.4	46	1.083	63	31.5
12	0.492	50	5.6	47	1.092	69	37.8
13	0.625	45	7.7	48	1.092	65	31.6
14	0.633	51	10.3	49	1.117	73	43.1
15	0.633	49	8.0	50	1.150	69	36.5
16	0.650	59	12.1	51	1.150	77	43.3
17	0.667	56	11.1	52	1.192	64	41.3
18	0.675	86	16.8	53	1.192	77	58.9
19	0.700	59	13.6	54	1.217	91	65.6
20	0.717	78	16.6	55	1.233	90	59.3
21	0.742	93	20.2	56	1.242	68	41.4
22	0.758	65	17.0	57	1.258	96	61.5
23	0.767	67	17.7	58	1.267	91	66.7
24	0.775	76	19.4	59	1.267	97	68.2
25	0.775	64	17.1	60	1.275	95	73.2
26	0.817	71	23.9	61	1.283	89	65.9
27	0.825	72	22.0	62	1.308	73	55.5
28	0.825	79	23.1	63	1.325	99	73.6
29	0.825	69	22.6	64	1.333	90	65.9
30	0.842	71	22.0	65	1.400	90	71.4
31	0.850	80	27.0	66	1.483	91	80.2
32	0.850	82	27.0	67	1.525	96	93.8
33	0.858	81	27.4	68	1.525	100	97.9
34	0.867	75	25.2	69	1.617	94	107.0
35	0.883	75	25.5	70	1.950	104	163.5

Paper Helicopter Experiment. Part A". *Quality Engineering* 18, pp. 413–430.

JOHNSON, J. A.; WIDENER, S.; GITLOW, H.; and POPVICH, E. (2006b). "A 'Six Sigma' Black Belt Case Study: G.E.P. Box's Paper Helicopter Experiment. Part B". *Quality Engineering* 18, pp. 431–442.

LIN, D. K. J. and SHEN, W. (2013). "Comments: Some Statistical Concerns on Dimensional Analysis". *Technometrics* 55(3), pp. 281–285.

MASON, S. F. (1962). *A History of the Sciences*. New York, NY: Collier Books.

MORRISON, S. J. (1957). "The Study of Variability in Engineering Design". *Applied Statistics* 6(2), pp. 133–138.

PIEPEL, G. F. (2013). "Comment: Spurious Correlation and Other Observations on Experimental Design for Engineering Dimensional Analysis". *Technometrics* 55(3), pp. 286–289.

SONIN, A. A. (2001). *The Physical Basis of Dimensional Analysis*, 2nd edition. Cambridge, MA: Department of Mechanical Engineering, Massachusetts Institute of Technology. web.mit.edu/2.25/www/pdf/DA_unified.pdf.

STAHL, W. R. (1962). "Similarity and Dimensional Methods in Biology". *Science* 137(3525), pp. 205–212.

SZIRTES, T. (2007). *Applied Dimensional Analysis and Modeling*, 2nd edition. Oxford, UK: Elsevier Butterworth-Heinemann.

TAYLOR, M.; DIAZ, A. I.; JODAR-SANCHEZ, L. A.; and VILLANUEVA-MICO, R. J. (2007). "100 Years of Dimensional Analysis: New Steps toward Empirical Law Deduction". arXiv: Physics 0709.3584v3.

VIGNAUX, G. A. and SCOTT, J. L. (1999). "Simplifying Regression Models Using Dimensional Analysis". *The Australian and New Zealand Journal of Statistics* 41(1), pp. 31–42.

$\sim$