NBER WORKING PAPER SERIES

EMPTYING THE TANK:
GETTING THE MOST OUT OF LIMITED DATA

M. Scott Taylor

Working Paper 24855
http://www.nber.org/papers/w24855

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

Thanks to A. Magesan, A. Jacobsen, J. W. Laliberte, K. Head, F. Mayer, and A. Whalley for helpful comments, and J. Taylor-McGregor for help with the title. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Emptying the Tank: Getting the most out of Limited Data
M. Scott Taylor
NBER Working Paper No. 24855
July 2018
JEL No. A2,Q0,Q4

ABSTRACT

All empirical researchers know that having more sources of variation in a dataset is valuable. What is not known is how valuable, and if the marginal value of adding another source of variation diminishes or increases. This note provides explicit answers to these questions. It defines "valuable" as the number of independent questions the data can potentially answer, and provides a surprisingly simple and useful rule that tells the researcher not only when they have "emptied the tank" of their data's valuable implications, but also the marginal value of further data collection. An illustration using home heating costs is provided.

M. Scott Taylor
IEE Canada Research Chair
Department of Economics
The University of Calgary
2500 University Drive, N.W.
Calgary, AB T2N 1N4
CANADA
and NBER
m.scott.taylor1@gmail.com

1 Introduction

Researchers are always limited by their data. And data issues are many: measurement error, limited observations, selective sampling, truncation, etc. are but a few of the common issues the empirical researcher has to surmount to proceed to analysis. This note investigates a different dimension of the data problem facing researchers – the limited sources of variation in their data. For example a dataset containing only a cross-section of GDP across countries has one source of variation; if we added industry or sectoral observations within each cross-sectional unit, we would have two; and if, for example, we had in addition observations over at least two time periods, then we would have three: country $\times$ industry $\times$ time.[1]

All empirical researchers understand that more sources of variation are useful and valuable because issues of identification and sources of variation are intimately connected. This is often reflected in the common seminar question: what source of variation are you exploiting to identify that parameter or effect?

While more sources are always good, what is less clear is how useful are they? Is it always useful to expand a dataset to include another source of variation? And further, how does this benefit diminish as our dataset becomes "wider"? This note provides explicit answers to these questions.

There is an alternative way to pose the question. In many cases, the sources of variation in a dataset are fixed and largely set by government agency procedures, administrative rules, historical accident, or, just as effectively, by steeply rising costs of further collection. In those cases, the question for the researcher is flipped on its head – it becomes how to maximize

[1]Some may prefer to substitute the term dimension for source of variation as they read. They are synonymous. I have chosen source of variation in the belief it is more familiar to most readers and clearly tied to data per se.

the information you can glean from a dataset with a given number of sources of variation.

To answer our two questions, consider the logical basis for the common seminar question mentioned above. Since the answer to any empirical question requires the identification of a parameter or effect from the data, at bottom the remark is asking what question can the data potentially answer. A successful answer links a source of variation in the dataset to the estimation of a key parameter or effect, that in turn answers the hypothetical question posed.

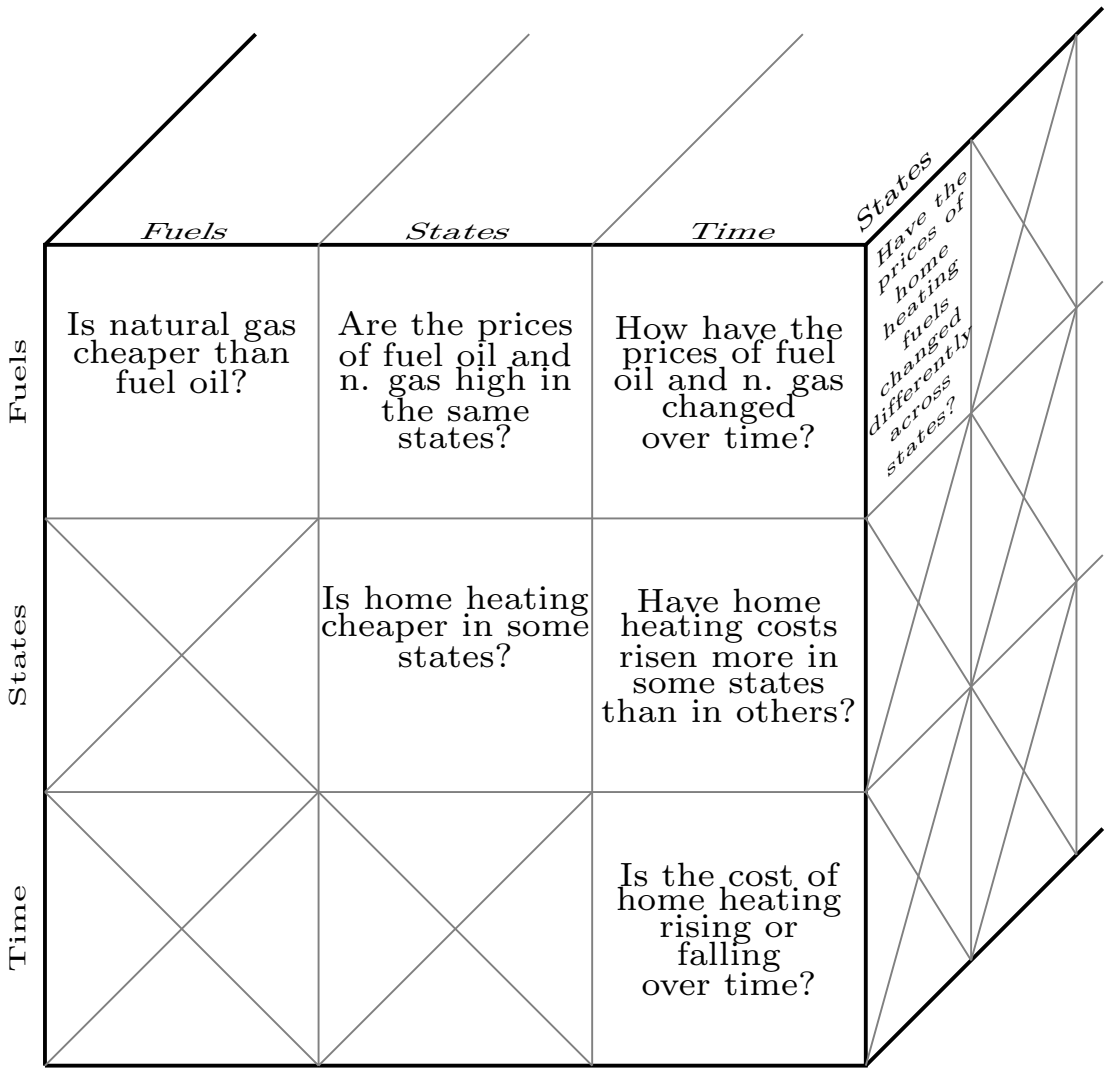
2 How many questions can you answer?

A concrete example will help fix ideas. Suppose we have a dataset with three sources of variation. We have prices for natural gas and fuel oil, across a country with many states, and we have this data collected, annually, over several years. By construction we have three sources of variation: fuel $\times$ state $\times$ time. Natural gas and fuel oil are used in home heating. The researcher sees the yearly home heating bill. When across fuel variation is available she also sees the breakdown across natural gas and fuel oil purchases. When across state variation is available she also sees the state of residence.

With this as our starting point let's ask how many independent questions we can ask of this dataset. A useful way to visualize this problem is shown in Figure 1 below. The schematic shows an $n \times n$ (3×3 in our case) face of a cube that has equal height, width, and depth. Down the side of the cube, across its face, and running through its depth are the labels for each source of variation: fuels, states, and time.

Within each face of the cube is written an example of the type of question the relevant

Figure 1



combination of variation in the data can answer. The X's signify combinations that provide no new information. Since having state × time variation is the same as having time × state variation only one of these faces carries a question. The reader should work through a couple of combinations to convince themselves of how the schematic works.

Along the diagonal, only one source of variation is exploited and the potential questions reflect this fact. So for example, if we pooled the data across states and collapsed its time

dimension, we are at the top left face and we can only ask whether on average natural gas (n. gas) is cheaper or more expensive than fuel oil. Similarly, the other diagonals lead to questions requiring only the remaining source of variation.

Moving off the diagonal implies the researcher exploits two sources of variation, and as a result the set of potential questions expands. Consequently, the entry on the second row and third column allows us to exploit across state and time variation to ask whether home heating costs have risen more in some states than others.[2]

Finally, consider the schematic's cut-away third dimension. As shown this cut-away contains all X's except for one entry. This one entry is where the researcher exploits all three sources of variation to answer a question about how the time series of home heating fuel prices across states differs (or not!).

The schematic makes it very clear how and why different sources of variation matter to empirical researchers, since it gives them the ability to pose and then potentially answer more questions of the data. Presumably the next step is to determine how these answers bear on the researcher's chosen hypothesis about home heating costs, their geographical variation or time-series properties. But the schematic also provides us with a hint as to how we could approach our question more generally. The astute observer may notice that in our simple case, with three sources of variation we can answer seven questions, and this follows because our schematic method of analysis adds up in the following way.

$$\binom{3}{1} + \binom{3}{2} + \binom{3}{3} = 7 \tag{1}$$

[2]Importantly all of these questions are independent in the sense that knowing any two answers on the face of the cube won't give you an answer to a third related question. Consider answers to questions at (2,2) and (3,3) and ask whether these determine our answer (2,3). They do not and hence our method identifies the number of independent questions the data could answer.

$\binom{3}{1}$ are the number of questions you can answer with one source of variation along the diagonal, $\binom{3}{2}$ the number of questions answered with two sources in play off the diagonal; and $\binom{3}{3}$ is the final combination that yields the final question answered by exploiting the third variation (dimension). Therefore, by construction, we know that a dataset with n sources of variation can provide, in theory, answers to M independent questions where M is given by:

$$M = \sum_{k=1}^{n} \binom{n}{k}. \tag{2}$$

Fortunately, if we recall the binomial theorem implies that

$$(1+x)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \tag{3}$$

and set $x = 1$ in (3) and recall $\binom{n}{0} = \binom{n}{n} = 1$, M becomes far simpler and equal to:

$$M(n) = 2^n - 1 \tag{4}$$

M is commonly referred to as a *Mersenne number* after the French polymath and Minim friar Marin Mersenne (1588-1648) who investigated its properties as specifically related to primes.[3]

Here our derivation of $M(n)$ shows it is simply the number of independent questions that can be posed and potentially answered by our data when it has n sources of variation. In our example $M(3) = 2^3 - 1 = 7$ questions.

[3]A Mersenne prime is a prime number one less than a power of 2. Mersenne primes are often used in cryptography.

While $M(n)$ provides a simple and clear answer to our first question, there are also different interpretations and potential uses. For example, while we arrived at the rule constructively by generalizing a purely graphical approach, the abstract minded may now recognize $M(n)$ as the Power Set of n elements minus one. Why? The Power Set of n elements gives us all the possible subsets of the n sources of variation that led to our questions; but since the empty set contains no variation, and therefore produces no questions, we need to subtract it from 2^n. Alternatively, to the very applied researcher $M(n) - 1$ is nothing more than the number of fixed effects a researcher can include in a regression framework before trivially explaining all of their data.

3 Is more always better?

With (4) in hand it is now easy to show that M is strictly convex in n (if we take n to be continuous). Alternatively, and preferably, we can take n to be an integer and note how the sequence of $M(n)$ given by $\{0, 1, 3, 7, 15, \ldots\}$ exhibits the property that

$$M(n+1) - M(n) = 2^{n+1} - 2^n = 2^n(2-1) = 2^n \tag{5}$$

and therefore we have an answer to our final question. The marginal value of adding further sources of variation to our dataset is positive, strongly increasing in n, and surprisingly equal to 2^n. Clearly from (5) this marginal value explodes with n. This implies, for our example, that the marginal value in terms of new questions we can pose of the data if we added one more source of variation is $2^3 = 8$ additional questions. And these together with those shown

in Figure 1 would allow us to ask $M(4) = 2^4 - 1 = 15$ questions in total.

Why is an additional source of variation so useful? Again our example is useful. Suppose we added to our dataset the knowledge of fuel purchase choices across those living in older homes (age $>$ 25 years) and those living in newer homes (age $\leq$ 25 years). This additional source of variation not only allows us to ask new questions about its own variation. For example, the $\binom{n}{1}$ terms in $M(n)$ rise linearly with n and allow us to now ask whether it is more expensive to heat an old house or a new house. Similarly, the new source of variation also allows us to ask a finely detailed question that exploits the new type of variation in conjunction with the earlier three. These $\binom{n}{n}$ terms also rise linearly with n, and allow us to ask, for example, how the time pattern of natural gas and fuel oil purchases across new and old homes differs across states. But the true benefit of adding a new source of variation comes from the many new combinations of variation we can now exploit to answer so so many more questions with the data. All of these new combinations exploits age in some way to ask and then answer new questions. These remaining new questions are represented by all the new $\binom{n}{k}$ terms in (2).

In our specific case, our new source of variation gives us the ability to ask in total 8 new questions. Example questions, together with the variation they exploit, are given below:

1. Are home heating costs more for older homes? (age)

2. What share of new homes are heated by natural gas? (fuel $\times$ age)

3. In states with high natural gas prices, do home owners of both old and new homes purchase less gas? (fuel $\times$ state $\times$ age)

4. How have home heating costs risen for older homes that primarily use fuel oil? (fuel

× time × age)

5. Have new homes heated exclusively by natural gas had their home heating costs rise less quickly than the national average during the shale gas boom? (state × time × fuel × age)

6. Is heating an old home less expensive than heating a new home in some states? (state × age)

7. Have newer or older houses seen home heating costs rise most in the last decade? (time × age)

8. Have home heating costs for old homes risen more in some states than others? (state × time × age)

It is now obvious that the researcher who can answer these questions has a far richer understanding of the available fuel switching possibilities, how these have varied geographically, and how they have or have not improved for all fuel and home types. What is less obvious, and perhaps surprising, is the scale of the change in information provided to the researcher. And only by knowing (4) and (5) can the diligent researcher exhaust all the possibilities and in doing so – empty the tank of all potential, but yet unasked, questions of the data.

Finally, a caveat. Although the simplicity of the rule is helpful in finding the volume of questions, the value of answering an additional question may decrease with n, or in the extreme only one question may be of value to the researcher. All of this is surely true, but (4) still plays a constructive role in providing guidelines for data collection and analysis. Some sources of variation are more useful than others, but diminishing returns in terms of

their value would need to be exceedingly severe to offset the exponential growth in volume shown by (4).

4 Conclusion

This note asked how valuable is further information by focusing on one dimension of the problem – the number of sources of variation in a dataset. While empirical researchers have always known more variation is better, I have shown exactly why this is true and provided a simple rule linking the sources of variation to new questions that can be answered. Surprisingly, the marginal value of additional sources of variation is increasing and not diminishing. With this knowledge in hand, empirical researchers can squeeze all the potential out of their current datasets – emptying the tank, while referees and advisors can ask how an author's theory matches with the perhaps unnoticed or conveniently ignored remaining $\binom{n}{k}$ implications of the data they employed. And theorists can now turn to examine whether the empirical researcher should prefer more data of given types to further sources of variation (i.e. dimensions). For example, if the researcher has access to N observations in total, and these could be split across $M \leq N$ sources of variation there is clearly a trade-off between the number of questions posed (increasing in M), and the ability to answer these questions well (increasing, presumably with N/M).[4] This issue awaits further investigation.

[4]Answering this question will be both difficult and context specific. Difficult because it involves weighing the benefit of further observations of a given type much like problems in the sequential testing literature, together with costs in terms of the degrees of freedom lost as the set of parameters to estimate expands (which can be especially severe in a non-parametric setting). Context matters as well because some questions are more valuable than others, and weighing these relative values will depend on the specific problem at hand.