# Regression in Analysis

2LT Kevin Burke 12 Nov 2008

VDP Class 08-A07

In regression analysis, the goal is to determine the values of parameters for a function to best fit a set of data observations. Put another way, regression attempts to best describe what inputs result in a given output. Though there are many complex forms or regression models, the simplest is a linear regression model. This is the model I will use for illustration purposes. I do this solely for the purpose of building a basic understanding of regression analysis. One must realize, therefore, that the type of regression analysis that could prove useful to the intelligence community would also prove far more complex.

Take, for example, the value of a car. If one assumes that said value decreases by a constant amount each year after its purchase, as well as for each mile it is driven, the linear function "value = price – (x)age – (y)miles" would predict its value. In this equation, "value" is the market value of the car, "age" is how old the car is, and "miles" is the number of miles that the car has been driven since its purchase. "X" and "Y" represent the relationship between the value of a car and its age and mileage respectively. In this case, one would expect the relationship to be negative. That is to say, one would expect the value of a car to decrease as age and mileage increase.

In any analysis of this type, one must provide a data file which contains the values for the variables. In this example, each data record would need to contain three numbers: value, age, and miles. For this example, the *classifieds* section of a newspaper might be a

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| 1. REPORT DATE | 2. REPORT TYPE | 3. DATES COVERED |
|---|---|---|
| **12 NOV 2008** | | **00-00-2008 to 00-00-2008** |

| 4. TITLE AND SUBTITLE | 5a. CONTRACT NUMBER |
|---|---|
| **Regression in Analysis** | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Military Intelligence,Fort Huachuca,Sierra Vista,AZ,85613** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | | |
| **unclassified** | **unclassified** | **unclassified** | **Same as Report (SAR)** | **5** | |

valuable source of data. The most important thing to remember about data sets, though, is that size *does* matter. The more observations provided, the more accurate the analysis.

At this point, it may still be unclear as to how regression analysis can be used throughout different fields. For Economists, one may want to get a better understanding of the way families spend money. In this case, the dependent variable might be a family's consumption expenditure and the independent variables might include the family's income, the number of children in the family, the amount of debt held by the family, and other factors that may affect the family's expenditures. For Sociologists, the interest may be in finding out what, if any, is the relationship between one's social status and one's occupation. Here, the independent variables might be inherent characteristics of one's job such as pay, qualifications, education, etc. In the intelligence world, especially given the current conflicts, one might be interested in determining what factors may or may not contribute to the emplacement of an IED.

Here, before I begin to lay out my initial equation, it seems quite important to point out two things. First, biases are an inevitability of everything we do. Whether it rears its head when one makes an assumption about the used car salesman in the ugly coat, or whether it's found in the assumptions that an analyst makes about the enemy, they are always there. This, however, does not mean all hope is lost. Rather, one need only to recognize one's biases, account for them, and move on. The point being that, even with biases, it is quite possible to provide quality analysis. Secondly, it is imperative that one thing be quite clear. I am not, in any way, arguing that regression

analysis is the end-all be-all of analytical tools. It is not the magical key that unlocks the enemy's secrets. It is not a discrete tool intended to be used by itself. And, above all else, it will not, under any circumstances, do the analysis for the user. Like anything else in the analyst's toolkit, regression analysis will only aid the analyst in making better, more predictive conclusions. Nothing more!

That being said, one more thing must be brought to light: my biases. My combat experience, thus far, has been that of a lower-enlisted gunner in an infantry unit. Specifically, I spent the better part of a year driving very slowly through southwestern Baghdad clearing routes. Put another way, I spent a year trying to find IED's before they found me. With that, one must realize that my proposed equation is very specific. Anyone who has spent anytime whatsoever in the area will realize that my equation is a very Baghdad-centric one. Again, though, that does not render it useless. Rather, it means only that the analysis tool I put forth cannot be used everywhere. Different areas, countries, enemies, etc will all require their own model. Just as the model used to fight the Russians during the Cold War contrasts dramatically with the model needed to fight an insurgency, so too does regression analysis require varying models for varying situations.

**B = a S + c P + d T + E**

**Where**:

**S** = socioeconomic status of an area
**P** = previous IED activity
**T** = time
**E** = unattributed factors

Above is what I suggest to be a beginning to understanding IED emplacements in and around Baghdad. As stated previously, it is not intended to be *the* answer. It is only put forth in order to encourage analysts to use regression models in their analysis.

To begin, it seems easiest to discuss the simplest of the variables. First, as is commonly known, it is often an enemy TTP to emplace an IED in a location that has had previous success. That is to say that the enemy, quite literally, often place a new IED in the crater formed by an earlier IED. Such was certainly my experience and, from talking to others, theirs as well. It seems like a very logical variable to be included.

The next simple variable is the "time" variable. Dusk and dawn, and not just with IED's, is often the time of enemy attack. Therefore, it seems important to incorporate that common occurrence into the equation. For me, it seemed like something always went "boom" when we had the sunrise patrol. Also, though, the "time" variable would begin to account for dates and time of year. Mondays always seemed more dangerous that Fridays. Uncomfortable months (too hot and too cold) seemed safer than comfortable ones. Holidays always seemed a bit more dangerous. "Time" attempts to account for the fluctuations that happen according to the calendar and watch.

Lastly, "socioeconomic status" (SS) attempts to quantify that gut feeling that every soldier has about an area, about a "bad" area. I can't recall one time ever being attacked, IED or otherwise, in a nice neighborhood. That's not to say that it doesn't happen, but it seems logical that it is far less likely to occur. In addition, unlike the previous variables, there is no single metric that comprises the data entry. On the contrary, the "SS" variable would certainly be a compound variable. That is, it would be

a variable, one data entry, comprised of a series of other measurements. For example, in constructing the "SS" variable, one might consider things like per capita income, reliability of public services, and the number of schools in an area. In constructing this variable, one would want to consider, or quantify, all of those things that make one neighbor hood "worse" than another.

In closing, it seems essential to reiterate several key points, not only to remind those readers who have determined regression to be hogwash, but also to ensure that there be no mistake as to what role regression analysis can play in intelligence. Put simply, regression is nothing more than another tool available to the analyst. It is not, and should not be *the* analysis. Just as a pattern analysis wheel is nothing more than pretty shapes and colors if there is no analysis, so too is a regression model nothing more than a headache-causing jumble of numbers without solid theory and analysis behind hit. Also, remember that a regression model is a very specific thing. While the intent is to create a model and analyze a data set in order to better predict, one must realize that, ultimately, the model only tells the analyst about that specific data set. It is up to the analyst, and those creating the models, to determine whether or not effective predictions can be made. Lastly, regardless of whether one understands regression a little bit or not at all, it is important to remember that the whole point is to make improvements. While the IED model presented above is clearly immature, incomplete, and overly simple, it is a beginning. "Good enough" is not a phrase that should ever enter the analyst's lexicon. To do so is to put soldiers at risk willingly. Improvement, not perfection, is the goal, and the addition of regression analysis to the analyst's toolkit would certainly be a vast improvement.