# Inferring the Degree of a Curve from Empirical Data

## An Apology for (some) Statistical Tools

The best statistical tools are those that provide a rigorous and standardized measurement for a concept that already exists in natural language. These include, for example, our measures of probability, correlation, significance, and similarity. Since linguistic meanings shift, and are relatively imprecise in comparison with mathematics, these identifications are never perfect. We are always left with inane caveats like the infamous "significant results may not be *significant*" or the badly abused phrase "correlation does not imply causality." And it is true that very few educated adults can provide a full and accurate explanation of Pearson's *r,* or even how to calculate a standard deviation. Yet those of us who have had any kind of successful math education are apt to remember that such calculations are fairly simple, that they are intuitively sensible approaches to the problems they address, and that at some moment in the past, we understood them intimately enough to pass judgement on their validity. This type of understanding can extend, I think, even to multivariate regression and cluster analysis, and it is of the utmost value in providing a common epistemology for public debate.

These tools, then, are distinctly different from many other forms of statistical analysis, which are so opaque that almost all readers must accept or reject their results on faith, as received knowledge from an authority. I've discussed the general problems with this recently, in the context of the "debate" between zetetics and spherecists. A more specific and more meaningful case would be John Lott, whose larger patterns of deceit and self-promotion put extraordinary strain on the the please-just-trust-my-opaque-analysis-of-X argument, as Ted Goertzel and others have pointed out. Again, modeling algorithms, while enormously useful, are often so complicated and nuanced that they can only really be understood by their creators, like the machines of some mad scientist in a Gothic horror novel. In arenas such as climate change, for instance, the strong emphasis on modeling has presented the public with mathematics as a black box, and the results have been disastrous.

I make these remarks, first of all, in apology for the fact that some of my own statistical methods have been fairly opaque. Recently I have set about re-working one data set, in particular, that I think I have sinned against in this vein. But in this article, I am proposing a specific tool, which I call a Bethel Test, and I want to make it clear why I think the tool is useful.

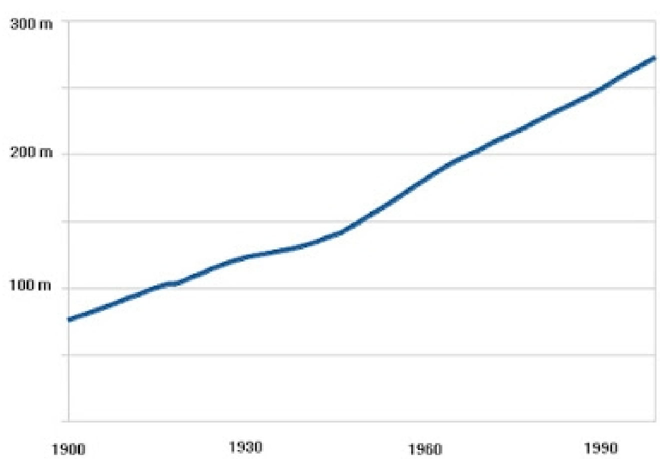
## Degrees and Derivatives as Metaphor

Lately I have been working on a set of problems in which the natural-language metaphors reference (more or less directly) the concepts of polynomial degree or the closely related concept of derivatives and antiderivatives in calculus. The underlying question is whether it is appropriate to describe a particular data set as patternless or as having a "trend", and if does have a trend, whether it is best described as linear, quadratic, cubic, or *n*th-polynomial-degree, or even as an exponential function (which in polynomial terms would have an infinite degree).

Polynomial equations do the yeoman's work of mathematical explanation in a range of fields, because they translate easily into simple narratives about trends, and in the other direction they translate nicely into calculus. Most commonly, this takes the form of spatial, temporal, and kinetic metaphors. In *Metaphors we Live By,* Lakoff and Johnson have suggested that such directional metaphors are irreducible parts of human thought; in any event they are certainly ubiquitous. For instance, the derivative chain *position, velocity, acceleration* is built into many other explanations. In popular discussions of the recession, it is common to hear references to the economy “at a standstill” or “in decline” or “in free-fall”: respectively, these reference a position value, and then its first and second derivatives. These notions can be intuitively daisy-chained to higher levels. Even though most of the public does not know or use the semi-official continuation of that derivative chain (*jerk, jounce, snap, crackle, and pop),* we hear a reference to the third derivative of position in phrases like Campbell Newman's claim that the economy is “on a power dive into the financial abyss”. Implicitly, a “power dive” is an accelerated free-fall, which is itself an accelerated decline, which is a continuous change in position. Such descriptions are distinct from one another in mathematically meaningful ways, and they are also easily perceived metaphors.

Similarly, discussions of Hubbert peak theory (i.e. peak oil) often highlight the fact that a curve can be trending in one direction while its derivative is moving in the opposite direction, a concept that goes beyond simple polynomials. Various phrases are used to reference this sort of dual trend, some of them rather oblique, but such discussions often resort to graphs of derivatives, such as energy-return-on-energy-investment charts. Since the conceptual popularization of peak oil in the last decade, Hubbert's language itself has become a referent metaphor for complex curves of all kinds: we can speak casually of “peak social media” or “peak Microsoft” or “peak Ron Paul”.

A related issue involves data that is informally or semi-formally described as “exponential growth”, rather than polynomial. This includes biological populations, economic indicators, and even industry benchmarks such as Moore's Law of microprocessor miniaturization. There are sometimes theoretical considerations for suspecting that a measurable phenomenon can be explained by an exponential function: for instance, it seems reasonable to model population growth as an iterated multiplication of the current population, and such a model yields an exponential curve. Nevertheless, other theoretical models for population growth are possible, which take into account carrying capacity or demographic transition effects. These models suggest their own curves, perhaps polynomial or logistic. The choice among these mathematical narratives is not always intuitive. For instance, the US population is often referred to as as an example of exponential growth. Yet if someone with no *a priori* beliefs was shown a graph of that population over the 20th century, as below, they would probably conclude that it approximated a merely linear function. (And I will argue that it is better described as quadratic.)



The belief that a curve is exponential can create a sort of extrapolative logical fallacy. All time series can be described as having an average rate of growth (or decline). It often seems innocuous to reify that average as a feature of the system in question, and then project it into the future. This is frequently a basic assumption in economic forecasting. Yet this implies that the system is necessarily best modeled by an exponential function, which is hardly ever the case. A fairly standard quantitative fallacy in the media is to focus on that implied exponentiation, rather than on the numbers themselves. Absurd possibilities abound: for instance, in their first year of life, an average male baby gains weight at about 10% per month. It follows that by the time the child is five years old, they will weigh more than a ton. We are familiar enough with human growth rates to know that they are not exponential, but the same logic may seem persuasive when applied to a less intuitive metric, like the GDP of China.

These matters of degree are of the utmost importance to policy initiatives, whether personal, corporate, or public. In a non-linear variation on Zeno's paradoxical race, we can note that the quadratic tortoise will *always* beat out the linear Achilles in the end. And the cubic snail will eventually beat the quadratic tortoise, and so forth. Thus it is crucial that political initiatives, on whatever level, operate in the same "polynomial degree" as the events they seek to address. If there is a second-degree influx of refugees to a region, a first-degree response will not be adequate for long. A community cannot hope to deal with a third-degree increase in solid waste by using a first-degree policy of reusable shopping bags. And so on and so forth. We like to think that "every little bit counts", and this is true enough in an additive sense, but addition is no use when your opponents are multiplying or using hyperoperations.

To my knowledge, Malthus' 1798 *Essay on the Principle of Population* is one of the first pieces of political economy to specifically highlight this issue. He compared the supposedly linear curve of food production with the supposedly exponential curve of human population, and drew a well-known and apocalyptic conclusion. Among the critiques of Malthusian doctrine are the claims that food production is non-linear, and that population is not in fact exponential. But how can we be sure? Recently, the North Carolina legislature, in one of the more striking moments of politicized mathematics, attempted to ban the use of non-linear equations in the context of climate change models. I am unclear if they wanted to ban all quadratic equations (a policy many frustrated algebra students might support) or only the ones they didn't like. At any rate, such questions beg for standardized metrics.

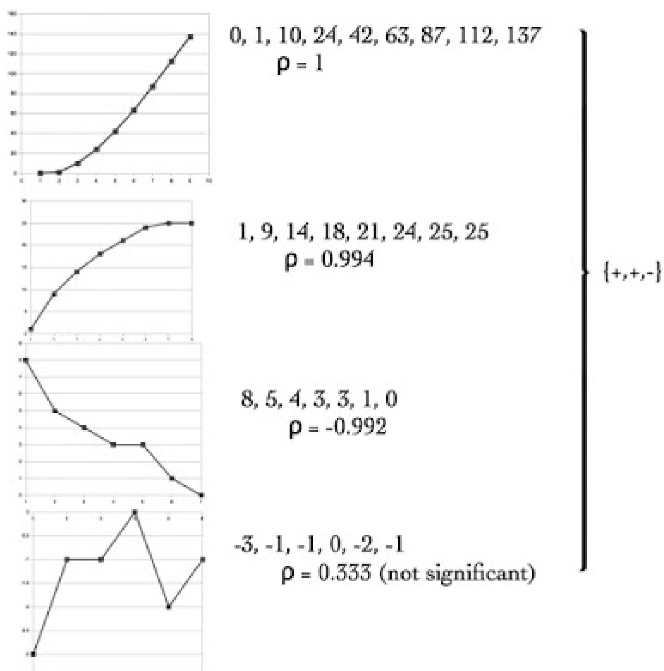
What is intriguing to me about these examples is that while we have various natural-language descriptions that reference the mathematical concepts, there is no immediately forthcoming mathematical tool to turn to in evaluating the natural-language claims! Textbook curve-fitting methods either rely on subjectively "eyeballing" the data, or else make *a priori* assumptions about the nature of the curve. Generally either of these methods tends to minimize the polynomial degree of the curve being fit, both to avoid over-fitting and to simplify the resulting explanation. While those are reasonable goals for most problems, they do not help us determine what the *most explanatory* degree for such a curve is. At the same time, the suite of more advanced tools we might use, such as log-linear analysis, becomes increasingly opaque (e.g. dummy variables to deal with negative values). More importantly, such tools tend to be specific to certain types of curves, adding a second level of opacity: perhaps the researcher is simply choosing the tool that will give them the result they desire.

My interest in creating the Bethel Test was to analyze the "degree" of trends within a variety of data series in a way that seems intuitively defensible, and relatively easy to understand. I assume a sequence of paired, real-valued data sets, H and V, ordered by H. The output of the Bethel Test is a claim that the data exhibits meaningful trends over some number of degrees, and the directionality of those trends.

**Simplified Bethel Test (where values of H are equidistant)**

The special case of a data series where the H values are equidistant helps to illustrate the principle of this test, and it is the standard for annualized time series data, which included most of my original candidates. In this case, *n* values of V are ordered by the corresponding equidistant values for H. Spearman's rho (ρ) is then calculated for H and V, as well as any of the standard tests of significance for ρ. (I will use the Fisher-transform method throughout). Spearman's rho is a common non-parametric measure of whether or not two series of numbers occur in roughly the same order. In this usage, it lets us know whether V is increasing, decreasing, or neither, while H increases.

If the value for ρ is not significant, the Bethel test stops. If it *is* significant, we record the positive or negative sign of ρ: a positive sign meaning that V tends to increase with H, a negative sign meaning that V tends to decrease as H increases. Then we iterate the test in the following manner. The *n* values of V are replaced by *n*-1 values of ΔV (i.e. V2-V1 , V3-V2, etc.). Spearman's rho is then calculated for this new data set, as compared to its original ordering, and the significance of the new ρ is calculated again. We continue this process, recording the sign of ρ each time, until we reach a point where ρ is insignificant.




Ultimately, the Bethel test produces a sequence of signs, perhaps {+,+} or {-,-,+} or {} or {+,-,+,+,-}. In the example shown above, we begin with a set of nine values (0, 1, 10....) These values are increasing, as is visually obvious in the graph. The difference between each pair of successive values (1, 9, 14....), shown below, is also increasing. The difference between *those* pairs of successive values (8, 5, 4....) is decreasing. And at the next iteration, there is no significant pattern of increase or decrease. Thus we have a Bethel result of {+,+,-}, which can be reasonably interpreted as having three “degrees of trend”: velocity, acceleration, and (negative) jerk, to use the physical metaphor. Notably, the ability of Bethel Tests to acknowledge curves with trends in multiple directions distinguishes it from log-linear analysis and similar techniques in question, and brings it closer to to the way that such curves are described in natural language narratives.

**Bethel Test with Non-Equidistant H**

For data sets where the values of H are not equidistant, a slightly more complex algorithm is needed. Instead of comparing ΔV with the original ordering, we will compare the slope values ΔV/ΔH with the original ordering of H. With each iteration of the test, *n*-1 new values for H are calculated by taking the midpoints of each sequential pair of H values ( (H1+H2 ) / 2 , etc.). Note that for special case of equidistant H values, the definition above provides the same result.

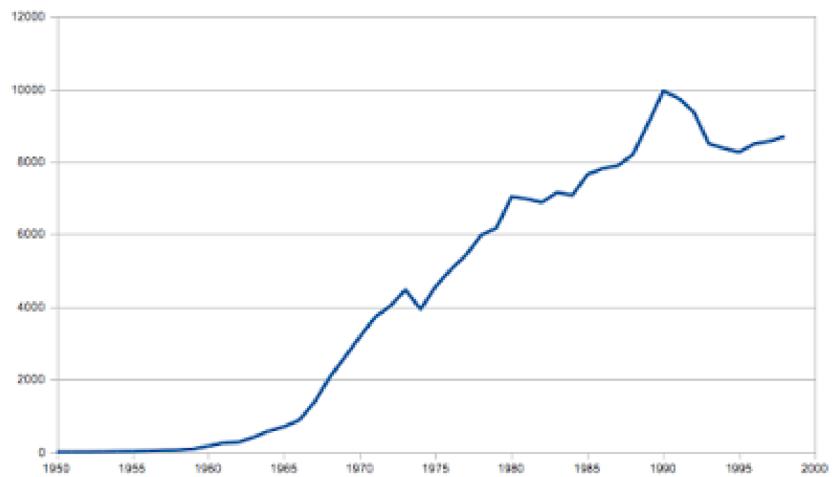
**Calculator**

I have created a web-based calculator for Bethel testing here. It uses comma-separated values; I hope it is fairly simple to use. Please contact me with questions, and please let me know if you find results from empirical data that exhibit more than three degrees of trend.

**Some Empirical Observations**

Looking over real-world historical data, it quickly becomes apparent that very few time series exhibit three or more degrees of trend in a Bethel test, even when they present phenomena that are popularly referred to as "exponential". For instance, the graph of US population discussed above is merely {+,+} or quadratic. The growth of Facebook from late 2004 to 2012 is also quadratic.

*Most* long-range human behavior appears to be either patternless, linear or quadratic. On the other hand, it is often possible to isolate sub-sections of a time series that have third degree trends. For instance, the graph below shows the production of personal cars in Japan, 1950-1998. Overall, at alpha = 0.05, the time series is {+}: car production was increasing, but not in an accelerated way. However, there are many sub-sections of this data which give {+, +}: accelerated growth. And there are 9 sub-curves that give {+,+,+} (hyper-accelerated growth) all of them in the 1950-1969 time-frame.




Such high-degree curves demand an explanation, and it is not hard to suggest one. In the 1950s and 1960s, Japan was rebuilding its pre-war industrial infrastructure with substantial help from the United States. During this recovery, Honda and Toyota both produced exceptional products, and began to serve both a substantial domestic market and, increasingly, an international market.

A similar confluence of factors can help to explain cubic growth in other situations. For instance, US government receipts display cubic growth between 1950 and 2007. This curve, however, subsumes two others: the population of the country was increasing, and the value of the dollar was declining. In constant dollars per capita, US government receipts only increase quadratically.

A third example of a cubic increase is more striking: the population of the city of Perm, in Russia, across a stunning 220 years. Without question, Perm (briefly known as Molotov) has benefited from constant and accelerated growth as one of Russia's major industrial centers, both under the Czars and the Soviet Union. But Perm is distinct from cities like Manchester or the Ruhgebeit in that it was essentially non-existent prior to the industrial revolution. Like the situation with post-war Japan, it attains cubic growth in part because of its humble origins.

Finally, there are examples of very high-degree curves over short time horizons, such as the Weimar hyperinflation, which has a degree of 17, and would probably be even higher if the metrical resolution allowed. I would suggest, however, that for most such events, we are recording a catastrophic breakage, rather than a continuous curve. When a currency depreciates a hundredfold in three weeks, the actual experience of market participants is not inflation but much more akin to confiscation.

**Examples**

*This list is always being added to. Please let me know if you find anything interesting*.

{} "trendless"

US homicide rate per capita, 1960-2011 (significant at 0.1)
Random number sequences

{-} "negative linear"

Raw copper production in Africa, 1992-2001 (at 0.01)
US voter turnout, 1960-2010 (at 0.1)

{+} "linear"

Mexican Immigration to the US, 1894-1973 (at 0.01)

{+,+} "quadratic"

Chinese energy production, 1978-2006 (at 0.01)
US Population 1900-2000 (at 0.01)
Population of Tiflis 1800-1990 (at 0.01)
Membership on Facebook, late 2004 to 2012 (at 0.01)
Automobile fatalities in the US, 1899-1932 (at 0.01)

Number of televisions in the US, 1945-2006 (at 0.01)
CREO list of mammal species extinct since 1500 (at 0.05)
Compact fluorescent lamp sales, 1988-2001 (at 0.01)
HIV infections, cumulative 1980-2001 (at 0.01)
Energy input needed to catch fish, Lake Chamo, 1998-2006 (at 0.05)
Number of Walmarts in the US, 1962-2006 (at 0.01)

{+,-} "diminished linear"

US reported rape rate per capita, 1960-2011 (at 0.01)
Population of London, 1750-2000 (at 0.01)
{-,+}

Decline in hits to a website for 15 days following a traffic spike (at 0.1)

{-,-} "negative quadratic"

Conversion efficiency of Jehovah's Witnesses (in hours of work per baptism), 1962-2011 (at 0.02) (notes)

{+,+,+} "cubic"

Population of Perm (Molotov), 1750-1970 (at 0.02)
Japanese Car Production, 1950-1969 (at 0.05)
US Government Receipts, 1950-2007 (at 0.1)
World photovoltaic energy production, 1971-2001 (at 0.05)
World wind energy production, 1980-2001 (at 0.01)
Height of the tallest building on earth, 2500 BC to 2014 (at 0.05)

{+,+,-} "diminished quadratic"

US Census world population estimate, 1950-2012 (at 0.01)

{+,+,+,+} "quartic"

Cellular phone subscribers, 1985-2001 (at 0.1)
Number of Walmarts in the US, 1962-1981 (at 0.1)
Maximum number of transistors per circuit, 1971-2014 (at 0.05) (notes)

{+,+,+,+,+,+,+,+,+,+,+,+,+,+,+,+,+,+}

Reichsmark exchange rate to gold, 1/18 to 9/5/23 (at 0.01)

{+,+,+,...} "exponential"

The Fibonacci Sequence.
Any exponential sequence in the format V = x^H, where x > 1.

{+,+,+,...,-,+,-,+,-,...}

Any exponential sequence in the format V = H^x, where the initial number of + degrees is the truncated value of x.

{+,-,+,-,...}

The Harmonic Sequence.
Exponential sequences in the format V = H^x, where x is between 0 and 1.

{-,+,-,+,...}

Exponential sequences in the format V = H^x, where x is less than 0.
Exponential sequences in the format V = x^H, where x is between 0 and 1.
Any section of the Fibonacci Sequence, run backwards.