The Challenges of Using Ranks to Estimate Sales

Researchers have frequently used publicly available data on product ranks to estimate nonpublic sales quantities, believing that there is a linear relationship between logged rank and logged sales values due to the assumption that sales follow a power law. However, using data on book sales, which are commonly thought to follow a power law, we find that the (double logged) relationship between ranking and sales is not linear, but actually concave. We demonstrate that this concavity is likely to cause poor predictions of sales in many instances. We also explore the use of nonlinear specifications as an alternative method to predict sales from ranks and find a simple specification that ameliorates many of these poor sales estimates. We illustrate some of the problems of applying a linear technique to this nonlinear relationship by examining the claim that the greater product variety made available to shoppers on the Internet has a large positive impact on social welfare, and also a claim about sales levels in top 20 and top 50 "charts."

I. Intro

Many market measurements, whether elasticity of demand or supply, income elasticity, the dispersion of market shares, sales diffusion patterns over time, or simply revenue estimates, require knowledge of the number of units sold. But because most companies treat data on sales as confidential, acquiring unit sales data is often difficult if not impossible.

Beginning in 2003, hundreds of research papers in all business fields have used product popularity ranking data taken from retail sites on the Internet as a proxy for unavailable unit sales.[1] This popular technique converts the ranks of products into unit sales by assuming a linear relationship between the logarithm of product ranks and the logarithm of unit sales, with the slope of this relationship (the "slope coefficient") being the key parameter to be estimated. We refer to this as the rank-substitution literature. This technique should work when product sales follow a power law and it has been presumed that sales in many entertainment and media markets, such as books, video, music, and apps follow a power law. The use of this technique has become so widespread that researchers (e.g., Benner and Waldfogel 2016) sometimes forgo any attempt to estimate the coefficient of the linear relationship between sales and ranks and just pick a value somewhere in the range of previous literature, being unaware that small changes in the coefficient estimates can cause dramatic changes in predicted sales. Our interest here is in examining whether the power law

[1] Brynjolfsson et al. (2003) and Chevalier and Goolsbee (2003) have over 1,600 and 600 Google Scholar cites respectively although many of the citing articles do not use the rank-substitution technique. We estimate usage of the rank-substitution technique by examining a sample of 141 articles citing Brynjolfsson et al. (2003). We find that 25% of these papers use the rank-substitution technique. Applying this percentage to the articles citing Brynjolfsson et al. (2003), we estimate that 375 articles use the rank-substitution technique. We then examined the most cited 40 papers citing Chevalier and Goolsbee (2003) and found that 5% of them used the rank-substitution technique and failed to cite Brynjolfsson et al. (2003). We conclude that over 400 articles are likely to have used this technique as of September 2021, with about half of those articles examining the book market and rest examining other industries such as software (apps), music, or video markets. As another example of the extent of the usage of the technique, a meta-analysis of the effect of WOM on sales published in the Journal of Marketing Research (Rosario et al. 2016) examining 96 studies covering 40 platforms and 26 product categories finds that the sales variable was operationalized using sales rank in 60.5% of all these studies.

assumption underlying this literature is reasonable for the product most commonly examined—books.

To the best of our knowledge, we are the first analysts to have what is essentially a complete distribution of a product category's (books) sales. Our data set allows us to examine whether the log-rank/log-sales plot is actually linear. We find that the relationship is not linear but is concave, indicating that the observations are not coming from a power law distribution. We then compare actual sales to the sales predicted using the assumed linear relationship between logged sales and logged ranks. We find that the errors in the predicted sales values can be very large, especially when an analysis tries to cover a wide range of ranks and sales. We also demonstrate how the chosen time period (e.g., weekly or yearly ranks) affects the measured relationship between ranks and sales, and how what might appear to be small differences in the estimated relationship between ranks and sales can have serious impacts on estimated results.

Because the log-rank/log-sales relationship is curved, estimates of the slope will depend on where on the curve the sample was taken. We demonstrate that the influential early estimates in this literature are based on unrepresentative samples skewed toward better selling books (but generally not the best selling books). We illustrate the potential for large errors in this literature by examining hypothetical causes of error due to trying to impose a linear structure on the curved relationship. Then we illustrate the nature of errors in two specific results found in the literature. First we show that the claim of large Internet-induced improvement in social welfare due to greater product variety in the book industry is an artifact of incorrectly assuming that sales follow a power law. Then we show that some predictions of sales in charts of top sellers have been inaccurate for the same reason.

Because the nature of the concavity changes little from year to year or across different categories of books, we explore the use of simple non-linear specifications and find that they provide a superior method to predict sales from ranks.

II. Pareto, Zipf, and Power Laws

Some apparent empirical regularities in the size distribution of physical and economic phenomena have fascinated researchers since the late 19th century. Vilfredo Pareto (1897) created theories to explain his observations that 20% of the population generated approximately 80% of the wealth (measured as land ownership) in many countries and in many time periods. Similarly, George Zipf (1949) found a regularity in the distribution of word usage such that the second most frequently used word was used about half as frequently as the most frequently used word, and the third most frequently used word was used one third as frequently, and so on. This regularity, known as Zipf's law, has been applied to factors as diverse as the population of cities and the viewership of television channels.

These intriguing relationships are based on the frequency of ordered observations. It has been generally proposed that these empirical regularities are related in the sense that they can be described as the outcomes of draws from a power law distribution.[2] Products and events that are thought to possibly follow power laws have several characteristics in common. One is that the range between the smallest and largest observations is very large (often many orders of magnitude). A second is that the distribution is skewed with the greatest frequency of observations occurring at the endpoint of the distribution with the smallest observations (e.g., lowest selling books).

[2] Various theories have been put forward to explain why such relationships might hold, such as preferential attachment, random walks, critical phenomena, or the principle of least effort. The interested reader is referred to section IV in Newman (2005) who discusses these theories at length.

These two characteristics are found for sales in entertainment markets where "hit" products produced by "superstar" creators dominate the much larger number of poorly selling items. These markets include books, movies, music, software and apps, and they have all had their sales estimated using the rank-substitution methods that are based on power law distributions, although power laws have been used to describe many other relationships such as crater sizes and citations to academic articles.[3]

One of the properties of power laws that intrigued Pareto, Zipf and their followers was that if logs were taken of both the ranks of the items as well as the units of the items and the two ordered series were plotted against each other, the result would be a straight line. Zipf called a plot of this relationship a "rank frequency plot" and the term is still used although Newman (2005) observes that the plots usually compare the rank of observations to the sizes of the item in question, so that a *rank-units plot* might be a more meaningful term. Because our units are sales, from hereon we will use the term *rank-sales plot*. There is also no single convention on whether (logged) rank goes on the vertical or horizontal axis.[4] Because the literature uses both possible versions of which variable goes on which axis (for example Goolsbee and Chevalier [2003] use Pareto's convention whereas Brynjolfsson, Hu and Smith [2003] follow Zipf), care needs to be taken when comparing slopes

[3] Many of the empirical instances that had been put forward as examples of power laws are now considered better explained by other distributions. Thus, for example, Clauset et al. (2009) examine twenty-four instances where the tail of observations had been thought to be consistent with power laws and conclude that seven of those examples are inconsistent with power laws, nine are consistent with power laws but also other distributions, and one is consistent only with a power law. In the market for super-long-term bestselling books (defined as 633 titles selling more than 2 million copies over a 70-year period), power law distributions were considered reasonable fits, but so were log-normal and stretched exponential distributions. For our purposes, what counts is whether the rank-sales plot is sufficiently linear throughout its range that the techniques used by rank-substitution researchers provide sufficiently accurate results to allow the imputation of sales value from ranks without causing significant error. This question is different although related to the one examined by Clauset et al. which is instead focused on the exact distributions that might best explain data in the tail.

[4] Newman (2005), referencing a web posting by Adamic (2000), notes that rank is on the horizontal axis under Zipf's analysis but on the vertical axis following Pareto. These are identical analyses but with their axes reversed, although this has not always been understood to be the case.

because the slopes of the lines under the two different systems are the inverse of one another. In a special instance, where the slope of the line is equal to -1, we have what is referred to as Zipf's law.

This linkage of the straight-line property between sales and ranks can be more formally illustrated, following Newman (2005). A power law distribution is typically represented as

(1) $$p(x) = Cx^{-\emptyset}$$

The cumulative distribution function of (1) is

(2) $$P(x) = \frac{C}{\alpha - 1} x^{-(\emptyset - 1)}$$

which can be rewritten as

(3) $$P(x) = Kx^{-\delta}$$

The cumulative distribution function tells us the probability of getting a value less than or equal to $x_{i,}$ which is the same as the rank of the x_ith observation relative to the total number of observations. Taking the log of (3) and replacing P(x) with the rank of the product and *x* with sales quantity, we get

(4) $$\ln(rank) = \ln(K) - \delta \ln\,(\text{sales})$$

Rearranging to follow the Zipf convention we get

(5) $$\ln(sales) = K' - \alpha \ln\,(rank)$$

where α is 1/δ and K' is $1/\ln(K)$. The shape parameter in which we are interested is α although those following the Pareto standard report its inverse, δ.

The regression that is estimated by researchers in this literature is of the form

(6) $$\text{Ln(sales)} = \beta 1 + \beta 2 \cdot \text{ln(rank)} + \varepsilon$$

(or its inverse). The main focus in the rank-substitution literature, and our focus as well, is on the slope coefficient, β2.

In the empirical work that follows, we shall report the slope coefficient, β_2, as a positive number although it is always in fact a negative number. This is to avoid having to constantly refer to absolute values when referring to which of two values is larger than the other, just as own price elasticity of demand values are often converted to positive numbers to simplify comparisons.

III. The Genesis of Using Ranks to Predict Sales

The rank-substitution literature begins with two seminal papers—Brynjolfsson, Hu, and Smith (2003) (BHS) and Chevalier and Goolsbee (2003) (CG)— providing a methodology that hundreds of papers have since followed. Both of these papers examine the book industry and assume that book sales can be represented by a power law. It is common for books to be considered as one of the industries where sales follow a power law as described by Anderson (2006) in his best-seller "The Long Tail."

The original insight behind the claim that ranks can be used as a sales proxy (given credit by both BHS and CG) is a 2001 PowerPoint presentation by Schnapp and Allwine at a web mining conference. Schnapp and Allwine examined the relationship of book ranks and sales that provided both an outline of the rank-substitution methodology and empirical support for its accuracy.[5] They found that the accuracy of predicting unit sales using ranks as a proxy for sales was within 15% of actual sales for the better selling books in their sample and within 25% for the lower selling books.[6] This appears to have been an important piece of evidence underpinning the belief that ranks can be

[5] Schnapp and Allwine also provide a slope coefficient (for which we have reversed the sign) of 1.11 for books with ranks up to 20,000 and 0.77 for the better selling books ranked below 2,000.

[6] These appear to be average deviations because when they list some actual deviations, whether in tabular or graphic forms, it is easy to find instances of individual books where the deviations are outside these ranges.

used to accurately predict sales. Nevertheless, as we discuss in Appendix 1, these conclusions are suspect because the ranks used by Schnapp and Allwine cannot be the correct ranks for books in their samples. Specifically, the relationship between ranks and sales must be monotonically negative but is not for many Schnapp and Allwine observations. These problematic ranks are likely due in part to Amazon sales ranks being reported in an inconsistent manner prior to 2004, also discussed in Appendix 1.

Building upon Schnapp and Allwine, BHS and CG proceeded to predict sales from ranks using data they were able to obtain from various sources. BHS obtained a sample of 321 books that span a large range (from rank 238 to 961,367) and estimate a slope coefficient of 0.871. CG also have an estimate (0.671) based on offline best-sellers listings in the WSJ (which included ranks and sales). The presumed power law induced constant slope coefficient eliminates any need for the analyst to try to attain a representative sample covering the full distribution of observations.

Due to the difficulty of obtaining any data on ranks and sales, however, another methodology, first proposed by CG, has also been used in this literature. In this methodology, the researcher conducts an "experiment" by watching how ranks change when some quantity of a book is purchased by the researcher. CG and BHS each conduct such an experiment and CG's estimate of the slope coefficient is 0.855 while BHS's is 0.916.[7] BHS note that their experimental estimate is "remarkably similar" (p 1588) to their estimate made with their 321-book dataset.

The CG "experiment" methodology is described as an inexpensive way for translating ranks to quantities sold since only a handful of units need to be purchased. But, if the experiment is to be

[7] CG use the Pareto approach whereas BHS use the Zipf approach which is the standard used in this paper. We translate the slope coefficients for studies using the Pareto approach into Zipf values by taking the inverse of the reported slope coefficients.

kept low cost, it is limited to purchases of books that do not sell a very large number of units in order for the additional units to have an outsized impact on the ranks so as to provide a high signal to noise ratio. Garg and Telang (2013) note the difficulty of generating even small samples of sales data or running "experiments" and instead propose an alternative methodology using downloads and revenue generation ranks that has been followed in papers such as Wang, Li and Singh (2018), although their methodology is still dependent on the power law assumption.

The general approach of assuming a linear relationship between the logarithm of product ranks and the logarithm of unit sales has recently been described as a "robust research tradition" by Benner and Waldfogel (2016), who use it for sound recordings:

> There is a robust research tradition of translating sales ranks into sales quantities using the following relationship: $q = Ar^B$, where q is the quantity sold and r is the sales rank. The parameter B reflects how quickly sales fall off at lower ranks (higher values of r). Studies generally find B to be in the neighborhood of -1. [p. 136]

After almost two decades, the slope coefficients have usually been found to be between 0.7 and 1.1, and researchers sometimes forgo any attempt to estimate the coefficient and just pick a value somewhere in the range of previous results (e.g., Benner and Waldfogel (2016)), thinking that the seemingly small difference between these coefficients cannot have much of an impact on results. But we will show that these apparently small slope differences will often have a large impact on results.

Unlike all prior researchers of whom we are aware, we have been fortunate to obtain a multi-year dataset covering the sales of virtually all consumer book editions sold in the US. This allows us to examine the complete relationship between ranks and sales, where ranks are consistently and appropriately calculated (since we have the data to calculate them). We can then directly test the fundamental assumption behind the rank-substitution literature: whether books sales follow a power law. After discovering that this assumption is not supported by the data, we investigate to what extent the relationship between ranks and sales may be mismeasured. We also propose rehabilitating the rank-substitution methodology using nonlinear estimation and consider its likely improvements to sales estimates, which appear substantial. We illustrate problems with linear estimation using two prominent the results currently in the literature. Finally, we consider whether some previous linear estimates might nevertheless be useful because researchers were only interested in bestsellers or because they hit a "sweet spot" where most of industry sales were included and yet the curvature of the log rank log sales relationship was small enough to allow sales estimates to be relatively unbiased.

IV. Data on the Book Industry

Our data on book sales comes from Nielsen BookScan, now known as NPD BookScan. Each observation is for an edition of a title (e.g., paperback or hardcover), and titles can sometimes have dozens or hundreds of editions. BookScan contains sales information (as well as variables such as list price, date of publication, publisher, and so forth) for all physical editions sold in BookScan's large panel of retailers (online and offline), which, given its high coverage of retailers, is likely to include just about every edition sold in the U.S.[8] While the data include online sales of print books (including self-published books), excluded are sales of e-books (e.g., the data include sales of print

[8] BookScan derives its data from point-of-sale transaction information (e.g., checkout scanners) reported by approximately 16,000 retail outlets thought to represent about 85% of the market, including online retailers such as Amazon. Since not every retailer is included, it is likely that the BookScan data will be somewhat short on the sales of each title/edition. Since BookScan includes retailers with large catalogs of titles, however, it is likely that almost all titles/editions that are sold in the market are included in the data.

books from Amazon.com but excludes sales from the Amazon Kindle Store) which only became an important part of the market after 2010.[9]

BookScan is normally considered the gold standard for data on physical book sales. The quality of the BookScan data is attested to by the fact that BookScan's subscribers are mainly book publishers wishing to keep track of how their editions are selling in the retail market since they do not know their actual sales until retailers finish returning the copies that do not sell, often months or years after taking delivery. Further evidence of the esteem in which BookScan data is held is the fact that Amazon, wishing to provide authors with detailed data on their sales, provides those authors access to BookScan sales data.[10]

Our purchased access to BookScan's data provided weekly and yearly sales information on edition sales for 13 years (2004 through 2016). For each year, our database includes information about the 250,000 top-selling editions in each major book category (although many of those editions will have empty data points for some variables, such as title or even sales). In early years, our data set contains the entire population of book editions sold at retail for the categories of Adult Fiction and Juvenile Fiction and Nonfiction. In later years, the 250,000 constraint starts to bind, meaning that we are missing the sales of editions selling one or two units, although we believe that we were able to estimate the number of such editions so that our data set is essentially complete for all years.[11]

[9] According to Statistica, only 69 million eBooks were sold in 2010 representing about 10% of the American market, https://www.statista.com/statistics/426799/e-book-unit-sales-usa/.

[10] This practice began in 2011 and continues to this day. See https://latimesblogs.latimes.com/jacketcopy/2010/12/amazon-gives-nielsen-bookscan-to-authors.html

[11] For works of adult fiction, the limit of 250,000 observations begins to bind starting in 2010 with the number of observations (book editions) selling 1 unit beginning to decline such that by 2014 there were no reported sales of editions with 1 unit. Nevertheless, we attempted to overcome this limitation by using information in the data to fill in synthetic observations for years when the number of observations were larger than 250,000. We did this by examining the growth rates of observations containing 1, 2, or 3 units during the years the data are not truncated. The correlations between the growth rates of observations with 1 and 3 units was 0.98 and between 2 and 3 units was 0.70. We added synthetic observations for the years where the data is truncated under the assumption that the growth in the missing

We chose not to include the Adult Nonfiction category in our main analyses because the 250,000 constraint leaves out too many editions for us to viably correct the numbers although we do examine how similar the distribution of Adult Nonfiction is up to the point where the 250k constraint begins to bind.

For the small portion of our analysis that examines the concentration of sales, we merge all versions of the same author-title combinations. For the main part of the analysis, however, where we are focused on the relationship between quantity sold and ranks, we do not aggregate author-title combinations since the literature we address treats all editions as separate items.

We note that although Books in Print and Amazon.com list millions or even tens of millions of mainly old and obscure editions, and that most prior research assumed that all of those editions have positive sales (e.g., BHS 2003), in fact many, perhaps most of those editions, do not sell any copies in any given year. For instance, Liebowitz (2023) examines a sample of 2,862 editions of very old former bestsellers listed as available in Books in Print in 2012 and finds that 83% of those editions do not sell any copies over a thirteen-year period beginning in 2004.[12]

The BookScan data do have characteristics that have been linked to power law distributions. First, there is a tremendous difference between the quantity sold of high and low selling editions, with yearly sales of best-sellers in the vicinity of 3 million units, and the lowest selling editions, of which there are tens of thousands per year, each selling 1 unit per year. Second, the peak of the distribution is at the left endpoint, not the interior, and the height of the distribution declines almost

observations with 1 or 2 units would have been the same as for observations with 3 units. For observations with 1 unit, we added back observations beginning with 2010 and for observations with two units we added back beginning in 2014. Similarly, we adjusted three years of data for juvenile editions when the number of editions selling 1 unit began to incongruously decline. These adjustments do not change the results in any important manner (i.e., the result is the same when studying years with no truncation).

[12] See footnote 10 in Liebowitz (2023).

monotonically throughout its range. There is nothing in this initial examination of data that would disabuse us of the notion that book sales follow a power law.

The great importance of the top sellers in overall sales is illustrated in Figure 1 which reveals 2009 cumulative unit sales for editions (ordered by sales from largest to smallest) as a share of total sales for all editions (424,653). The editions are ordered from highest selling to lowest selling.

Figure 1: Cumulative Share of Sales 2009

The 20% of editions that sell the most copies account for virtually all the sales in the industry, with the top 10% accounting for about 93% of sales and the top 5% accounting for about 84% of sales. The takeaway from this figure is that a small percentage of editions account for a very large share of total sales, and no reasonable economic discussion of the overall industry can occur if data on the better selling editions, the top 3% of the industry editions, say, are not included. Very similar levels of concentration are found when we perform these calculations for each year.

V. The Nonlinearity of the Double Logged Rank-Sales Relationship

We begin by visually examining the shape of the rank-sales relationship to see if it conforms to the linearity assumption that has been assumed throughout the rank-substitution literature. Figure 2 reveals the empirical relationship between the logged rank and logged (unit) sales in the American Adult Fiction and Juvenile book market in the years 2004 and 2016.

Figure 2: Relationship between Logged Sales and Logged Ranks

The relationship in each year appears, even to the unaided eye, clearly concave to the origin. An overwhelming majority of the editions sell very few copies and are represented on the right-hand side of the figure where the curves are relatively steep. The editions with the greater sales are represented on the left-hand side where the curves are relatively flat. The two curves, 12 years apart, are very similar to one another except at the rightmost portions. Because there are many more editions in 2016 than in 2004, the 2016 line must lie to the right of the 2004 line near the rank axis.[13]

Table 1 provides greater detail on the slope of the double logged sales/rank curve at various points along the curve. It was created by running separate regressions for each similarly ranked group of editions for the complete (Adult Fiction plus Juvenile) data pooled over the entire 2004-2016 interval (although the range of coefficients is very similar when individual years or weeks are examined). The table tells us, for example, that when the regression is limited to the 486,879 [14] editions selling between 55 [ln(4)] units and 148 [ln(5)] units in a year, the slope coefficient is 2.97. The better selling books with low ranks have a slope coefficient that is small, indicating a relatively flat relationship between sales and ranks, whereas the lower selling books have a much steeper relationship between rank and sales.

Table 1 reveals a considerable range, almost 8:1, between the largest and smallest slope coefficients which vary from 0.54 to 3.91. This slope is somewhat lower than the average of this ratio when computed for the individual years.[15] This changing slope over different portions of the distribution, of course, is also what Figure 2 revealed: these coefficients do not come from a straight line. Only the bottom three rows of the very best selling editions indicate relative linearity, and those editions account for only 13% of overall sales. And as we will show in Table 8 below, even changes in the coefficient of a few tenths can have very large changes in predicted sales. Nevertheless, if

[13] The growth in editions was 128% whereas the growth in sales was 14%.

[14] When the titles are pooled in this manner each edition exists as multiple observations depending on how many years that edition has positive sales in our data.

[15] The average value of the ratio for the individual years is 8.9 with the highest value of 14.3 in 2005 and the lowest value of 7.3 in 2014.

researchers are interested in only a very small portion of the distribution, say the market sales made up of bestsellers, and take their samples only from that portion of the distribution, the assumption of linearity might not cause large problems for the conversion of ranks to sales, about which we will have more to say below. However, researchers frequently don't have access to data from the bestselling products, as in the case of some of the influential early estimates that we mentioned in the introduction which used unrepresentative samples including the better selling books but excluding the bestselling books.

Table 1: Coefficients Along the Curve for Complete Data 2004-2016

We find similarly concave shapes for each of our 13 years, for each week we have examined, and for each of the separate subcategories of Juvenile Fiction, Juvenile Nonfiction, and Adult Fiction. We also find similar concave shapes for each fiction major genre, for each major publisher, and for each major binding type (paperback, mass market paper, hardcover). Similarly, we find concavity for Nonfiction editions down to the point where the 250k limit begins to bind.

The use of the rank-substitution methodology has become a tradition that has continued through the current moment, in spite of some earlier evidence hinting at its problems. BHS, in an unpublished 2010 working paper using a larger than average but still small sample of observations, found a nonlinearity, although they do not examine the extent to which the nonlinearity would affect results estimated under the assumption of linearity. Similarly, Duch-Brown and Martens, in a 2014 working paper, find a similar looking concave relationship between log sales and log rank for purchased downloaded music (e.g., iTunes).

VI. Explaining Early Results from the Rank-Substitution Literature

With the assumption of a linear log-rank/log-sales relationship holding throughout the full range of values, it has been common to just run a simple OLS regression of log sales on log rank (or vice versa) to measure the (assumed linear) rank/frequency slope coefficient. Although there has been some criticism of this general method, it does not seem to have affected the rank-substitution literature.[16]

Table 2 presents the slope coefficients (adjusted to be positive) that form the basis for the early rank-substitution literature, and the range of slope coefficients found here is also where most later slope coefficient found in the literature fall. The range in Table 2 runs from 0.671 to 1.111. The authors substituting ranks for sales take comfort from the fact that most slope estimates in this literature seem close to one another. CG state (p 209) "Thus all of these experiments suggest fairly consistent estimates of [slope coefficient] Ɵ in the relatively tight range of 0.77 to 1.11 [0.9 to 1.3]."[17] BHS state (p. 1587) "These estimates also favorably compare with Pareto slope parameter estimates obtained by Chevalier and Goolsbee…Weingarten (2001) and Poynter (2000) ...It is significant that ...these parameter estimates …are remarkably similar."

Table 2: Early Estimates of the [-] Sales Rank Slope coefficient

Several of the slope coefficients in this table are actually consistent with the concave shape found in Figure 2. For example, the two Allwine and Schnapp slope coefficients in Table 2 indicate a larger

[16] The use of OLS to measure the rank/frequency slope has been called into question by Goldstein, Morris and Yen (2004) who suggest, based on numerical simulations, that using OLS to measure the slope of the line provides biased results. Gabaix (2009) suggests that the standard error is calculated incorrectly. Newman (2005) and Clauset et al. (2009) argue that OLS is a poor technique for measuring the slope of the line, proposing a maximum likelihood estimator instead (since OLS is not equivalent to a maximum likelihood estimate when the data come from a non-normal distribution). Nevertheless, the rank-substitution literature assumes that the power laws hold throughout the full distribution.

[17] We have inverted the numbers in the quote (original numbers in brackets) to match the Zipf convention of slope coefficient that we are using throughout the paper.

value for the sample that included poorer selling editions (found on the steeper right-hand portion of Figure 2) as opposed to the smaller sample excluding poor selling editions. Similarly, the CG coefficient from best-selling ‘offline’ data, which would be on the leftmost portion of the curve, is flatter than any of the other estimates in Table 2, which do not consist of best-sellers.

None of the estimates of the slope coefficients shown in Table 2 are based on the full range of book editions. We know this because the empirical methodology used to estimate these coefficients are heavily weighted toward moderately successful editions.[18] Of course, if the rank-sales relationship were linear, it would not matter from which portion of the distribution a sample was taken.

Using our far more complete data, Table 3 provides slope coefficients from regressions over various book classifications during the year 2009, which was chosen because it is prior to when the 250,000 constraint starts to bind, it is in the middle of our data period, and it is prior to eBooks becoming important.[19] The slope coefficients from such regressions are all statistically significant and the R-squareds are all very high, as is found the rank-substitution literature. But of greater interest, all these coefficients are considerably larger than those in Table 2.

Table 3: Regression of logsales on logrank, 2009

The first three rows in Table 3 show the slope coefficient for the three major genres for which we have complete data, Adult Fiction, Juvenile Fiction, and Juvenile Nonfiction. The coefficients are

[18] The CG experiment altering weekly Amazon sales of an edition from 14 to 20 units, when converted to yearly sales of the edition from all retail outlets (as opposed to just Internet sales which made up less than 5% of overall sales at the time) was the equivalent of changing yearly sales from 14,600 units to 20,800 units, putting the edition in the top 5% of all editions by sales. The CG offline data (i.e., not their experiment) consists of bestsellers only, which clearly are not typical editions. The Weingarten experiment, based on altering the one-day sales ranking, increased sales by an annualized amount (adjusted for Amazon’s small share) of an additional 182,500 copies over the course of a year. Thus, these analyses are based on editions selling many more copies than the median edition’s sales of 8 copies a year.

[19] The slope coefficient for the 2009 merged juvenile and adult fiction categories is 2.54. The coefficient for the merged data averaged over the 13 years of data is 2.45. There is little change in slope coefficients during 2004-2009 but beginning in 2010 the coefficients decline until reaching 2.31 in 2016.

fairly close in magnitude.[20] Rows 4-6 in Table 3 provide information from populations that limit the number of slow-selling editions in some way and thus have lower coefficients. The fourth row provides *weekly* sales information in early 2009 for Adult Fiction.[21] Note that there are only 66,180 editions in this weekly population although the yearly data for Adult Fiction contains 226,920 editions. The smaller weekly number of editions can be viewed as a form of truncation due to the exclusion of editions that do not sell any copies in a particular week.[22] If editions had sales measured as a weekly moving average based on a longer interval, providing fractional values such as 1/3 of a sale per week, say, many more slow selling editions would remain in the weekly sample, leading to higher slope coefficients as the regression analysis gets populated with more observations in the steep portion of the relationship.

The fifth and sixth rows provide the slopes based on populations truncated at the low sales level of 69 copies a year where one truncation (Adult Nonfiction works) is unavoidable due to the 250,000 limit on yearly editions and the other (Adult Fiction) is intentionally created to match the truncation in the Adult Nonfiction category. The similarity of the coefficients strongly implies that if we had the complete population of Adult Nonfiction, it's slope coefficient would be similar to those found in the first three rows.

The linear assumption underlying the rank-substitution literature implies that truncating the data would have no effect on the slope coefficient since any portion of the straight-line rank-sales plot

[20] The size of these coefficients varies somewhat across different populations or different years. The Adult Fiction coefficients over the 13-year period of yearly data range from 2.14 to 2.66, with the larger values occurring in the earlier years and falling throughout the period. Coefficients for Juvenile range from 2.22 to 2.58 with a decline over the period.
[21] This coefficient come from the seventh week of 2009. The other coefficients for weekly data in 2009 fit in a rather tight bound. The smallest of the coefficients is 1.66 and the largest is 1.82, all well below the yearly coefficient.
[22] The difference in number of titles is caused by the fact that the majority of editions published in a year sell fewer than 52 copies, meaning that many editions will not have any sales in a particular week. For example, in 2009 only 29% of fiction editions sold more than 51 copies in a year, implying that at least 71% of yearly editions will not have positive sales during every week of the year. Editions selling many copies during the year, by way of comparison, will tend to sell copies every week and not drop out of the weekly samples.

should have the same slope as any other. But we now know that the removal of low selling editions from the sample, as is done when slow sellers are left out of the sample, lowers the values of these coefficients. Because the coefficients in Table 2 are based on time intervals much short than a year, they have lower values than those in Table 3, due to the removal of many slow selling editions from the sample. The different coefficients between Tables 2 and 3 make perfect sense and are entirely consistent with one another when it is understood that the rank-sales plot is curved and not linear.

VII. Prediction Errors Due to Linearity Assumption

Predicting values on a concave line, when a straight line had been assumed, clearly would be expected to lead to errors that are larger than would be found if the actual relationship were a straight line. Given the data that we are dealing with, we can describe the nature of the errors that are likely to arise, and these are illustrated in Figure 3. The blue dots, which tend to form a smooth curve except at the edges, are sales and rank data for all the Fiction and Juvenile editions in 2009.

If we run a linear regression such as that in equation (6), we get predicted values of sales based on ranks very much like the steep gold line in Figure 3. The gold line continues beyond the confines of this figure to reach the sales axis at a value of about 275 trillion units (not visible in the figure). The very large number of book editions on the right side of the figure lead to the steep estimates of the slope coefficient as found in the top rows of Table 3.

This gold line makes it clear that there are two main sources of prediction error due to the concavity of the relationship. On the left-hand side of the figure we see that a small number of best-selling editions are predicted to sell very many more copies per edition than they actually sell. Since the bestselling edition actually sells somewhat more than 2 million copies, it is clear that there will be major errors in predicting total industry sales and that average residuals will be very large for better selling editions. On the right-hand portion of the figure, another source of prediction error comes from the sales of the slower selling editions which are predicted to sell fewer copies per edition and although the deviations are small there are very many more such editions.[23] Overall, as we empirically demonstrate below, predicted sales on the gold line are very far from actual sales for the industry as a whole, dominated by the enormous errors from the bestselling editions.

Figure 3: The tradeoffs when using linear methods to predict values on a curve

Figure 3 also shows an illustrative flatter red line, which consists of the predicted values (many out of sample) from a linear regression run on the top 45k better selling editions. The slope is less steep, leading to smaller errors for better selling editions (compared to the gold line), but the errors for the slowest selling editions are considerably higher than was the case for the gold line, and there are great many of those editions. We take up issues involved with such truncated data in Section X but want now to focus on the problems using the complete data represented by the gold line.

We have already seen how the nonlinearity of the rank-sales plot is likely to cause estimates of its presumed linear slope to have very different results depending on from where in the distribution the sample was being taken or the time interval over which the measurements occur. But more important than the slope coefficients are the predicted values of sales given ranks, the centerpiece of the rank-substitution literature. After all, estimating the slope coefficient was merely a means to an end, with the end being the conversion of an edition's rank into an estimate of that edition's sales.

[23] If this minimization of the sum of squared residuals used in the regression line seems counter-intuitive, remember that the regression is minimizing the sum of the squared residuals of logged values so that the any decrease in residuals on the left-hand side from a flatter estimated line, say, based as they are on logged values, are balanced by the increased residuals from the much larger number of relatively small squared residuals on the right-hand side. The much greater misestimates on the left-hand side that seem unbalanced compared to the smaller misestimates on the right only appear that way when the unlogged values are compared.

If we are going to evaluate the degree of measurement error, we need to create measures of "how close" the estimates are from the actual sales values in a way that is useful for researchers contemplating the use of ranks to predict sales quantities. The most natural measurement is to compute the 'residual' between the values of the predicted (not logged) unit sales and the actual unit sales. The sign of the residual is irrelevant when measuring accuracy, so we will use the absolute value of these residuals. Summing the absolute values of residuals provides our first measure of accuracy, A1, as shown in (7). This measure is most useful when comparing predicted values over the same set of observations for different methods of converting ranks to sales.

(7) $$A1 \equiv \sum_{i=1}^{n} |Resid_i|$$

When the samples or populations differ, we will want to normalize the measure of accuracy to allow useful comparisons. For one thing, we need to adjust for the number of observations since larger populations will tend to have a larger sum of absolute residuals, all else equal. We also wouldn't expect the residuals of million-selling books to be the same as the residuals for books selling one or two units and that also need to be accounted for.

For these reasons, we will take the ratio of the residual relative to the actual value. We can do this for the individual observations or in the aggregate. First, we calculate the ratios of residual/sales for individual editions, which we refer to as $A2_i$.

(8) $$A2_i \equiv \frac{Resid_i}{Sales_i}$$

Then we take either the mean or median value of the $A2_i$s. $A2_i$ appears to be the measure that Schnapp and Allwine (2001) used when they argued that the predicted sales values of books in their sample were close to the actual values.

We can also form a ratio with the sum of absolute residuals as the numerator and the total sales as the denominator. We refer to this measure as A3.[24]

$$(9) \qquad A3 \equiv \frac{\sum_{i=1}^{n} |Resid_i|}{\sum_{i=1}^{n} Sales_i}$$

With these measures in hand, we can address how well or how badly various methods work when trying to predict sales from ranks.

We begin our analysis by estimating the accuracy scores when a linear estimation technique is imposed on a nonlinear log sales log rank relationship, using 2009 data although any other year would provide similar results. Instead of using an unbiased random sample, as researchers with limited data would normally attempt to do, we use the entire population.

We run the simple regression in equation (6) just as others in this literature have done. As we have already seen in Figure 3, we should expect some highly inaccurate predictions of sales. Table 4 provides the details of these inaccuracies in terms of the unlogged values. The first row of Table 4 provides A1, the sum of the absolute value of residuals, where residuals are measured in terms of the number of books sold. The value rounds to an extraordinarily high 366 trillion books. This result is driven by a small number of the best-selling editions where, for example, the largest residual for an individual book edition is greater than 275 trillion copies.

The second row indicates that the residual/sales ratio, averaged over the 424,653 editions, is 33,640%, indicating a massive level of mismeasurement. The edition with the largest ratio of residual to sales has a value of over 95 million (although there are also very many editions with very small ratios). Such large ‘outliers’ have an undue influence on the overall average but even the

[24] A3 is identical to the percentage difference between the predicted sales and actual sales if the difference between predicted and actual sales is the same sign for each observation.

median value of .5 is large since this means that half of the observations had predicted values that differed from the actual value by at least 50%.

Table 4: Simple Regression, 2009 All Juvenile and Adult Fiction, 424,653 obs

The third row indicates the size of the aggregated residuals relative to total sales. The level of mismeasurement at the aggregate level is much greater than for the average over individual editions since this measure is barely influenced by the very large number editions with low residuals. The predicted and actual industry sales associated with these results are shown in the fourth and fifth row. The actual number of books sold (fourth row) is almost 424 million units whereas the fifth row indicates that the predicted sum of books sold is seen to be an astonishing 366 trillion books in 2009, which would absurdly amount to roughly a million books purchased per person per year in the US.

In other words, the current methodology to convert ranks to sales can provide extremely flawed, results if the sample used to estimate slope coefficients is random and unbiased, thus mimicking our population. Equivalent results are found for separate components of the industry (Adult Fiction, Juvenile Fiction and Nonfiction) and for multiple years as well.

Nor is this problem avoided by those studies which do not convert logged sales back to ordinary sales, a procedure followed in studies focused on elasticities. Although the percentage measure of deviation for logged values may not be as large as for the unlogged values, the elasticity estimates can still be off to the same extent that the slope coefficient is mismeasured. In Table 1, the range of the rank-sales slope coefficients was seen to be almost 8:1, but in particular years, such as 2005, the ratio between the larger and smaller slope coefficients is greater than 14:1. Thus the predicted values of log sales are also potentially unreliable even if not as strongly unreliable as the predicted value of sales.

These results should make it obvious that a researcher using random samples risks generating very misleading estimates of the industry size and the sales quantities of editions. But we have already noted that previous researchers did not use unbiased random samples because they thought the linearity assumption removed the need for such a sample and thus they used, quite by accident, samples that contained more of the successful editions.

Additionally, as Figure 1 make clear, the top 5%-10% of editions are responsible for the large majority of sales, so it is natural to wonder if eliminating many of the smallest selling editions might provide for more accurate estimation of sales for the editions which are most important to characterizing the overall industry. Perhaps the errors that come from using unbiased samples are fortuitously ameliorated through the use of biased samples favoring better selling editions.

We return to this general question after exploring whether non-linear estimation can provide a superior method to predict sales from ranks and examining two instances of how actual predictions are affected by the curvature of the log-rank/log-sales relationship.

VIII. Overcoming these problems through nonlinear estimation

Given the nonlinear empirical relationship between (logged) sales and (logged) ranks it might seem natural to change the regression specification to allow for curvature. We compared the performance of various nonlinear models, adding polynomial terms to the regression model (e.g., adding an additional quadratic term) or using spline regressions.[25] From this comparison we found that adding a polynomial term to the power of five performed best and thus present these results in what

[25] Spline regressions require a larger sample and provide poor out of sample estimates compared to polynomials.

follows.[26] Since there is no commonly accepted terminology to describe raising the power of a variable to the fifth power, we will merely refer to this as the polynomial specification in the text below. Specifically, we run a regression of the form:

$$\text{Ln(sales)} = \beta 1 + \beta 2 \cdot \ln(\text{rank}) + \beta 3 \cdot (\ln(\text{rank}))^5 + \varepsilon \tag{10}$$

Figure 4: Raw 2009 Data and Nonlinear Fitted Values

The estimated nonlinear relationship and the raw data for adult fiction and juvenile editions in 2009 are shown in Figure 4. It is clear from the figure that the sales of the bestselling editions are underestimated, and this will lead to relatively large residuals for the top few hundred high-selling editions, with the predicted values for these editions lower than the actual values (whereas they were higher in the linear specification). Sales of smaller selling editions, however, are closely matched.

This intuition is confirmed by Table 5 which provides the identical accuracy measurements as those in Table 4, but is based on the polynomial specification. The errors in the fitted values are far lower than was the case for the linear regression. Clearly, despite the relatively poor accuracy for best-selling editions, the overall prediction of sales for the 425,653 editions is quite good, particularly compared to the linear results in Table 4.

Table 5: Polynomial Specification, 2009, 424,653 obs

[26] We are not claiming that n=5 is in anyway ideal, but merely that it performs somewhat better than n=2. We found that the average residual (A2) was about twice as high using n=2 as opposed to n=5. By way of comparison, as we shall soon illustrate, the use of a nonlinear as opposed to a linear specification led to a much larger improvement, usually by an order of magnitude.

The average residual/sales ratio (A2) is much smaller than for the linear case, having a value of 5% (as opposed to 33,640% in Table 4) and a median value of 3.5% (as opposed to 50%), which appears to be in the ballpark for the precision touted by Schnapp and Allwine and the authors in the literature that followed. The estimate of total industry sales is too small by only about 12%. We note that the edition with the largest ratio of residuals/sales has a value of .98, which is many orders of magnitude below the largest values for the linear specification. Clearly, these are much more reasonable estimates than was the case for the linear specification.

Our results above are based on using sales and rank data for the market population of books although researchers typically only have access to sales and rank data for a sample, usually a fairly small sample, of the items in their market of interest, otherwise they could just use the sales data they have and not need to estimate them. This raises the question of how accurate the polynomial estimation of sales from ranks will be when using a sample. We can use our population level data to simulate the accuracy of predicting sales from ranks using different models with synthetic samples of varying sizes from our population. Our analysis indicates that if researchers can get a moderate number of observations, several dozen to one hundred observations are all that are required, they may be able to use simple non-linear estimation to predict sales with reasonable accuracy based on the ranks of book editions.

In Table 6 we show results from Monte Carlo simulations exploring prediction accuracy (measured by A2) using the polynomial model. We run our regressions for 1,000 random samples of different sizes to estimate the relationship between sales and ranks and then use the estimated relationship calculated for each sample to compute the prediction error for each observation in the full population so as to find the average value of A2 for that run. After calculating these A2 values for each of 1,000 runs, we then find the median, mean, min, max, and 95th and 99th percentile values over those 1,000 A2 values. In Appendix 2, we illustrate such simulations for the linear model, which, not surprisingly, performs very poorly.

The average and median values of A2 in the samples are quite small, even with samples as small as 20 observations. With a sample of only 20 observations there is a 1% chance of the average A2 being greater than 56.4%, although the odds are 50-50 that it will be smaller than 5.3%. With a sample of 50 editions there is a 99% likelihood that the average A2 will be less than 8.7% although the simulations indicate that there is a small chance that the value might be as high as 16.4%. With 100 editions in the sample the worst result in 1,000 runs is an A2 value of 7.9%. There are minor improvements when the sample is increased but samples sizes of about 100, or even 50, appear to provide results that many researchers are likely to consider "accurate enough" for their purposes.[27]

Table 6: Monte Carlo Simulations of Average A2 –Year 2009 – 1,000 Iterations

Another complication regarding sampling methods is that researchers rarely have access to sales and rank data over the full range of values, including the highest-selling and lowest-selling editions. For example, BHS (2003) use data from a sample of books selling between 1 and 481 copies a week which is not a representative sample because they exclude very slow selling editions and very high selling editions. Therefore, we also examine how well random samples will do when the samples are taken from somewhat symmetrically truncated section of the population. The results are found in Table 7 which uses random samples of books within a range of 10 to 250,000 units sold.[28]

Table 7: A2 Values, Sales Truncated to range 10-250,000, 2009, 1,000 Iterations

[27] We also explored the prediction error using data from 2004 and 2016 and the conclusions are similar.

[28] As in Table 6, we run our regressions for 1,000 random samples of different sizes to estimate the relationship between sales and ranks and then use the estimated relationship calculated for each sample to compute the prediction error for each observation in the full population so as to find the average value of A2 for the run.

The overall accuracy (A2) is poorer when the samples are truncated, as one would suspect, but remains at the seemingly reasonable level of about 10 percent. Samples of 50 are only slightly less accurate than those of 2,500. We conclude that polynomial regressions, even with this level of truncation in the data, continue to perform quite well.

IX. Some Actual Prediction Errors in the Literature

We have seen that the red fitted line in Figure 3, based on data biased in favor of better selling editions, has sales predictions for the high-selling editions that are much closer to the actual values than is the case with the gold fitted line based on unbiased data. The flip side is that the red line's predicted values for the slow selling editions will be considerably higher than their true values.

This flip side is at work in the notable BHS (2003) result that the Internet was greatly increasing social surplus due to the much larger variety of obscure products it made available, leading to increased sales from these obscure editions. The results from that analysis are greatly influenced by the size of the estimated linear slope coefficient, as demonstrated in Table 8.[29]

Table 8: Predicted Market Share of Obscure Editions for Different Slope Coefficient Values

Along the header row we find four possible the slope coefficients. In column 1 we find the coefficient BHS derived from their dataset, as found in Table 2. We then provide three other possible slope coefficient, the last two of which better reflect the values that would have been found if BHS had a random unbiased sample that would be expected to mimic the results found for the nontruncated populations found in Table 3. Column 2 uses the value from Schnapp and Allwine's

[29] These results are derived using BHS's equation 13. The slope coefficient of the (presumed) linear rank/sales relationship provides theoretic sales quantities for every edition based on its rank, so the theoretic market share of sales for editions of any rank can be easily calculated.

larger sample. Column 3 uses the value for weekly adult fiction found in Table 3, and column 4 uses a value approximately equal to the yearly population slope coefficients found in Table 3.

Using the BHS slope coefficient in column 1, the predicted market shares of obscure editions according to BHS (2003) is 29.33%, a value much higher than the actual offline shares of such editions. This highly cited simulation is taken to imply that consumers prefer a larger variety of books if only given the chance to purchase them online, generating a potential large increase in consumer surplus. But as the next column shows, this simulated value will markedly change for even a relatively minor change of slope coefficient. For example, the higher value from Schnapp and Allwine found in column 2 (which is within Chevalier and Goolsbee's "tight range" as shown in Table 2) results in the share of obscure editions being smaller by a factor of around 4. This would mean that BHS's estimate of social welfare increases from Internet book sales would also be too large by a factor of 4.

The coefficients in columns 3 and 4 of Table 8, based as they are on the population of books, have unbiased and somewhat larger slope coefficients. When a slope coefficient of 1.7 is used (based on weekly sales), the market share for the 'obscure' editions and the concomitant social gain from Internet induced variety, would be lower than the BHS result by a factor of several thousand. When the yearly slope coefficient (2.5) is used, BHS' results are too large by a factor of several million, as seen in column 4.

The reason for the BHS overestimate is easy to understand. The coefficient used in the BHS calculation, 0.871, is somewhat flatter than the red fitted line in Figure 3. We have already noted that the predicted sales represented by such a line will lead to large overestimates of the sales of the slow selling editions on the right-hand side of the diagram. But the very exercise that BHS are conducting is based on the predicted sales of these low-selling editions and any overestimate in the shares of these editions will distort their conclusions.[30]

Note that we are *not* suggesting that the BHS result is off by a thousand or a million as might seem to be implied by the last two columns of Table 8. Because all the values in Table 8 are based on the incorrect power law assumption that sales are linearly related to ranks, they are all untrustworthy.[31]

A completely different empirical estimate of sales based on ranks can be found in Ferreira and Waldfogel (2013). They (FW) look at the list of songs making up charts of top 20, top 50 or top 100, for different countries and different time periods. They wish to convert these ranks into sales of the individual songs so they can attribute the sales to the nationality of the singer/group. They note the results of Chevalier and Goolsbee (2003) and BHS (2003), as well as their own examination of record sales in South Korea. Their discussion, as elsewhere in the literature, essentially treats the slope coefficients for books and music as interchangeable. They chose a slope coefficient of 1 (Zipf's law) when converting ranks to sales.

Because FW are unaware of the curvature of the relationship, they do not realize that the value they generate is somewhat out of line since best-sellers tend to have slope coefficients closer to .5 (as seen in Table 1).

[30] Another issue facing BHS and other researchers in this literature is determining the number of editions in the market. BHS conclude that the total number of book editions is 2.3 million which they state is the number of editions in print. Their analysis apparently assumes that all of the books in print are actually sold in the time period of interest, which is generally not the case as reported by Liebowitz (2023). If BHS had chosen 1 million as the number of editions sold, instead of 2.3 million, the shares (compared to column 1 of Table 5) would have been 20% (instead of 29%), so the assumption that all listed editions are sold clearly affects the results.

[31] A recent estimate of the share of obscure editions in an Internet enabled world is found in Liebowitz, Ward, and Zentner (2023). The most likely impact of online access was to increase the sale of obscure editions by 2%-6%, implying that the BHS predicted values are too high by a factor of about 10. Quan and Williams (2018) also find a much smaller gain in surplus due to additional variety, although they look at shoes, not books.

Table 9: Accuracy of FW predicted shares

How accurate is their methodology? We cannot know for certain since neither they nor we have record sales data for the countries they examined. But we can use their methodology on our weekly book sales data to gauge how accurate their sales predictions would be if books could be interchanged for sound recordings, as FW assume can be done.

One of the magnitudes examined by FW is the share of the top 10, top 20, or top 50 recordings relative to the top 100. Those results are reproduced in Column 1 of Table 9. Column 2 represents the actual share of the top 20 or 50 editions. There is a considerable difference between the predicted shares and actual shares. Column 3 presents results using a more relevant slope coefficient of 0.5, which provides much more accurate predictions.

We can dig a little further into these predicted values and instead of looking at relative shares of groupings, examine the absolute value of the deviation of actual sales of individual units from the values predicted by the slope coefficient of 1. Figure 5 reveals that the errors range from 6% to 93% for the top 10 editions and increase from there up to above 250% for the more poorly ranked editions in the top 100. It is easy to imagine these levels of error leading to misleading results in many circumstances.

Figure 5 Accuracy of FW Predicted Sales for Editions ranked 2-100

Fortunately for FW, the actual level of song sales plays almost no role in their results. The position in their music charts by nationality appear to be fairly random so that it matters little what the actual sales values are. FW even state that their results would hardly change if all songs in the charts were given the identical sales value.[32] In a more recent paper citing our findings, Reimers and Waldfogel (2021) use a more reasonable slope coefficient of 0.47 for top selling books.

X. Can a narrow focus on better selling books redeem the linear method?

The log rank log sales relationship is concave, which means that sales estimates based on the assumption of a linear relationship are likely to be inaccurate. This has been demonstrated both in general and in the specific cases discussed. To address this issue, a polynomial specification can be used, as we have seen.

It is possible to consider the possibility that a linear analysis with a narrow focus on leading editions may produce reasonably accurate results under certain conditions. This is because there is not a great deal of curvature in any narrow portion of the distribution and the sales share of a relatively small number of top-selling editions typically accounts for the majority of sales, as shown in Figure 1. Additionally, the top-selling editions in Figure 3 appears to be on a relatively flat portion of the curve. Therefore, it might be thought that focusing on this part of the distribution may allow for relatively accurate predictions for editions that make up a large portion of industry sales.

There is also a body of research supporting the belief that only the tail of the distribution representing the largest items of an ordered series are likely to appear to follow a power law, even if items are drawn from a power law distribution.[33] For instance, Clauset et al. (2009), in a highly influential paper, make this argument and provide an algorithm for determining a cutoff value beyond which observations in a series no longer appear to follow a power law. When this algorithm was applied to our data (as implemented by Brzezinski [2014]), this cutoff value occurred at very

[32] This claim is found in the 2010 working paper version of their published paper, in the section on robustness checks.

[33] Newman (2005), using random draws from a power law distribution, concludes that only the larger values (e.g., better selling) reveal their power law origin. Reed (2001), however, claims that both tails empirically exhibit power law behavior for male earnings and US settlement (towns) sizes.

low rankings (averaging 1,761 out of approximately half a million editions per year) for each of the 13 years of data, with the average sales of the cutoff edition being around 65,000 copies per year.[34] For 2009, the Clauset method generated a cutoff rank of 1,396, indicating that sales of editions beyond this rank (with yearly sales less than 43,001) do not appear to come from a power law distribution.

Obviously, this method removes more than just the very small sellers. In fact, it removes 99.67% of editions which are responsible for 62% of sales in 2009, and over the years of our data it removes an average of 67% of industry sales.[35] Ignoring such a large portion of the market appears to be a case of throwing out the baby with the bathwater for any analyst interested in other than just the very top selling editions. But for analysts who are interested only in top sellers comprising a minority share of industry sales, however, it seems that the linear rank-substitution method might provide useful sales data.

For example, Table 10 reports the accuracy of estimating sales from ranks in 2009 when running the linear specification estimated from this small portion of the population using the Clauset et al. method to find a cutoff point. When examining the accuracy of predicted sales over only the 1,395 top selling editions, A2 and A3 values are quite small, as seen in column 1, indicating accurately predicted sales from ranks (although we must note that the values in Table 5, using a polynomial specification, are not that much worse in the much more demanding instance of predicting sales over the entire industry sales distribution).[36] But if we try to predict sales for the full industry using

[34] The lowest value for the cutoff ranking is 193 in 2015 and the highest value is 10,383 in 2006. The lowest cutoff sales value is 6,861 in 2006 and the highest is 149,834 in 2015.

[35] The smallest share of sales removed is 30% in 2006 and the largest is 86% in 2015.

[36] Using polynomial specifications often improves the within sample accuracy of Clauset et al. samples, as might be expected, but care needs to be taken because the exponentiated term, which is always negative when using larger portions of the distribution, is sometimes positive for the seemingly linear part of the curve identified by the Clauset et

this limited sample, the predictions are extremely poor, as seen in column 2. Nevertheless, where analysts are only interested in the top selling editions, they might not care about out-of-sample predictions for the entire industry.

Table 10: Accuracy of Predicted Sales from Ranks using top 1,395 ranks, 2009
Linear Specification

Because the Clauset et al. method throws out too much information for a general examination of the industry, it seems worth investigating whether a truncation of slower selling editions, somewhat less restrictive than that proposed by Clauset et al., might provide accurate sales predictions and still cover enough of the sales in the industry to satisfy researchers. The use of a wider portion of the sales distribution will introduce more curvature into the sample and decrease the accuracy of the within sample predicted sales compared to the Clauset et al. method.[37] But there might be a sweet spot where the within sample share of sales is large enough to represent the entire industry and the sales estimates remain reasonably accurate.

How much industry data would an analyst feel comfortable ignoring, hoping that it doesn't matter to the results? Ignoring the smallest observations representing only a few percentage points of sales, say, might be acceptable to many analysts unless the focus of the analysis was on the smallest selling observations, such as when studying the long tail hypothesis. But, on the other hand, ignoring a large fraction of sales, as is the case when the Clauset method jettisons a majority of sales, is unlikely to be considered acceptable.

al. methodology. The exponentiated term is sometimes measured as being significantly positive (2007, 2008, 2012, 2015), and in those years predictions outside the sample are far more wildly inaccurate and there is no improvement in accuracy by the nonlinear specification within sample.

[37] In general, as more of the sample is added we get more curvature. The estimated log rank log sale line will be steeper as additional, slower selling observations are added, meaning that the overestimated predicted sales will be smaller for the slower selling editions but higher for the higher selling editions.

Table 11: Tradeoffs from Truncating Sample to only Include Larger Sellers, 2009

Table 11 illustrates the tradeoffs that an analyst may encounter when attempting to estimate sales from ranks. The first two unnumbered rows show the quantity of top-selling editions that the analysis is based on, and the percentage of industry sales that these editions represent, for different truncations in the year 2009. Rows 1-2 show the predicted sales errors for linear estimation, using either the A2 or A3 measurement (the percentage difference in predicted versus actual sales is very close to the A3 values so we do not show it separately). Rows 3-4 demonstrate the improved accuracy of nonlinear estimation using our polynomial model. It's clear that nonlinear estimation is more accurate than linear estimation, usually by a factor of at least 5 or 10.

The data restrictions reported in the first two columns exclude approximately 1% and 5% of sales, respectively, likely to be considered an acceptable loss by many researchers unless the focus is on the tail. The problem with using these limited data restrictions is that the predicted sales are highly inaccurate. In the first column, the A3 error is exceptionally large, indicating that the overall predicted sales errors are 3,300% higher than actual sales, and the average misprediction of sales per edition is around 46%. These errors are too large to be acceptable. While the errors in column 2 are smaller, they are still quite large, especially the A3 errors. The nonlinear estimates in both of these instances have much smaller errors, making the use of these samples much more viable.

For the table as a whole, we see that the errors in predicted sales decrease as we move from left to right, but the percentage of excluded sales increases.[38] This reflects the tradeoff that analysts must weigh when using the linear rank-substitution method. Is there a goldilocks "sweet spot" in the

[38] Although the predicted sales errors increase going from right to left in the table, it is *not* the case the prediction errors always increase as we include a larger share of industry editions beyond the range of this table.

linear analysis where the included market share is not too small and the errors in predicted sales are not too large? That will depend on the preferences of the researcher. Do the lower sales misestimates in the last three columns justify the exclusion of 15%-30% of industry sales? While we cannot provide a definitive answer to this question, we note that this is a choice that does not need to be made. The nonlinear specification allows equivalently small misestimates of sales without having to remove a large share of industry sales, so that is obviously the path to choose.

For previous linear estimates, as opposed to new research, might those results be relatively accurate because they were, through luck, in some sort of “sweet spot”? It's difficult to say for certain, but it is a possibility, although an unlikely one. Rerunning past estimates using a nonlinear specification would be a simple way to test for that possibility.

XI. Conclusion

Accurate prediction of sales for products based on their ranks requires a predictable relationship between ranks and sales. Previous research using the book industry has assumed that such a relationship exists and follows a power law, implying a linear relationship between the log of ranks and the log of sales. Support for this power law assumption comes from the enormous range of sales values and a highly skewed distribution leading to a small number of bestselling books being responsible for a large share of sales.

This power law assumption has been used in hundreds of research papers, but our analysis using full data sets for the book industry shows that the relationship is actually concave, not linear. Our findings hold for each year of the data, each individual week examined, and for each genre of book. This nonlinear relationship can lead to significant distortions when using linear statistical techniques to estimate sales based on ranks. Nonlinear estimation methods provide superior results.

The nonlinear nature of the relationship between book sales and rank means that the slope of the line representing this relationship will vary depending on the portion of the sales distribution being considered. This further implies that short time periods such as a week will have a flatter average slope compared to longer time periods such as a year because many slow-selling books are excluded from weekly sales figures. These findings have implications for previous research that used rank-substitution methods, as such studies may not be reliable unless they only focused on a narrow range of the distribution (e.g., bestsellers). Because these problems are due to the erroneously assumed linear relationship, it is generally more reliable to use nonlinear methods to analyze this type of data.

It is worth noting that while these findings are specific to the book industry, it has been widely assumed that other industries with similar characteristics (e.g., music, software) may exhibit similar trends. However, further research is needed to confirm this.

Adamic, Lada, and Bernardo Huberman. "The nature of markets in the World Wide Web." *Quarterly Journal of Electronic Commerce* 1, no. 1 (2000): 5-12.

Adamic, Lada. “Zipf, Power-laws, and Pareto - a ranking tutorial” found at: http://www.hpl.hp.com/research/idl/papers/ranking/

Benner Mary J. and Joel Waldfogel (2016) “The Song Remains the Same? Technological Change and Positioning in the Recorded Music Industry.” *Strategy Science* 1(3):129-147.

Brynjolfsson, Erik, Yu Hu, and Michael D. Smith. "Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers." *Management Science* 49, no. 11 (2003): 1580-1596.

Brynjolfsson, Erik and Hu, Yu Jeffrey and Smith, Michael D., The Longer Tail: The Changing Shape of Amazon’s Sales Distribution Curve (September 20, 2010). Available at SSRN: https://ssrn.com/abstract=1679991

Brzezinski, Michal. "Do wealth distributions follow power laws? Evidence from ‘rich lists’." Physica A: *Statistical Mechanics and its Applications* 406 (2014): 155-162.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. "Power-law distributions in empirical data." *SIAM review* 51, no. 4 (2009): 661-703.

Chevalier, Judith, and Austan Goolsbee. "Measuring prices and price competition online: Amazon. com and BarnesandNoble. com." *Quantitative marketing and Economics* 1, no. 2 (2003): 203-222.

Chevalier, Judith A., and Dina Mayzlin. “The Effect of Word of Mouth on Sales: Online Book Reviews.” *Journal of Marketing Research*, vol. 43, no. 3, American Marketing Association, 2006, pp. 345–54.

Duch-Brown, Néstor, and Bertin Martens. Search costs, information exchange and sales concentration in the digital music industry. No. 2014/09. Institute for Prospective Technological Studies Digital Economy Working Paper, 2014.

Ferreira, Fernando, and Joel Waldfogel. "Pop internationalism: has half a century of world music trade displaced local culture?." *The Economic Journal* 123, no. 569 (2013): 634-664.

Gabaix, Xavier. "Power laws in economics and finance." *Annu. Rev. Econ.* 1, no. 1 (2009): 255-294.

Garg, Rajiv, and Rahul Telang. "Inferring app demand from publicly available data." *MIS quarterly* (2013): 1253-1264.

Goldstein, Michel L., Steven A. Morris, and Gary G. Yen. "Problems with fitting to the power-law distribution." *The European Physical Journal B-Condensed Matter and Complex Systems* 41, no. 2 (2004): 255-258.

Liebowitz, Stan J., and Alejandro Zentner. “The Pecuniary Motivation to Create.” 2020. Available at https://ssrn.com/abstract=3195384

Liebowitz, Stan J., Michael R. Ward, and Alejandro Zentner "Only a "Longish" Tail." 2023. Available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3770454

Liebowitz, Stan J. "Have we misunderstood the Copyright 'Tradeoff'? Copyright's impact on Sales, Prices, and Availability of Books." 2023.

Newman, Mark EJ. "Power laws, Pareto distributions and Zipf's law." *Contemporary physics* 46, no. 5 (2005): 323-351.

Pareto, V. (1897). Cours d'Economie Politique. F. Rouge, Lausanne.

Quan, Thomas W., and Kevin R. Williams. "Product variety, across-market demand heterogeneity, and the value of online retail." *The RAND Journal of Economics* 49, no. 4 (2018): 877-913.

Reed, William J. "The Pareto, Zipf and other power laws." *Economics letters* 74, no. 1 (2001): 15-19.

Reimers, Imke and Waldfogel, Joel 2021. "Digitization and Pre-purchase Information: The Causal and Welfare Impacts of Reviews and Crowd Ratings," *American Economic Review*, vol. 111(6), pages 1944-1971, June. Online Appendix: https://www.aeaweb.org/content/file?id=14673

Rosario, Ana Babic, Sotgiu, Francesca, De Valck, Kristine and Bijmolt, Tammo. "The Effect of Electronic Word of Mouth on Sales: A Meta-Analytic Review of Platform, Product, and Metric Factors." *Journal of Marketing Research* 53, no. 3 (2016): 297–318.

Schnapp, Madeline and Tim Allwine. (2001). ''Mining of book data from Amazon.com'', Presentation at the UCB/SIMS web mining conference, https://web.archive.org/web/20070710123934/http://www2.sims.berkeley.edu/resources/affiliates/workshops/webmining/Slides/ORA.ppt.

Wang, Quan, Beibei Li, and Param Vir Singh. "Copycats vs. original mobile apps: A machine learning copycat-detection method and empirical analysis." *Information Systems Research* 29, no. 2 (2018): 273-291.

Zipf, G. K. Human Behaviour and the Principle of Least Effort. Addison-Wesley, Reading, MA (1949).

Table 1: Coefficients Along the Curve for Complete Data 2004-2016

logsales interval	coefficient	editions	mean sales/edition	Min Sales	Max Sales	total sales
logsales 0-1	3.91	1,768,498	1.33	1	2	2,346,107
logsales 1-2	3.48	1,148,128	4.55	3	7	5,224,174
logsales 2-3	3.24	830,091	13	8	20	10,445,442
logsales 3-4	3.24	597,625	34	21	54	20,379,329
logsales 4-5	2.97	486,879	92	55	148	44,867,127
logsales 5-6	2.54	393,799	251	149	403	98,675,476
logsales 6-7	2.04	316,387	680	404	1,096	215,297,335
logsales 7-8	1.53	239,337	1,830	1,097	2,980	437,974,504
logsales 8-9	1.16	151,134	4,896	2,981	8,103	739,894,029
logsales 9-10	0.91	72,442	12,865	8,104	22,026	931,988,063
logsales 10-11	0.77	26,588	34,666	22,027	59,871	921,694,025
logsales 11-12	0.62	7,777	91,487	59,878	162,692	711,491,833
logsales 12-13	0.52	1,615	241,601	162,756	441,112	390,185,454
logsales 13-14	0.52	232	629,718	442,512	1,173,978	146,094,622
logsales 14-15	0.54	43	1,899,093	1,280,247	3,218,384	81,660,999

Table 2: Early Estimates of the [-] Sales Rank Slope coefficient

Sample	coefficient	R^2
Schnapp and Allwine larger sample [CG]	1.111	
Weingarten [CG]	0.952	
BHS [experiment]	0.916	
BHS [dataset]	0.871	0.80
CG Experiment	0.855	
Poynter [CG]	0.834	0.97
Schnapp and Allwine better selling small sample [CG]	0.769	
CG (offline bestselling data)	0.671	0.94

Table 3: Regression of logsales on logrank, 2009

	Slope Coef [-]	Std error	t-statistic	Constant	Adj Rsq	Observations
1. Yearly Adult Fiction	2.559	0.0015	1757	31.793	0.931	226,920
2. Yearly Juvenile Fiction	2.596	0.0025	1052	31.058	0.904	117,190
3. Yearly Juvenile Nonfiction	2.283	0.0021	1063	26.212	0.933	80,040
4. Weekly Adult Fiction	1.705	0.002	875	19.216	0.920	66,180
5. Yearly Adult Nonfiction Truncated	1.223	0.0005	2490	19.763	0.962	244,134
6. Yearly Adult Fiction Truncated	1.536	0.0019	792	21.916	0.912	60,447

Table 4: Simple Regression, 2009 All Juvenile and Adult Fiction, 424,653 obs

Accuracy Measure for predicted Sales	Value
A1: Total Abs Residual	366,275,716,972,544
A2: Average\Median Absolute Residual/Sales	33640% \ 50%
A3: Total Absolute Residual/Total Sales	86,396,600%
Actual Industry Unit Sales	423,946,802
Predicted Industry Unit Sales	366,276,086,071,296

Table 5: Polynomial Specification, 2009, 424,653 obs

Measurement	Value
A1: Total Abs Residual	122,430,942
A2: Average/Median Absolute Residual/Sales	5.0%/3.5%
A3: Total Absolute Residual/Total Sales	28.87%
Actual Industry Sales	423,946,802
Predicted Industry Sales	373,737,105

Table 6: Monte Carlo Simulations of Average A2 –Year 2009 – 1,000 Iterations

Sample Size	Median	Mean	Max	95th Percentile	99th Percentile
N=20	0.053	0.089	7.929	0.190	0.564
N=50	0.051	0.053	0.164	0.061	0.087
N=100	0.051	0.052	0.079	0.056	0.062
N=1,000	0.050	0.051	0.058	0.053	0.056
N=2,000	0.050	0.051	0.055	0.052	0.054

Table 7: A2 Values, Sales Truncated to range 10-250,000, 2009, 1,000 Iterations

Sample Size	Median	Mean	Max	95th Percentile	99th Percentile
N=50	0.108	0.108	0.146	0.132	0.139
N=100	0.104	0.106	0.138	0.125	0.131
N=1000	0.102	0.102	0.111	0.107	0.109
N=2500	0.102	0.102	0.108	0.105	0.107

Table 8: Predicted Market Share of Obscure Editions for Different Slope Coefficient Values

	(1)	(2)	(3)	(4)
Slope Coefficient	0.871	1.11	1.7	2.5
Share of editions outside top 250,000	29.33%	7.47%	0.01%	0.00000041%

Table 9: Accuracy of FW predicted shares

	(1) FW coef =1	(2) actual	(3) coef=0.5
Top 10 share (out of 100)	56%	39%	38%
Top 20 share	69%	51%	52%
Top 50 share	87%	76%	76%

Table 10: Accuracy of Predicted Sales from Ranks using top 1,395 ranks, 2009
Linear Specification

Accuracy Measurement	(1) For 1,395 obs	(2) For all observations
A2: Average\median Absolute Residual/Sales	2.7% \ 1.6%	37,765% \ 16,930%
A3: Total Absolute Residual/Total Sales	7.2%	207.6%
Actual Industry Sales	162,761,664	423,946,802
Predicted Industry Sales	160,120,208	1,295,319,552

Table 11: Tradeoffs from Truncating Sample to only Include Larger Sellers, 2009

Limiting Sample to X Top Editions:	110,000	50,000	30,000	20,000	10,000
Share of total sales included	99.02%	94.85%	89.08%	82.97%	70.88%
1. Linear Average A2	45.99%	18.45%	11.38%	8.42%	5.91%
2. Linear A3	3354.55%	177.65%	65.14%	38.05%	19.28%
3. Nonlinear Average A2	6.78%	3.53%	1.47%	0.76%	0.72%
4. Nonlinear A3	21.24%	8.72%	3.65%	2.94%	3.30%