Received: 28 September 2018 | Revised: 9 April 2020 | Accepted: 12 April 2020 | Published on: 10 July 2020

DOI: 10.1002/smj.3199

RESEARCH ARTICLE

STRATEGIC MANAGEMENT JOURNAL WILEY

# Binned scatterplots: A simple tool to make research easier and better

**Evan Starr** | **Brent Goldfarb**

Robert H. Smith School of Business, University of Maryland, College Park, Maryland

**Correspondence**
Evan Starr, Robert H. Smith School of Business, University of Maryland, College Park, MD.
Email: estarr@rhsmith.umd.edu

[Correction added on 26 July 2021, after first online publication: The copyright line was changed.]

**Abstract**

**Research Summary:** We seek to diffuse a graphical tool—binned scatterplots—which we argue can dramatically improve the quality and speed of research in strategic management. In contrast to the current practice of showing plots of predicted values, binned scatterplots graph the nonparametric relationship between two variables, either unconditionally or conditional on a set of controls, for multiple subgroups. This allows researchers to quickly detect the shape of that relationship, examine outliers, and assess which part of the support may be driving a relationship. We propose that the adoption of binned scatterplots will lead to the identification of new and interesting phenomena, raise the credibility of empirical research, and help create richer theories.

**Managerial Summary:** Regression analysis often assumes linear or quadratic relationship forms between two variables. We seek to diffuse a graphical tool—binned scatterplots—that allow the researcher and reader to evaluate whether such assumptions are maintained throughout the data. For example, binned scatterplots may clarify that a regression relationship is nonlinear, or driven by an exceptional firm or a small set of firms. We propose that using binned scatterplots will improve the transparency and quality of empirical work, as well as aid in the development of new phenomena and enhance theory development.

  

**KEYWORDS**

binned scatterplots, data visualization, graphical analysis, least squares regression, nonparametric analysis

## 1 | INTRODUCTION

Recent scholarship in the field of strategic management has drawn attention to the need for improved graphical representation of evidence (Greve, 2018, Levine, 2018; Schwab, 2018) and the importance of reducing scientific apophenia, the likelihood of drawing conclusions that the underlying data do not support (Goldfarb & King, 2016). In this manuscript, we reiterate these challenges by reviewing the most common graphical practices in management and how those practices may be misleading. We then present a practical solution to both challenges: *binned scatterplots* (Cattaneo, Crump, Farrell, Fang et al., 2019, hitherto CCFF). Binned scatterplots provide a graphical representation of the conditional, nonparametric relationship between two variables, and allow the quick detection of nonlinearities, outliers, distributional concerns, and the best fitting functional form. Binned scatterplots can also easily examine heterogeneous relationships for different subgroups. Given these qualities, we seek to diffuse their use in management research to improve the quality and speed of empirical work, to assist in establishing new facts and phenomena, and to generate new explanations via abduction (Heckman & Singer, 2017, King, Goldfarb & Simcoe, 2020), which may then be taken to alternative data for hypothesis testing.

To illustrate the necessity of graphical visualization, British statistician F. J. Anscombe (1973) created four datasets (each with 11 observations) with apparently identical linear relationships, as shown in Table 1. The exercise elucidates a weakness in regression methodology that becomes apparent when one graphs the actual data, as we have done in Figure 1. The actual relationship is roughly linear in the first data set, curvilinear in second, larger in the third because of an outlier, and entirely driven by an outlier in fourth. Anscombe's point is that without visualizing the raw relationship between two variables, regression analyses assuming any functional form may be misleading.

**TABLE 1** Four nearly identical data sets

| | Data set 1 | Data set 2 | Data set 3 | Data set 4 |
|---|---|---|---|---|
| Coefficient on *x* | 0.50 | 0.50 | 0.50 | 0.50 |
| | [0.23, 0.77] | [0.23, 0.77] | [0.23, 0.77] | [0.23, 0.77] |
| Constant | 3.00 | 3.00 | 3.00 | 3.00 |
| | [0.46, 5.54] | [0.46, 5.55] | [0.46, 5.55] | [0.46, 5.54] |
| R-squared | 0.63 | 0.63 | 0.63 | 0.63 |
| N | 11 | 11 | 11 | 11 |

*Note:* Regression results from a simple regression of $y = b_0 + b_1x + e$. Confidence intervals are presented around the point estimates. Data is available from Anscombe (1973).

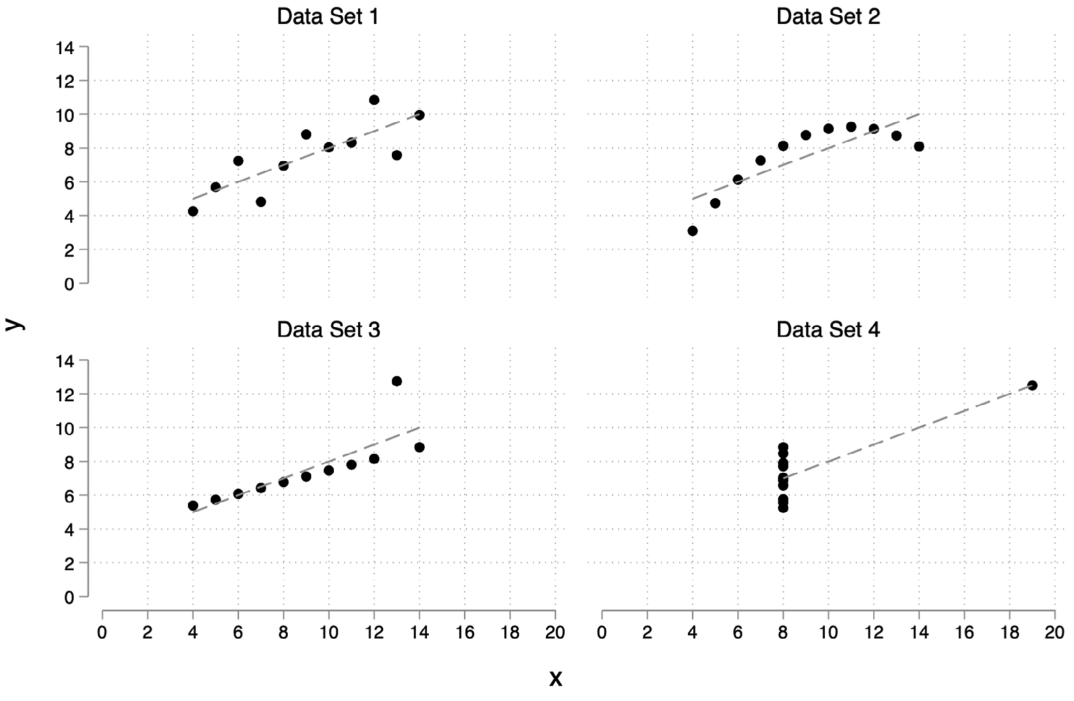


**FIGURE 1** Anscombe (1973) data sets

In strategic management research, the graphical representation of results is uncommon, appearing in only approximately one-third of the 100 most cited empirical papers in leading strategic management journals from 2014–2016 (see Table 2 for details).[1] When results are shown graphically, predicted values of either the main effect of interest (27%) or the interaction effect (30%) are the most likely to be shown.[2] Scatterplots of the unconditional IV-DV relationship are rare (4%), and graphs of the conditional, nonparametric relationship are exceptional—appearing in only one of these 100 most cited papers. Moreover, despite the fact that 94% of the papers involved a continuous independent variable, across those papers an average of only 1.07 functional forms were reported, with 94% reporting only one functional form, and only 18% of those papers explored sensitivity to outliers. Given that even the most popular graphs shown in management journals—plots of fitted values—may poorly reflect the actual relationship of interest (as Figure 1 and Table 1 make plain), it is clear that we, as researchers, reviewers, and editors, need better graphical tools to improve the credibility and viability of empirical research.

Binned scatterplots are one such tool. In the following section, we describe how binned scatterplots are created. We then illustrate the value of binned scatterplots by exploring

[1]We chose 20 papers from the leading journals *Strategic Management Journal, Academy of Management Journal, Administrative Science Quarterly, Management Science* and *Organization Science* by ranking papers by the number of citations on Google Scholar, conditional on having empirical content and being published between 2014 and 2016. On the particular numbers cited in Table 2, we only focus on what appears in the text of the published manuscript. As a result, we only capture what appears in the published studies, and not all the analyses the authors may have examined. It is, of course, possible that authors explore additional functional forms but fail to report this activity in the paper. However, a failure to report such data exploration makes it difficult to interpret published statistics (King, Goldfarb & Simcoe, 2020).

[2]Among the 63% of studies that proposed interaction effects, 48% presented predicted values for different subgroups.

**TABLE 2** Graphical analyses in top Management Journals

| *Panel A. graphs in the manuscripts* | *Mean* | *N* |
|---|---|---|
| Was there a graph at all in the manuscript? | 46% | 100 |
| Was the distribution of a variable graphically shown? | 2% | 100 |
| Was the relationship between *x* and *y* shown at all? | 35% | 100 |
| *Panel B. graphing the relationship of interest* | | |
| Was a scatter plot between *x* and *y* shown? | 4% | 100 |
| Were the fitted values from the estimated model graphed? | 27% | 100 |
| Was the raw data shown, after conditioning out observables? | 1% | 100 |
| Was the relationship between x and y shown in any other way? | 10% | 100 |
| *Panel C. interaction or subgroup analyses* | | |
| Was there an interaction hypothesis? | 63% | 100 |
| Were the interaction results displayed graphically? | 49% | 63 |
| Were the fitted values for the interaction model graphed? | 48% | 63 |
| *Panel D. outliers and functional forms* | | |
| Were any the independent variables of interest continuous? | 94% | 100 |
| How many functional forms were examined for these variables? | 1.07 | 94 |
| Was sensitivity to outliers explored? | 18% | 94 |

*Note:* The sample comprises 20 empirical manuscripts from each of the following journals: *Strategic Management Journal, Organization Science, Management Science, Administrative Science Quarterly*, and *Academy of Management Journal*. Within each journal, the 20 empirical papers were selected by ranking papers based on their citations, conditional on being published between 2014 and 2016 and having empirical content.

relationships in a dataset of computer workstation firms from 1980–1996 used in Sorenson (2000) and other papers, and which Sorenson has made available on the FIVES website.[3] We do not replicate prior work. Rather, we use the data to illustrate the power of binned scatterplots in the context of questions of interest to the strategic management community. Our exposition emphasizes not only how binned scatterplots can best be used to understand the nature of a relationship between two variables in a dataset, but also how their use can help elucidate when the data are too sparse to make strong statements at all. We describe the use of a new routine available in both Stata and R, `binsreg` (Cattaneo et al. 2019), which should be used instead of a popular older Stata command `binscatter` (Stepner, 2014). We summarize the main discrepancy between these approaches in Appendix A, while the technical proofs can be found in CCFF. We conclude by highlighting the limitations and value of binned scatter plots as it relates to both empirical and theoretical research.

## 1.1 | What are binned scatterplots and how do they work?

For small samples, scatterplots may informatively depict the relationship between two variables, $x$ and $y$.[4] But, as samples get large or if variables are categorical, patterns may be difficult

[3]The publicly available data can be found at: http://five.dartmouth.edu/datasets.

[4]Formally, given $N$ realizations of two random variables $x$ and $y$, a scatterplot displays each combination $(x_i, y_i)$ for $i = 1, \ldots, N$ observed in the data.

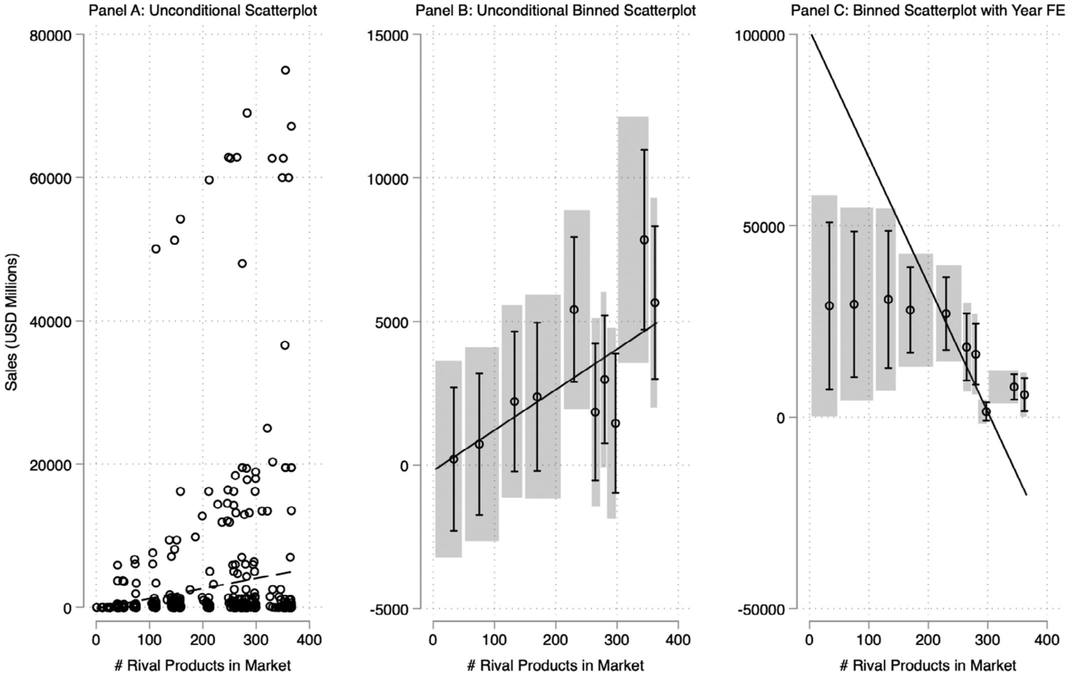


**FIGURE 2** Scatterplots, binned scatterplots, and nonlinear relationships

to observe. Using the sample of computer workstation firms, Figure 2a depicts a traditional scatterplot (and best-fit regression line) of the relationship between the number of rival products in the market (as a measure of competition) and annual firm sales between 1980 and 1996. It is difficult to discern any relationship in the scatterplot.

A binned scatterplot condenses the information from a scatterplot by partitioning the $x$-axis into bins, and calculating the mean of $y$ within each bin.[5] The resulting plot shows the nonparametric relationship between $x$ (in bins) and the mean of $y$.[6] Figure 2b shows the same plot as in Figure 2a, except now the number of rivals is divided into 10 quantiles and average sales is shown in each bin. The best-fit regression line is also shown. In contrast to Panel (a), in Panel (b) a clear, positive relationship between the number of rivals and average firm sales is evident. Note also how in Panel (b) the $y$-axis became compressed relative to Panel (a), since means are less variable than the raw data. Practically speaking, this unconditional relationship is not intuitive: why was more competition associated with more firm-level sales in the early computer workstation industry?

## 1.2 | Binned scatterplots and multivariate models

One plausible explanation for why more rivals was associated with higher sales relates to the stage of industry evolution. As the industry enters a growth phase, both the number of rival

[5]Theoretically, this can also be done for quantiles in examining bivariate relationships without controls. We know of no theoretical result that would allow quantile-based binned scatterplots with control variables.
[6]By nonparametric, we mean that we are not assuming any (parametric) functional relationship between $x$ and $y$.

products and firm sales may rise together. This suggests that the observed relationship between the number of rival products and firm sales may simply reflect a growing industry; a rising tide raises all boats. One way to assess this explanation is to include year fixed effects in the specification, such that the relationship between the number of rivals and firm sales is driven only by within-year variation in the number of rival products and sales. Figure 2c reports the binned scatter plot including year fixed effects. A linear regression would find a negative relationship (shown by the solid, downward sloping line), which might lead a researcher to conclude that competition hurts sales. However, the binned scatter plot reveals a more nuanced, nonlinear relationship: Between 0–250 rivals the relationship is relatively flat, but once there are more than 250 rival products further increases in rival products are negatively correlated with mean firm sales.[7]

Our example raises the question of how binned scatterplots incorporate control variables, such as year fixed effects, so that we can observe the conditional, (semi) nonparametric relationship between $x$ and $y$, as in Figure 2c. Binned scatterplots do this by first partitioning the independent variable of interest into bins. A regression model is then estimated, allowing the relationship between $x$ and $y$ to be different for each bin, while also controlling for any other variables of interest. Using the resulting coefficients on each bin, and setting other covariates to specific values (often the default is to set them at zero[8]), the binned scatterplot shows the expected value of $y$ for each bin.[9]

In practice, only 1 of the 100 manuscripts we reviewed showed the conditional nonparametric relationship between $x$ and $y$. However, as Figure 2c makes evident, binned scatterplots can show conditional relationships just as easily as they can show unconditional relationships. This is exceedingly useful, when it is unclear to what extent other covariates may be responsible for changes in the conditional relationship between $x$ and $y$.

Before we discuss the details of how to implement binned scatterplots, including the number of bins, confidence intervals/bands, clustering, and shape testing, we first provide several examples of their usefulness in practice.

## 2 | APPLICATIONS IN PRACTICE

### 2.1 | Nonlinear relationships

Most directly, binned scatterplots show a nonparametric estimate of the conditional expectation of $y$ given $x$. While OLS can create best-fit lines, including polynomial fits, the basic binned scatterplot makes no assumptions about the relationship between $x$ and $y$.[10] Subsequent tests

[7]Note that in Panel C the regression line does not fit the effects in each bin. While the best-fit line will coincide when the model does not include controls, with control variables the best-fit line will not generally fit the binned scatter plot.
[8]As a result, mean-centering control variables can improve interpretability of the binned scatterplot. Similarly, any fixed effects will be set to the omitted (base) category, which may need to be adjusted for interpretability. `binsreg` estimates point estimates for each fixed effect, and hence can be difficult to use when there are many fixed effects to be estimated.
[9]Note that this approach to incorporating controls in binned scatterplots (i.e., estimating them with the bin-specific coefficients in the same model) is distinct from a prior approach described by Stepner (2014), which relies instead on residualization via the Frisch–Waugh–Lovell theorem. As discussed in Appendix A, and in CCFF, the residualization approach is generally inaccurate when controls are included and they are correlated with the independent variable.
[10]When controls are used, assumptions are made about how those controls are related to the dependent variable (e.g., the assumption could be linear, or quadratic, etc.). Because of these often parametric assumptions, when controls are used the binned scatterplot produces a semi nonparametric graph, where the "semi" comes from the parametric assumptions relating the other controls to the dependent variable.

(discussed later) can confirm the existence of any nonlinearities, but binned scatterplots can help identify whether assuming the relationship is linear, quadratic, log linear, and so on, is appropriate in the first place.[11] For example, in Figure 2c, as discussed above, the mean relationship between rivals and sales is flat when there are a few rivals, but then decreases when rivals are high. The assumption of a linear functional form obscures this pattern.

## 2.2 | Outliers and subgroup analyses

As shown in Table 2, only 2% of the manuscripts we reviewed showed distributions of the variables of interest. Regressions simply fit lines to the data, thus obscuring any information about the distribution of the independent variable. Binned scatterplots make plain whether the distribution of the independent variable is driving a particular relationship—as may be the case when there are outliers. By identifying these outliers, researchers, and reviewers can understand whether a small number of data points are producing a relationship of interest. Given the importance of firm heterogeneity, this makes the technique a natural fit for the strategy field. Here is one such example: The binned scatterplots in Figure 3 examine the relationship between the number of products the firm produces (i.e., a measure of corporate scope) and firm R&D expenditures. Intuition suggests that firms that have more products will be engaged in more R&D. However, Figure 3a reveals an outlier (IBM) that has a very high level of R&D and few products (in this market), and as a result the best-fit line suggests an imprecise negative relationship. When we exclude IBM, however, the slope becomes positive and precisely estimated, as we might expect (Figure 3b). Without seeing the binned scatter plot in Figure 3a, we would not necessarily recognize that the negative relationship is driven by a single outlier.[12]

More than just examining outliers, binned scatterplots can be easily created for different subgroups. To illustrate distributional differences for different subgroups, Figure 3c breaks out the relationship between the number of products and R&D expenditures by ownership type. The binned scatterplot indicates that private firms tended to have between one and eight products and this variance was evenly distributed across firm-years. Public companies were similarly evenly distributed between one and 10, with an outlier that had 20 products. Similarly, subsidiaries had between one and five products, though there was another outlier at 15 products. While a traditional predicted values plot would obscure the distribution of $x$—sometimes leading to predictions about relationships where there is no actual data—this scatterplot informs us that these data only support statements about the relative patterns of product offerings by ownership type for, essentially, 1–5 products.

## 2.3 | Time trends

Many studies employ analyses examining changes over time, often between a control group and a treatment group in a difference-in-differences analysis. Binned scatterplots can be employed

[11] In Appendix B we describe how binned scatterplots can help prevent the author from confusing nonlinear effects for interaction effects.

[12] One important point in these graphs is that when the data are sparsely distributed, a single bin may encompass a single observation. This is the case in Panel B of Figure 3, where every bin above eight reflects only one firm; this is why there are no confidence intervals for these points. We further discuss issues related to the number and placement of bins in Section 3.1.

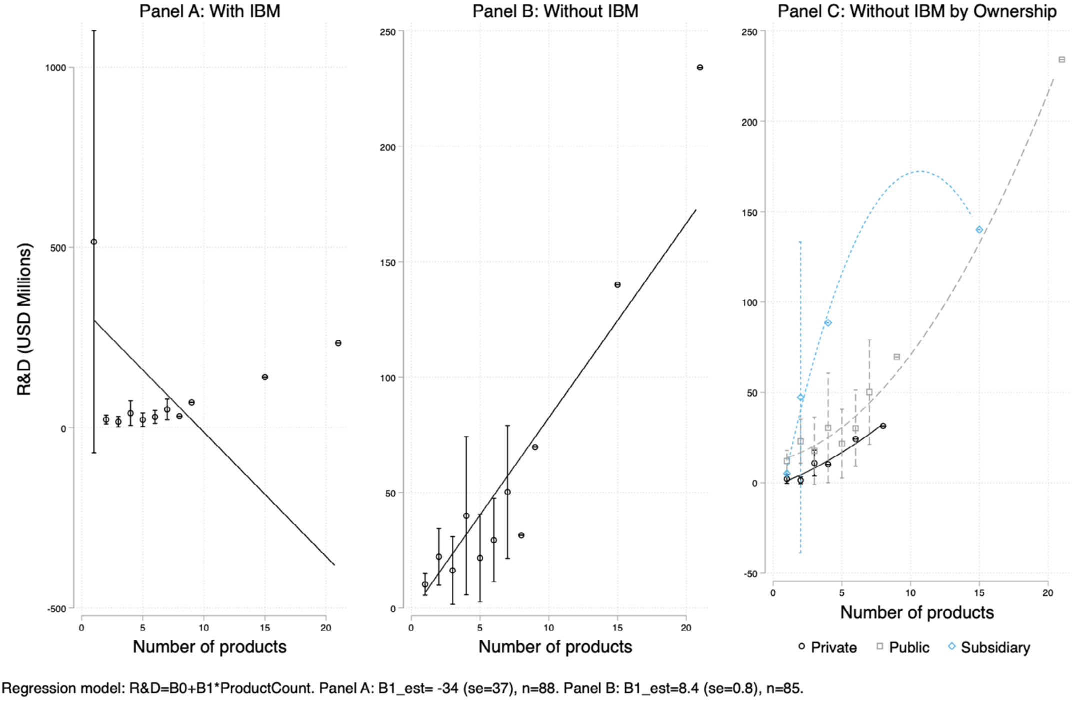


**FIGURE 3** Outliers and subgroups [Color figure can be viewed at wileyonlinelibrary.com]

to show whether the pretreatment trends are parallel between treatment and control. As an example, consider the question of firm diversification over time, in terms of the number of distinct products that they produce. Figure 4a looks at the sample as a whole, tracking the average number of products offered by a firm in each year. There appears to be a break in the trend in 1985, before which the average number of products was rather constant. After 1985, however, the average number of products produced increased dramatically. Figure 4b breaks down these differences by ownership type. Before 1985, each type of firm was on similar, relatively flat trajectories. After 1985, however, public firms were the ones who dramatically expanded their product offerings, whereas private firms and subsidiaries expanded their portfolios at a much slower rate.

## 3 | IMPLEMENTATION

In practice, binned scatterplot routines `(binsreg)` have been programmed into Stata and R.[13] These commands easily handle control variables, create binned scatterplots for different groups, and include options for clustering, weighting, confidence intervals and bands, and parametric

[13]An alternative command `binscatter` was initially written by Stepner (2014) in Stata, and extended to R in 2017 by Timothy Lin (see https://www.timlrx.com/2017/06/11/binscatter-for-r/). However, these routines do not correctly account for control variables. The `binsreg` routine does, and we will focus on its use here. See the Appendix A for more details contrasting these two approaches.

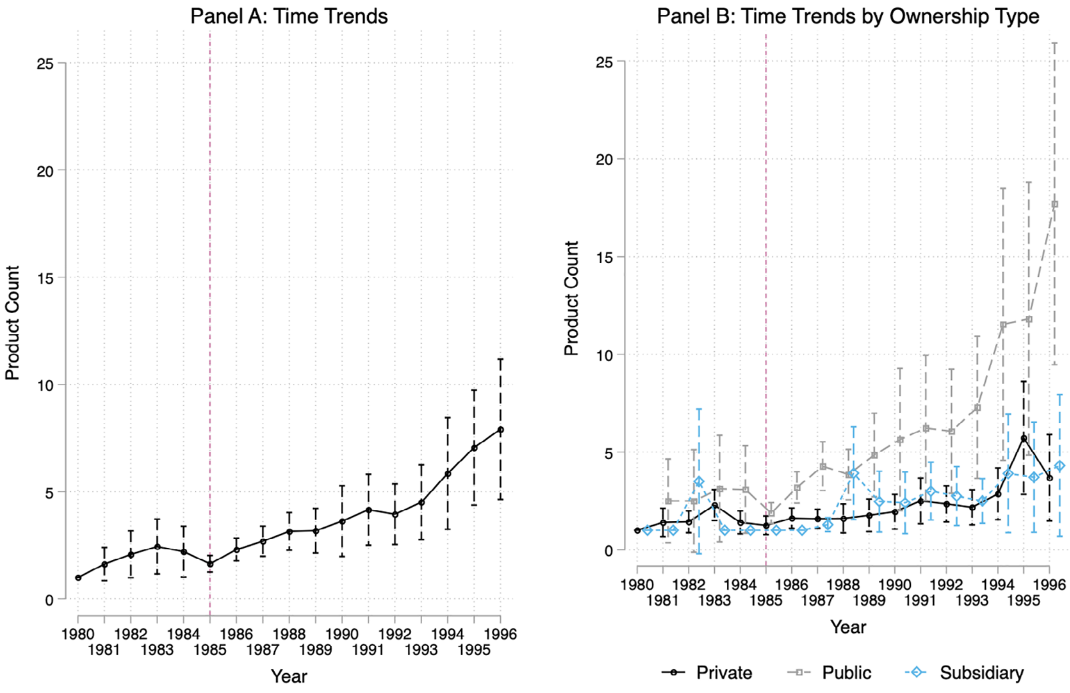


**FIGURE 4** Time trends and product diversification [Color figure can be viewed at wileyonlinelibrary.com]

shape testing.[14] In this section we discuss choices related to the implementation of binned scatterplots.

## 3.1 | Number of bins

One of the most important choices in constructing a binned scatterplot is the number of bins. More bins will allow a researcher to identify more curvilinear patterns, but because each bin has fewer data points there will be more idiosyncratic variance. In contrast, fewer bins include more data points, leading to more precision, but may be less effective in identifying nonlinearities. This is the traditional tradeoff between variance and bias, which is similarly found in selecting the bandwidth in matching (Smith & Todd, 2005) and in regression discontinuity (Cattaneo, Idrobo, & Titiunik, 2017).

CCFF's `binsreg` routine allows several possibilities for choosing the number of bins and for the position of the bins. The default picks the number of bins of the independent variable of interest in a data-driven way—minimizing the integrated mean squared error of the binned scatterplot—and places each bin by dividing the distribution of $x$ into equal quantiles corresponding to the chosen number of bins. This approach trades off the bias and variance in an objective way. At times, we may wish to forego the default choices to enhance the interpretive meaning of a binned scatterplot. For example, when examining trends across years,

[14]CCFF also discusses within and cross-bin smoothing, which we discuss in Appendix C.

even-spacing of the bins across years (as opposed to even distribution of the data across years) may be more intuitive (as in Figure 4).

In the basic binned scatterplot (i.e., without any smoothing), CCFF proves that the number of bins that minimizes the integrated mean squared error is proportional to $n^{1/3}$, where $n$ is the number of observations, such that more observations will generally lead to more bins. However other elements are also important. For example, holding the distribution of $x$ constant, the more curvilinear the true relationship between $x$ and $y$ is, the more bins the algorithm will select (otherwise mean squared error will increase). This implies that even with large $n$, few bins will be chosen for relatively flat relationships. The calculation of the optimal number of bins in a basic binned scatterplot thus takes into account the amount and location of variation in the data available to identify the relationship between $x$ and $y$. Importantly, if the data do not support such a calculation, the researcher can then understand that there is relatively little information in the data to provide strong statements about differences in effects across the support of the data.[15]

## 3.2 | Confidence intervals and bands

Binned scatterplots estimate a different mean of the dependent variable for each bin, but they can do more. First, it is possible to calculate the 95% confidence intervals of the estimated means. This will give the researcher a sense of the precision of the dependent variable of interest within each bin. This is a particularly important capability, as a mean is a summary statistic that hides variation in the variable of interest. An example of this can be seen in Figure 3a, where the first bin has an extremely large confidence interval due to the IBM outlier. However, as previously noted, researchers should be cautious if there are few or even single observations in a bin, which can lead to very tight or singular confidence intervals, as in Figure 3b.

Confidence intervals around a single point may be somewhat misleading, however, when we consider that in a binned scatterplot each point represents the data from an entire *interval* of data. Accordingly, it is possible to calculate a confidence *band* across that whole interval. To illustrate this in the data, we have applied both confidence intervals (at a single point, the mean within the bin) and confidence bands (across the whole bin) to the relationship between rivals and sales in Figure 2b,c. Note, however, that the routines employed by CCFF rely on having enough data in each bin to estimate these bands. Accordingly, when there are few observations, plotting such bands will be infeasible (which is the case in Figure 3). This is a feature not a bug, as it reflects the amount of information in the data.

In addition, creating confidence intervals and bands requires making assumptions about the extent to which error terms are identically and independently distributed. In many cases, the standard errors may need to be clustered, for example, at the level at which "treatment" is determined (Abadie, Athey, Imbens, & Wooldridge, 2017; Cameron & Miller, 2015).

[15]In the extreme case, the `binsreg` routine will not calculate the optimal number of bins, and will also turn off all inferential procedures. See pages 12 and 13 of CCFFb for specific details on the implementation in the case where there is "not enough variation" in $x$. Relatedly, binned scatterplots work best when the $x$ variable is continuous or has a large number of categories, since they show the nonparametric relationship of interest across the full domain of $x$. In the case where $x$ is dichotomous, a binned scatterplot will only have two bins at most. In contrast, when $y$ is dichotomous but $x$ is continuous, binned scatterplots can still be run as normal, where the mean $y$ presented in each bin now reflects a probability, as in a linear probability model.

Accordingly `binsreg` allows confidence intervals and bands to be calculated while clustering at a specified level.

### 3.3 | Testing parametric and shape restrictions

CCFF develops methods to formally test for shape and directionality restrictions.[16] For example, if we had hypothesized that the link between sales and rivals in Figure 2c was nonlinear, a test of a null hypothesis of linearity gives a p-value of 0.15, and a test of a null hypothesis of a quadratic relationship gives a p-value of 0.464. The confidence bands hint at this visually: we could draw multiple shapes within the confidence intervals in the binned scatterplot in Figure 2c.

In terms of directionality, a test that the linear slope across the support of the data is equal to zero gives a p-value of 0.04. However, as the binned scatterplot and confidence bands clarify, there is only evidence of a negative relationship in one region of the data. So should this relationship be interpreted as decreasing or not? This example highlights a cautionary note about hypothesis testing: there may be hidden researcher degrees of freedom associated with simple, linear directional hypotheses. Absent a clear statement about the nature of a negative relationship across the support of $x$, and the associated test, a researcher could choose to present the linear estimate as evidence of support. Binned scatterplots thus help constrain such selective reporting.

This is a general problem with null hypothesis testing: the interpretation of p-values will be distorted by selective reporting as well as by making informed choices from or within a dataset (Gelman & Loken, 2013; Goldfarb & King, 2016). Accordingly, prespecification of hypotheses, analyses, data sampling, specific tests, controls to be included, etc., are required to interpret a p-value correctly (Spanos, 2010; King, Goldfarb, & Simcoe, 2020). What does prespecification imply then for the value of binned scatterplots as it relates to hypothesis testing? Our view is that given the well-established challenges with prespecification (Heckman & Singer, 2017; Olken, 2015), and that binned scatterplots bring to light much more detailed patterns than are typically articulated, or perhaps even imagined, in hypotheses (e.g., Figure 2c), binned scatterplots are best used for exploratory, abductive analyses (i.e., question/phenomenon-based research that does not seek to test hypotheses) and hypothesis development. We expand on this point in our discussion below.

## 4 | DISCUSSION

In this discussion we address practical guidelines for creating and reporting binned scatterplots and discuss how to incorporate them into empirical research. Given the myriad ways to implement a binned scatterplot, we suggest that, as a baseline, practitioners follow this guideline: Create and report a canonical binned scatterplot with confidence bands and with the bins chosen by minimizing the integrated mean squared error (i.e., the default in CCFF).[17] Relative to our discipline's current approach of reporting results—that is, plots of fitted values—this binned scatterplot will depict the nonparametric relationship of interest and allow for the detection of

[16] CCFF implemented these tests in the command `binsreg` and in a separate command `binsregtest`.

[17] In Stata, the precise command would be `binsreg y x z, dots(0,0) cb(0,0) title("Canonical Plot")`, where $y$ is the dependent variable, $x$ is the independent variable of interest, and $z$ is the vector of controls.

outliers as well as rich and impoverished parts of the data, or, if the command fails, whether there is sufficient information in the data to say much at all. The canonical binned scatterplot will also reveal whether a polynomial fit can easily summarize the relationship between the variables across the entire support of the independent variable.[18]

How then might binned scatterplots be used in empirical work? The power of a binned scatterplot to identify patterns in the data makes it a natural fit for fact-driven or other exploratory inquiries, including post hoc analyses, robustness tests, and hypothesis development. As an example of the latter two, if a supported hypothesis suggests a positive relationship, a binned scatterplot may confirm that the relationship is robustly, globally positive. Or, it may show that the positive relationship is driven by an outlier or is only found for a portion of the data. The resulting facts derived from binned scatterplots may thus show that relationships observed in the studied sample are consistent with existing theories, or they may challenge such theories, requiring the search for alternative explanations.[19] It is important to emphasize that such rich discoveries are exploratory in their nature, and a failure to describe them as such would be misleading (King et al. 2020): binned scatterplots are revealing patterns in a given sample, but say little about the population from which the sample is drawn.

Binned scatterplots may also be a powerful complement to epistemic maps (King et al. 2020; Leamer, 1985). Epistemic maps depict the robustness of a regression of $y$ on $x$ across permutations of functional form, measurement, controls, and subsamples. Binned scatterplots could then be used to drill down to understand the nature of the relationship for preferred points on the map. In combination, the tools would provide a powerful mechanism to develop a rich and accurate understanding of which relationships are found in a sample and the consistency of these relationships.

While it may be unconventional to prespecify directional and shape-specific hypotheses, the ability of binned scatterplots to identify rich patterns may also allow their testing when used with larger datasets that allow sample splitting (Han, Kamber & Pei, 2011). In sample splitting, one sample can be mined to identify patterns in the data and to develop explanations for these patterns. These patterns and explanations can then be the prespecification for the second, hold-out sample. Thus, so long as the sampling plan is prespecified, if the patterns from the first sample are replicated in the second, the author may make a stronger claim. Of course, a hold-out sample can only be used only once: if the researcher iterates between the datasets, then this undermines the interpretation of test statistics in the hold-out sample.

To clarify the interpretation of binned scatterplots for the reader, we suggest that authors present their reasoning and findings as they were discovered, and avoid misleadingly presenting discoveries as hypotheses (Bettis, 2012; King, Goldfarb, & Simcoe, 2020). If, for example, a binned scatterplot shows that an outlier is driving a relationship of interest, and the fact that the result was driven by outliers was used to develop a new explanation, then we cannot use the fact that there are outliers as *evidence* for this explanation. Straightforward descriptions of

[18]CCFF encourages smoothing as a way to identify patterns in the data, and this is an integral part of their routine and exposition. However, smoothing should be done carefully. We provide an example of this and potential pitfalls in the Appendix C. Note that the canonical binned scatterplot that we recommend does not incorporate any smoothing.

[19]For example, we might have hypothesized that increased competition would reduce sales, but in Figure 2, Panel C we observed a flat pattern between the number of rivals and firm sales until there are about 250 rival products in the market, but then decreasing thereafter. One possible explanation might be that the rival products are in different submarkets, and it is not until there is a large enough number of products in each submarket that competitive pressures kick in. Evaluating this new conjecture would require further evaluation perhaps with the current dataset, or perhaps adjudication would require additional data.

research processes allow the reader to interpret reported findings correctly and avoid unjustified claims of generalizability.

Finally, binned scatterplots elucidate the form of the relationships between two variables; they are not tools to resolve issues associated with sample selection, endogeneity, missing variables, or similar problems.

## 5 | CONCLUSION

We seek to diffuse the practice of displaying binned scatterplots in empirical work in strategic management. Being able to quickly visualize a relationship of interest, by subgroup if applicable, even conditional on other controls, is a powerful tool. The power of binned scatterplots lies in its *efficiency*. It allows researchers to *simultaneously* evaluate whether a host of alternative, unspecified, and even unimagined patterns are found in the data. This efficiency, coupled with its ease of use, will allow authors, reviewers and editors to improve the predictive validity of offered theories and the credibility of tested theories while enriching theory development and establishing new facts to explain (Graebner, Knott, Lieberman, Mitchell, & Helfat, 2018).

## ACKNOWLEDGEMENT

The authors would like to thank Yuheng Ding for excellent research assistance. [Correction added on 28 August 2020, after first online publication: acknowledgement section has been added.]

## ORCID

*Evan Starr* https://orcid.org/0000-0002-4368-1710
*Brent Goldfarb* https://orcid.org/0000-0002-8828-9018

## REFERENCES

Abadie, A., Athey, S., Imbens, G. W., & Wooldridge, J. (2017). When should you adjust standard errors for clustering. In *NBER working paper No. w24003*. Cambridge, MA: National Bureau of Economic Research.

Anscombe, F. (1973). Graphs in statistical analysis. *American Statistician*, *27*(1), 17–21.

Bettis, R. (2012). The search for asterisks: Compromised statistical tests and flawed theories. *Strategic Management Journal*, *33*(1), 108–113.

Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, *50*(2), 317–372.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2017). A practical introduction to regression discontinuity designs. In *Cambridge elements: Quantitative and computational methods for social science*, Cambridge, MA: Cambridge University Press I. https://www.cambridge.org/core/elements/practical-introduction-to-regression-discontinuity-designs/F04907129D5C1B823E3DB19C31CAB905.

Cattaneo, M. D., Crump, R. K., Farrell, M. H., Feng, Y. (2019). Binscatter Regressions. arXiv:1902.09615.

Gelman, A., and E. Loken. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time. Department of Statistics, Columbia University. https://osf.io/n3axs/download.

Goldfarb, B., & King, A.A. (2016). Scientific Apophenia in Strategic Management Research: Significance Tests & Mistaken Inference. *Strategic Management Journal*, *37*(1), 167–176.

Graebner, M., Knott, A. M., Lieberman, M., Mitchell, W., & Helfat, C. (2018). Question-driven and phenomenon-based empirical strategy research. *Strategic Management* Journal call for papers for a special issue.

https://onlinelibrary.wiley.com/pb-assets/assets/10970266/Question-Driven_and_Phenomen-Based_Empirical_Strategy_Research-1509475745000.pdf.
Greve, H. R. (2018). Show Me the Data! Improving Evidence Presentation for Publication. *Management and Organization Review*, *14*(2), 423–432.
Han, J., Micheline, K., & Jian, P. (2011). *Data Mining: Concepts and Techniques*, 3rd edn, Waltham, Massachusetts: Morgan Kaufmann.
Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques* (3rd ed.). Waltham, Massachusetts: Morgan Kaufmann.
Heckman, J. J., & Singer, B. (2017). Abducting economics. *American Economic Review*, *107*(5), 298–302.
King, A., Goldfarb, B., & Simcoe, T. (2020). Learning from testimony on quantitative research in management. *Academy of Management Review*. in press.
Leamer, E. (1985). Sensitivity Analyses Would Help. *The American Economic Review*, *75*(3), 308–313.
Levine, S. S. (2018). Show Us Your Data: Connect the Dots, Improve Science. *Management and Organization Review*, *14*(2), 433–437.
Olken, B. A. (2015). Promises and perils of pre-analysis plans. *Journal of Economic Perspectives*, *29*(3), 61–80.
Schwab, A. (2018). Investigating and communicating the uncertainty of effects: The power of graphs. *Entrepreneurship Theory and Practice*, *42*(6), 823–834.
Smith, J. A., & Todd, P. E. (2005). Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics*, *125*(1–2), 305–353.
Sorenson, O. (2000). Letting the market work for you: An evolutionary perspective on product strategy. *Strategic Management Journal*, *21*(5), 577–592.
Spanos, A. (2010). Is Frequentist testing vulnerable to the base-rate fallacy?. *Philosophy of Science*, *77*(4), 565–583.
Stepner, M. (2014). Binned scatterplots: Introducing -binscatter- and exploring its applications. In *2014 Stata conference 4, Stata users group*. https://ideas.repec.org/p/boc/scon14/4.html.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Starr E, Goldfarb B. Binned scatterplots: A simple tool to make research easier and better. *Strat Mgmt J*. 2020;41:2261–2274. https://doi.org/10.1002/smj.3199