Visually-Weighted Regression

Solomon M. Hsiang*

Princeton University, NBER

May, 2013

Abstract

Uncertainty in regression can be efficiently and effectively communicated using the visual properties of statistical objects in a regression display. Altering the "visual weight" of lines and shapes to depict the quality of information represented clearly communicates statistical confidence even when readers are unfamiliar with the formal and abstract definitions of statistical uncertainty. Here we present examples where the color-saturation and contrast of regression lines and confidence intervals are parametrized by local measures of an estimate's variance. The results are simple, visually intuitive and graphically compact displays of statistical uncertainty. This approach is generalizable to almost all forms of regression.

*Email: shsiang@princeton.edu. The programs described in this paper are available at the author's website: www.solomonhsiang.com/computing/data-visualization. I thank Marshal Burke, Andrew Gelman, and Reed Walker for helpful suggestions.

The Problem

Applied statisticians exert substantial effort calculating statistical uncertainty when they estimate parameters. Nevertheless, the results of this exercise are often overlooked by readers with limited statistical training or by readers who focus their attention on point-estimates or statistical significance. Nonparametric techniques are particularly vulnerable to misinterpretation since sampling error can introduce large but statistically irrelevant structures in a regression. "Edge effects" that arise when data becomes sparse near the edge of the data's support are especially problematic since these artifacts often distract readers from the central region of the support, where the quality of estimates and inference is actually higher.

We would like a method for intuitively presenting regression uncertainty to all readers. In particular, we want an approach that intuitively communicates to readers which portions of a regression are uninformative because the results are too imprecise, while focusing readers' attention on those regions in a graph where the informational content is highest.

Currently, the most widely used approach is to plot confidence intervals or standard errors using additional curves, shading or error bars. These displays present exact quantitative information that is essential to the proper quantitative interpretation of results, so they are the important in many contexts. However, readers who do not use precise quantitative interpretations may be less careful, relying more heavily on the visual impact of a data graphic or the emotional response it elicits (Cleveland and McGill (1985); Gelman and Unwin (2011)). For example, the color red often makes readers think that something in a graphic is "bad" or "dangerous," and large, dark or visually conspicuous elements of a graphic inform a less careful reader that those elements are "important" (Tufte (1983)).

Realizing this, it seems that the current practice of displaying uncertainty using additional lines and shading has the exact *opposite* effect on readers from what the author of a regression graphic would like. Highly uncertain regions, which are less important, have more linage and more coloration, grabbing the readers attention. In contrast, regions which are more certain, and thus *should* command more of the readers attention, become less conspicuous and sometimes are ignored.

Figures 1-2 present an example. In Figure 1 we generate a sample of one-hundred data points, drawing X from a normal distribution and and generating Y by adding a nonlinear transformation of X to errors that are symmetric, have zero mean and are heteroscedastic. In Figure 2A, we use these data points to try and recover the expected value of Y conditional on X. The upper panel depicts the conditional mean computed using a kernel-weighted moving average (Nadaraya (1964); Watson (1964)) with 95% confidence intervals shaded, while the lower panel depicts the histogram of the underlying data for reference. This standard presentation makes the flaring confidence intervals near the edge of the data large and dark, drawing the readers attention away from the central portion of the regression, which looks like a simple and inconspicuous line. This is unfortunate, since the attention-grabbing confidence intervals near the edges are supposed to communicate "do not pay much attention to this region."

Proposed Solution

We propose that statisticians leverage the concept of "visual weight" to communicate uncertainty in regression results. Visual weight describes the amount of a viewer's attention that a graphical object or display region attracts, based on its visual properties. Artists and designers understand this concept well and use it to their advantage when trying to express ideas or direct a viewer's attention (Arnheim (1954)). Statisticians could do the same. In general, large, interestingly shaped, colored and high-contrast objects in a graphic are the things that attract a viewer's attention. This means that in an image that is mostly white, like Figure 2A and most regression displays, any dark lines or shading that contrast with the white background attract the reader's attention. In regressions with poorly behaved edge effects, the size, darkness, visual contrast and the curved shapes of flaring standard errors distract the viewer from the center of the display, where most of the information is contained.

If the goal of displaying regression results is to share information with a reader, then the reader's time will be used most efficiently if the author of a display directs the readers attention towards the portions of the graph containing the most information. In regression, informational content is driven by the level of certainty, and in graphics, attention is directed by visual weight. So our simple solution is to equate visual weight with statistical certainty when designing regression displays.

Under the standard approach to presenting regression results, graphical objects ("ink" in the language of Tufte (1983)) are added to portions of a graph to convey uncertainty: error bars, confidence limits and shading. This *addition* of visually interesting elements to convey uncertainty is what skews the visual weight of a display towards its regions with the lowest informational content. Instead of *adding* graphical ink or contrast, and thus visual weight, to uncertain portions of a graph, we propose that graphical ink and contrast are *removed* when results become more uncertain. Doing so will cause a reader's attention to increase in regions of the graph that contain more information, and to decrease in regions with less information.

To implement this idea, we must formalize the artist's classical notion of visual weight so that we can draw visually-weighted images using a computer. To do this, we propose defining a "visual-weighting function" $c(.)$ that is used to determine the visual weight of statistical objects, such as a line-segment or confidence interval. Once c is chosen to reflect the quantity of information underlying the statistical object or the level of certainty associated with that object, it can also be used to parametrize the visual properties of the object in a graphic. For example, if color saturation is set equal to c, then more informative regions of a graphic will be darker, thereby attracting the viewers attention to those regions.

For sets of paired observations (X, Y), we use regression to recover the conditional expectation function

$$\hat{f}(X) = \mathrm{E}[Y|X] \tag{1}$$

which has several measures of certainty at every point. For example, one could use a standard error, a confidence interval, the level of statistical significance, the local sample size or the inter-quartile range of a posterior distribution to summarize the level of certainty in the estimate $\hat{f}(X)$. Selecting any one of these (or other) summary metrics for certainty, we set c equal to this measure over the entire

support

$$c(X) = certainty(\hat{f}(X)). \tag{2}$$

Using the data to estimate $\hat{c}(X)$, we visually weight the regression results by using $\hat{c}(X)$ to determine the color-saturation or other parameterizable visual feature of the regression results that increases the visual weight of the plotted object. A "visually-weighted regression" is simply a plot of the vector-valued function $\{\hat{f}(X), \hat{c}(X)\}$ over the values of X, where $\hat{c}(X)$ is depicted using the visual weight of the graphical objects used to display $\hat{f}(X)$.

The concept and utility of visually-weighting regression results is best communicated with examples. Below, we provide two examples that are immediately accessible to readers[1]: visually-weighting a kernel-based mean regression line and simultaneous visual-weighting of both a regression line and its confidence interval. In addition to these examples, other implementations should be possible for all types of regression.

Example 1: Visual-weighting of a regression line using color-saturation

We present an example of visually-weighted regression in Figure 2B. The graph summarizes the data from Figure 1 and displays the same regression results as Figure 2A, however visually distracting confidence intervals have been removed. Instead, to convey uncertainty the regression line has been dimmed where estimates are expected to be less certain, so that it contrasts less with the white background and claims less of the reader's attention. The high contrast in the center of the regression line pulls the reader's attention towards the center of the graphic, where most of the information is displayed. As a reader tries to examine the edges of the regression, they struggle to make out the shape of the line, feeling a bit uncertain the same way one feels when we try to make out distant shapes in a fog. This emotional *feeling* of uncertainty is familiar to everyone, regardless of our statistical training, so we intentionally make readers *feel* uncertain when looking at the uncertain portions of the regression in order to communicate formal uncertainty to readers who have no formal training.

We encourage readers to examine both panels of Figure 2 with their eyes and their mind relaxed. As one looks at Panel A, our eyes are draw towards the edges of the image, where the flaring and twisting behavior of the confidence intervals is interesting to look at. As one shifts to looking at Panel B, our eyes are drawn inward toward the center of the image, where the sharp contrast between the regression line and the background is attractive to look at. By visually-weighting a regression display, we take advantage of the natural algorithms that our brain uses to search for visual information. If we "reward" our brain with the feeling that it has discovered more visual information when it is viewing more statistical information, we create a more intuitive data graphic.

In Figure 2B, $\hat{f}(X)$ is drawn as a line whose color-saturation (darkness) is parametrized to match the values $\hat{c}(X) = \sqrt{N(X)}$ where $N(X)$ is the number of observations (after weighting) used to compute the mean at X, since this is inversely proportional to the expected variance in the mean[2]. For display

[1] Plotting functions in Matlab and Stata have been written by the author and are available for free online. They can be downloaded at www.solomonhsiang.com/computing/data-visualization.

[2] Because $\hat{c}(X)$ must be displayed over line segments, rather than at exact points, each segment's color is scaled to

purposes or different data, alternative definitions of $\hat{c}(X)$ may be preferred. In this example, visual-weighting is achieved by altering the color-saturation of the line, however it could also be done by increasing the thickness of the line, changing its color, altering its pattern, or changing the properties (eg. size) of markers between line segments.

Graphical compactness The primary goal of visually-weighting a regression display is to align readers level of attention and feelings of certainty with actual statistical confidence, however this approach has the additional benefit of being graphically compact: it allows for observational density and confidence to be displayed without needing to introduce additional graphical elements that clutter graphics and confuse readers. In Figure 2A, confidence intervals are added to convey certainty and the histogram in the lower panel is used to depict the observational density. In this figure, these elements are clear, but if multiple regressions were to be shown on a single set of axes, these components would quickly become overwhelming. In contrast, visually-weighted regressions can easily be overlaid with one another. Figure 3 makes this point clear. Panel A displays four sets of observations (with the same error structure as that in Figure 1) and the functions that underlie their data generating processes. The data are closely packed together, so displaying four sets of confidence intervals, error bars, or histograms, in addition to four regression lines, would be extremely confusing. Instead of doing this, Panel B displays four visually-weighted regressions for this data. The graphic is visually clear, allowing the reader to focus on the similarities (in the middle) and the differences (at the edges) across these functions while also providing the reader with a sense of which structures may be less reliable. For example, the data underlying the left end of the blue curve and the right end of the black curve deviate substantially from the true means; fortunately, the lines are lightly colored to convey the potential unreliability of the resulting portions of the regression.

Example 2: Visual-weighting of a regression line and confidence intervals using "watercolor regression"

The example above conveys an intuitive sense of uncertainty to the casual viewer, however it does not provide trained viewers with the formal measures of uncertainty commonly used in evaluating regression analysis. By visual-weighting both regression lines and confidence intervals, we can achieve both of these objectives in a single graphic. To achieve this, we propose using the color-saturation of confidence intervals to depict statistical certainty – with an overlaid regression line that is reverse-colored so that it too is visually-weighted via its contrast with the confidence interval. Figure 4A depicts a simple implementation of this approach using a bootstrapped 95% confidence interval for synthetic data. The color saturation of the confidence interval is set to be inversely proportional to the width of the bootstrapped confidence interval at each value of X:

$$\hat{c}(X) = \frac{1}{\hat{Q}_{0.975}(X) - \hat{Q}_{0.025}(X)} \tag{3}$$

$\hat{c}(X)_{segment} = (\sqrt{N_{x_1}} + \sqrt{N_{x_2}})/2$ where N_{x_1} and N_{x_2} are the number of observations used to compute $\hat{f}(X)$ at the two endpoints of each line segment.

where $\hat{Q}_\alpha(X)$ is the α-quantile of resampled estimates for $\hat{f}(X)$ at each value of X. When the estimate is more precise, the confidence interval is darker and its contrast with the white regression line is stronger, causing both features to have additional visual weight that draws the viewer's eye towards those portions of the display. The inverse-proportionality in Equation 3 is chosen because it conveniently sets the total amount of ink-coloration to be constant for any vertical line-integral through the plotted confidence intervals[3]. One can imagine that a uniform amount of ink is "squirted" onto the page along the regression line and then "spread out" vertically (in the Y-dimension) to fill in the confidence interval. Thus, even if a causal viewer does not know that confidence intervals are defined such that there is a constant probability (e.g. 95%) that the true expectation of X falls within the interval, the intuition of this idea is communicated by "ink-preserving" confidence intervals that fix the quantity of ink within the interval for each value of X.

With more effort, we can construct a regression display that (i) has a visually-weighted confidence interval and regression line, (ii) preserves ink and (iii) uses ink to depict the probability that the true value for $\mathrm{E}[Y|X]$ passes through a given location (X, Y). We prefer this approach to the simple approach in Equation 3 because it is generally not the case that a regression line is located within a confidence interval with uniform probability, however specifying that color saturation is uniform in Y (Figure 4A) could be misinterpreted as suggesting this idea. To satisfy (iii), we specify the visual-weighting function over both X and Y:

$$\hat{c}(X, Y) = \hat{\Pr}(Y = \mathrm{E}[Y|X]) \tag{4}$$

subject to the ink-preserving constraint. Equation 4 can be implemented multiple ways, for example using asymptotic distributions or Bayesian estimation. Here, we demonstrate an implementation that uses the bootstrap to estimate Equation 4. We first resample the data M times, estimating $\mathrm{E}[Y|X]$ each time. Figure 4B displays the "spaghetti plot" of these M estimates. This spaghetti plot approximates the distribution of $\mathrm{E}[Y|X]$ and is somewhat visually-weighted because dense clusters of estimates appear darker, however the less-dense clusters also have a high visual weight because they contrast with the light background – an undesirable property because there is low certainty in these regions. To overcome this drawback, we apply a smoother (in Y) to the spaghetti by defining a mesh over the estimates (Figure 4C) and applying a kernel-weighted moving average to each column of the mesh[4]. The resulting graphic is shown in Figure 4D. In regions where there were many spaghetti lines close together, the confidence interval is dark and the contrast with the regression line is sharp, giving both the confidence interval and regression line heavy visual-weight. Where the spaghetti lines were diffuse, the confidence interval is washed out and contrast with the regression line is low, clearly indicating that results are uncertain in these regions. Because $\hat{c}(X, Y)$ also becomes small near the upper and lower bounds of the confidence interval, this approach indicates to untrained viewers the low likelihood that a conditional expectation is near these limits of the confidence interval. Since its appearance

[3]This "ink-preserving constraint" is $\int_{-\infty}^{\infty} \hat{c}(X, Y) dY = \textit{constant}$ for all values of X.

[4]To smooth the data in this example, we use the same Nadaraya-Watson estimator that is used to estimate the non-parametric regression line, although other estimators could also be used. We note that we found it computationally convenient to use a mesh that scaled with the extent of the data (in Y) so that a constant bandwidth (in mesh units) could be applied to all columns of the mesh – a procedure that retains details in the distribution even when the confidence interval becomes narrow.

resembles a watercolor painting, we term the above approach a "watercolor regression."

Generality

The general approach of using visual weight to direct reader's attention and convey statistical confidence is generally applicable to almost all forms of regression. It can be applied to all types of graphics, from displaying the results of ordinary least-squares estimation to maximum-likelihood estimates of quantiles to the posterior distributions of Bayesian estimates. Here, we demonstrated the approach for specific measures of uncertainty applied to nonparametric mean regression, altering the color-saturation of a line (or confidence interval) and the contrast between the line and its background to influence its visual weight. However, there are numerous metrics for uncertainty that may be recombined with various estimation procedures and depicted using different alterations to regression lines, many of which may provide similar or better results.

References

Arnheim, R. (1954). *Art and Visual Perception: A Psychology of the Creative Eye*. University of California Press.

Cleveland, W. S. and McGill, R. (1985). Graphical perception and graphical methods for analyzing scientific data. *Science*, 229(4716):828–833.

Gelman, A. and Unwin, A. (2011). Visualization, graphics, and statistics. *Statistical Computing and Graphics*, 22.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and Its Applications*, 9:141–142.

Tufte, E. R. (1983). *The visual display of quantitative information*. Graphics Press, Chesire, Connecticut.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya*, 26:359–372.

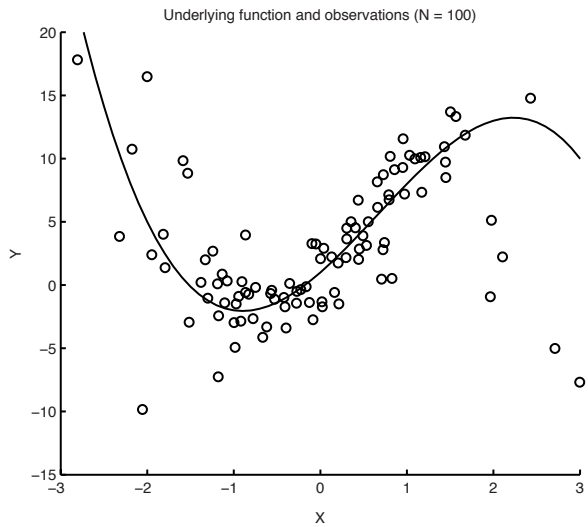

Figure 1: The function $Y = 1 + 6X + 2X^2 - X^3 + \epsilon$ and 100 observations where $X \sim N(0, 1.2)$ and $\epsilon \sim N(0, 1.5(1 + X^2))$.

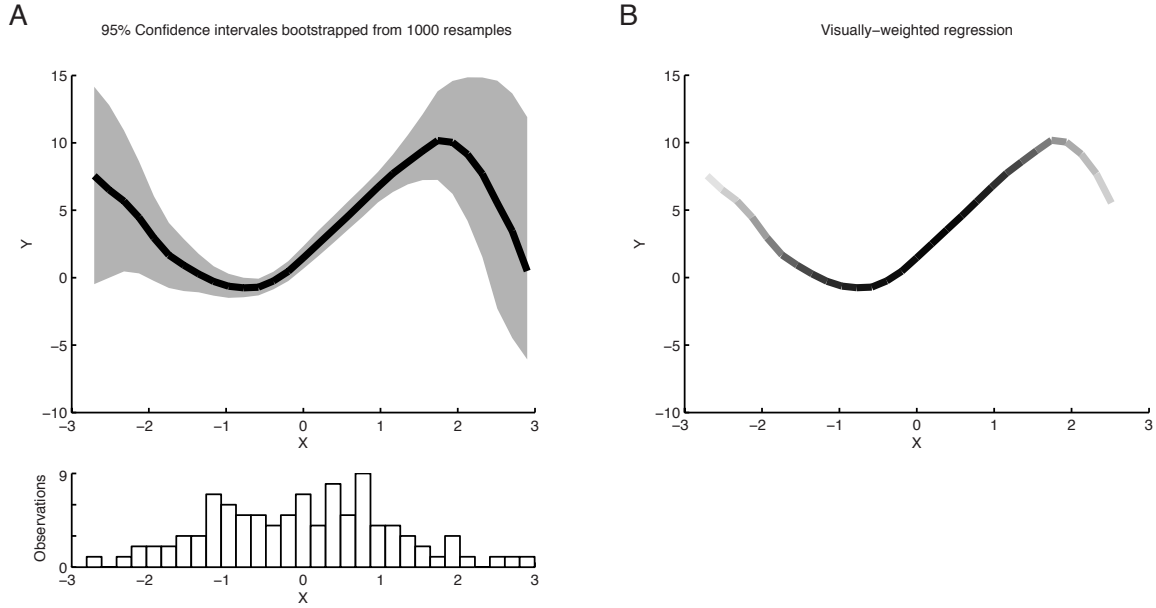

Figure 2: (A) A kernel-weighted moving average regression of the data in Figure 1 using an Epanechnikov kernel with bandwidth = 1. 95% confidence intervals are computed from 1000 resamplings of the data with replacement. Histogram displays the density of underlying observations. (B) Visually-weighed version of the regression used in Panel A. The darkness of the line is scaled by the quantity of information underlying each portion of the curve. In this example, the darkness is inversely proportional to average expected variance of each segment's endpoints $\sqrt{(N(X)}$, where $N(X)$ is the effective number of observations (after kernel-weighting) used to compute the mean at X.

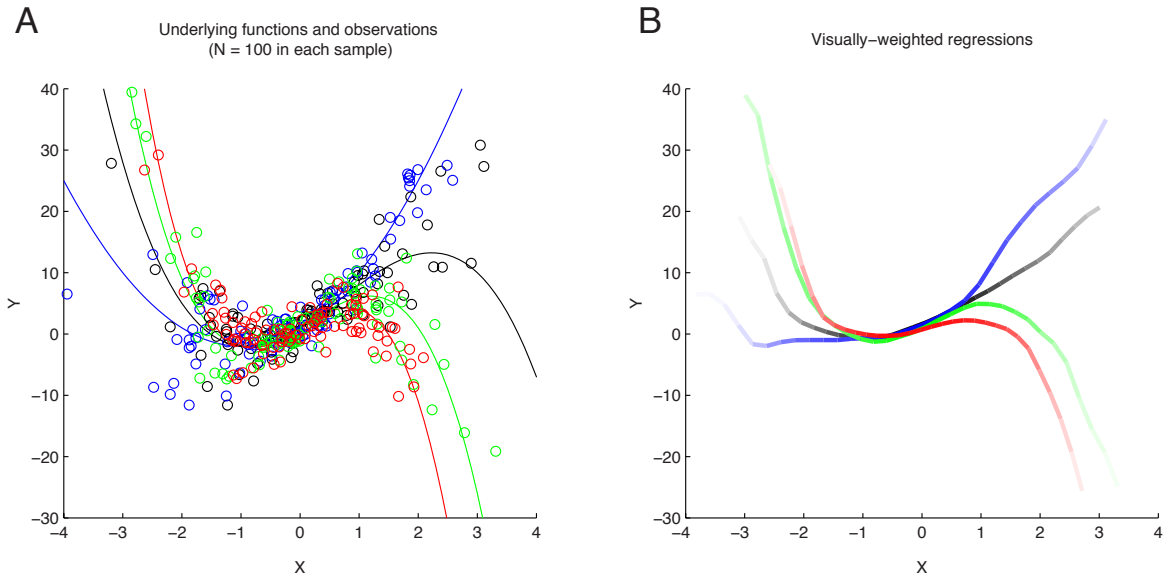

Figure 3: (A) Four data sets with different data generating processes that are similar to that in Figure 1. Regressions of this data that display confidence intervals or histograms rapidly become visually complex or consume multiple display panels. (B) An overlay of visually-weighed regressions is visually clear and informative about the reliability of the various estimated structures.

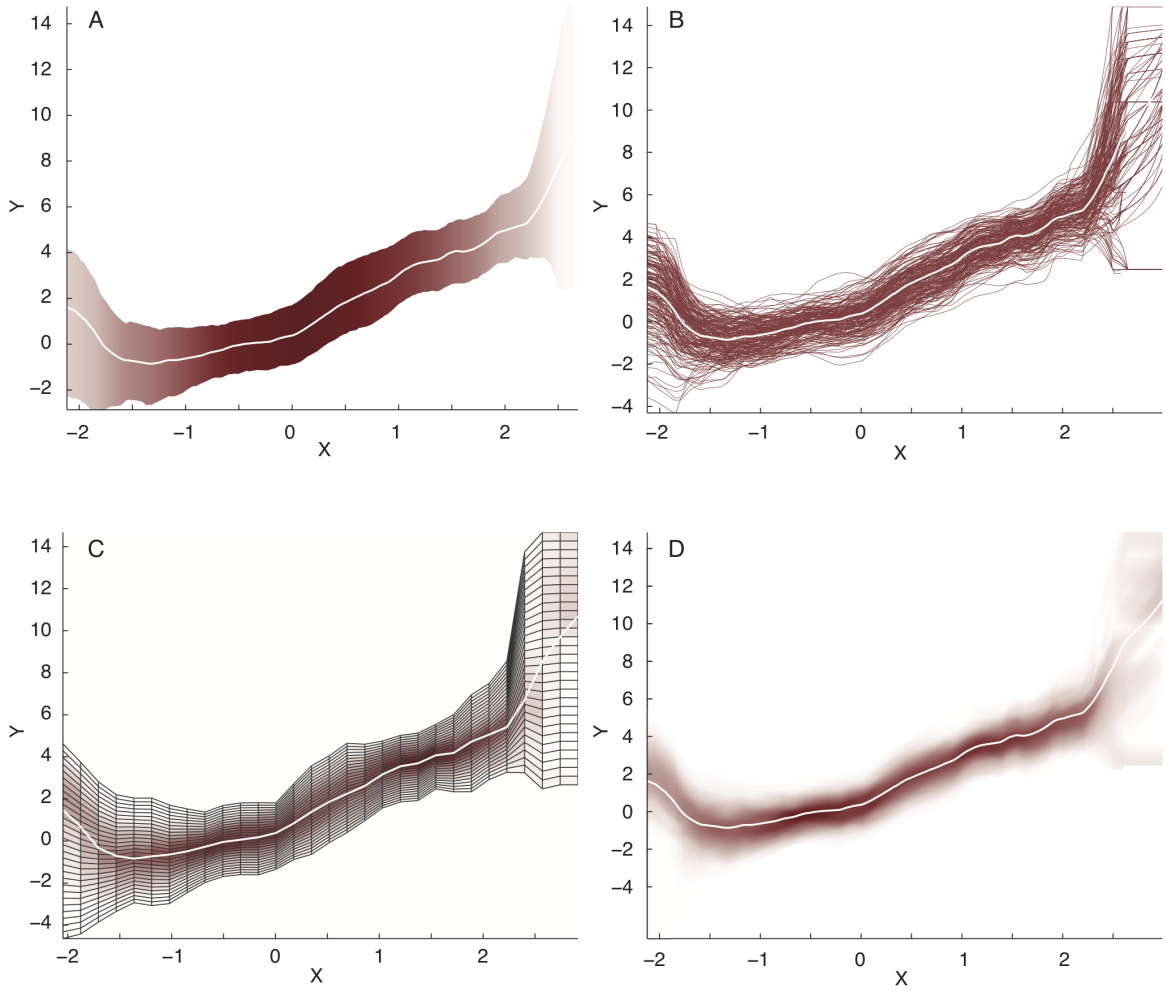

Figure 4: (A) Simultaneous visual-weighting of a confidence interval and regression line using the visual-weighting function in Equation 3. The coloration of the confidence interval is "ink-preserving" because the quantity of ink used to color the graph, integrated along a vertical line, is constant across all values of X. The regression line is visually-weighted because its contrast with the confidence interval increases in regions where the confidence interval is narrower. (B) "Spaghetti" plot of regression results obtained by resampling the data. (C) Applying a mesh over the resampled estimates in B allows us to smooth over the spaghetti in the Y-dimension. (D) Smoothed results for continuous visual-weighting of the confidence interval in both the X- and Y-dimension, as described in Equation 4. This "watercolor regression" uses visual-weighting to depict the certainty of the regression line (via contrast with the confidence interval) as well as certainty within the confidence interval itself (via color saturation).