Visually-Weighted Regression

Solomon M. Hsiang[*]

Princeton University

June, 2012

Abstract

Uncertainty in regression can be efficiently and effectively communicated using the visual properties of regression lines. Altering the "visual weight" of lines to depict the quality of information represented clearly communicates statistical confidence even when readers are unfamiliar or reckless with the formal and abstract definitions of statical uncertainty. Here, we present an example by decreasing the color-saturation of nonparametric regression lines when the variance of estimates increases. The result is a simple, visually intuitive and graphically compact display of statistical uncertainty. This approach is generalizable to almost all forms of regression.

[*] Email: shsiang@princeton.edu. The program and simulation described in this paper are available at the author's website: www.solomonhsiang.com/computing/data-visualization. I thank Marshal Burke and Reed Walker for helpful comments.

The Problem

Applied statisticians exert substantial effort calculating statistical uncertainty when they estimate parameters, however 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 statistical 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

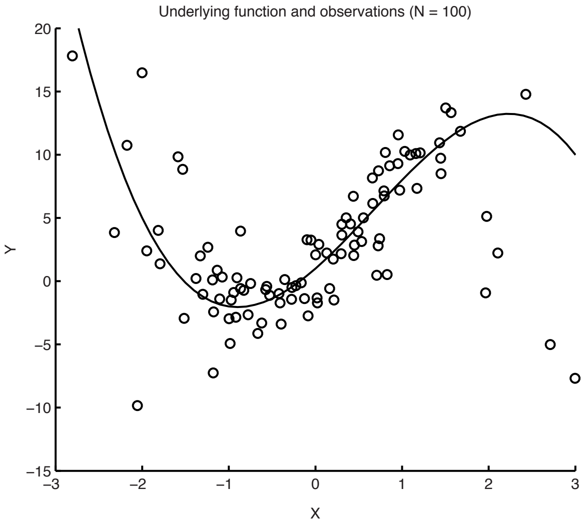

Figure 1: The function $Y = 1 + 6X + 2X^2 - X^3 + \epsilon$ and 100 observations where $X \sim N(0, 1.2)$ and $\epsilon \sim N(0, 1.5(1 + X^2))$.

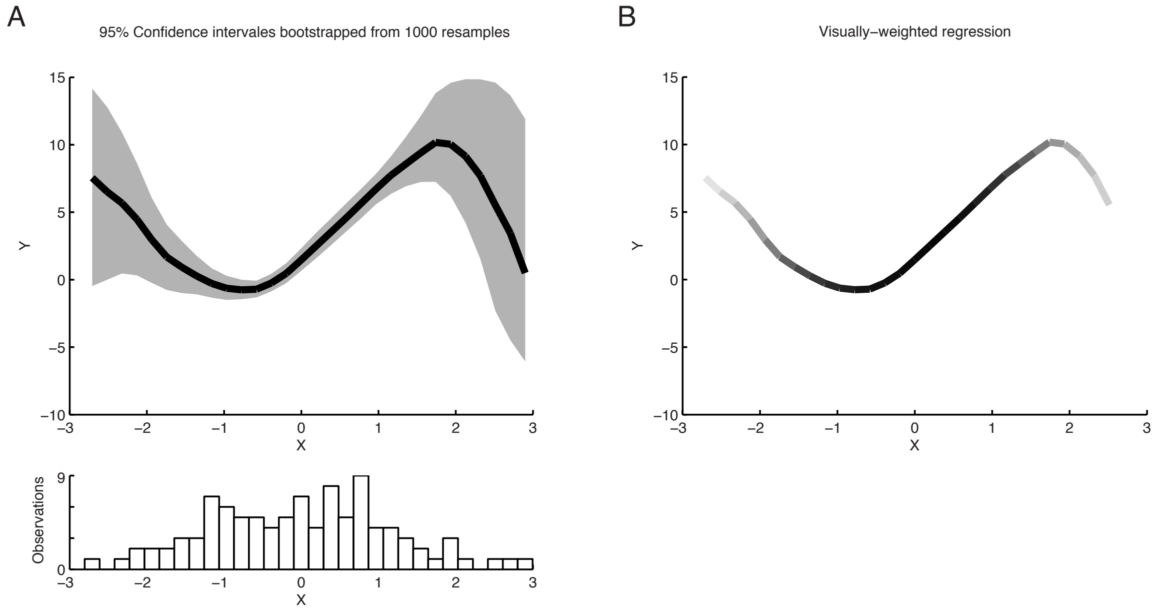

Figure 2: (A) A kernel-weighted moving average regression of the data in Figure 1 using an Epanechnikov kernel with bandwidth = 1. 95% confidence intervals are computed from 1000 resamplings of the data with replacement. Histogram displays the density of underlying observations. (B) Visually-weighed 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.

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, and thus visual weight, to uncertain portions of a graph, we propose that graphical ink is *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.

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.

From a computational perspective, visually-weighting regression is trivial; although, to our knowledge, it is not currently an option in any statistical package. 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

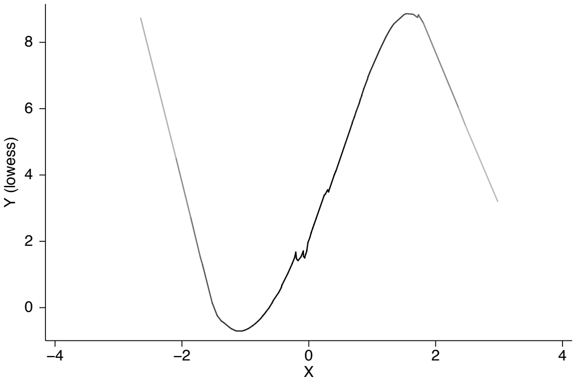

Figure 3: Visually-weighted lowess in Stata.

one of these summary metrics for certainty, we can describe a function of this measure over the entire support

$$\hat{c}(X) = certainty(f(X), X) \qquad (2)$$

which we use to visually weight the regression results. A visually-weighted regression is simply a plot of the vector-valued function $\{\hat{f}(X), \hat{c}(X)\}$ over the values of X. We propose that in a two-dimensional graphic, $\hat{f}(X)$ is drawn as a line whose visual-weight is parametrized to match the values in $\hat{c}(X)$. For the example presented in Figure 2B, we select $\hat{c}(X) = \sqrt{N(X)}$ 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[1]. For display purposes or different data, alternative definitions of $\hat{c}(X)$ may be preferred and another example is given below (Figure 5D). 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.

To make application of this approach immediately accessible, a Matlab function for visually-weighting kernel-based mean regression (`vwregress.m` shown in Figure 2B) and a Stata function for visually-weighting local linear regression (`vwlowess.ado` shown in Figure 3) have been written by the author and are available for free online[2]. Similar implementations should be possible for various types of regression by composing new program commands and command options, or by using clever combinations of existing programs.

Graphical compactness and generality

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 4 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

[1] Because $\hat{c}(X)$ must be displayed over line segments, rather than at exact points, each segment's color is scaled to $\hat{c}(X)_{segment} = (\sqrt{N_{x_1}} + \sqrt{N_{x_2}})/2$ 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.

[2] Plotting functions can be downloaded at `www.solomonhsiang.com/computing/data-visualization`.

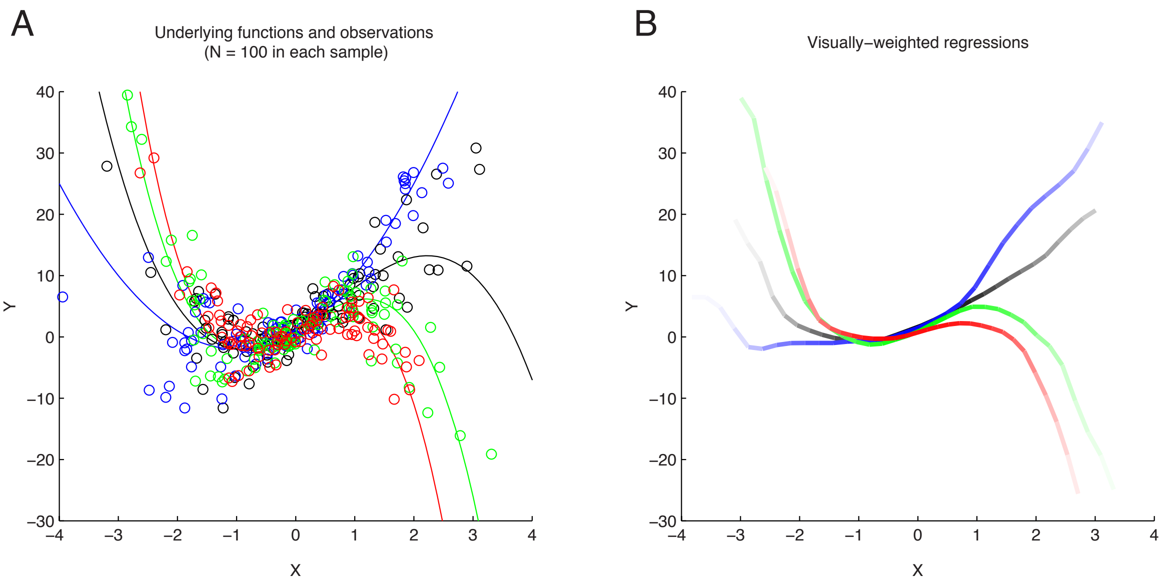

Figure 4: (A) Four data sets with different data generating processes that are similar to that in Figure 1. Regressions of this data that display confidence intervals or histograms rapidly become visually complex or consume multiple display panels. (B) An overlay of visually-weighed regressions is visually clear and informative about the reliability of the various estimated structures.

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.

One drawback of using only a visually-weighted regression line is that hypothesis testing cannot be done visually. However, two variants on this approach can be used to allow readers to test hypotheses. First, if lines denoting confidence limits or shaded regions denoting confidence intervals are introduced, they can also be visually-weighted. Adding in confidence bounds introduces some additional visual weight to regions of the graphic where certainty is low, however visually dimming these structures partly counteracts this effect[3], as demonstrated in Figure 5. A second approach is to *use visual weighting to denote the outcome of hypothesis tests*. For example, if an author is interested in testing the null hypothesis that the conditional mean is zero, they could set $\hat{c}(X) = 1 - \Pr(f(X) = 0)$. Alternatively, the author could vary the color of a line, eg. making the line less red and more blue when a hypothesis test is more statistically significant, while still using the color-saturation of the line to denote the certainty of the regression values themselves.

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

[3] The function `vwregress.m`, provided online, contains an option for including visually weighted confidence limits and an option for visually weighting a regression by the size of the confidence intervals rather than $\sqrt{N(X)}$.

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 one measure of uncertainty applied to nonparametric mean regression, altering one property of a line 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.

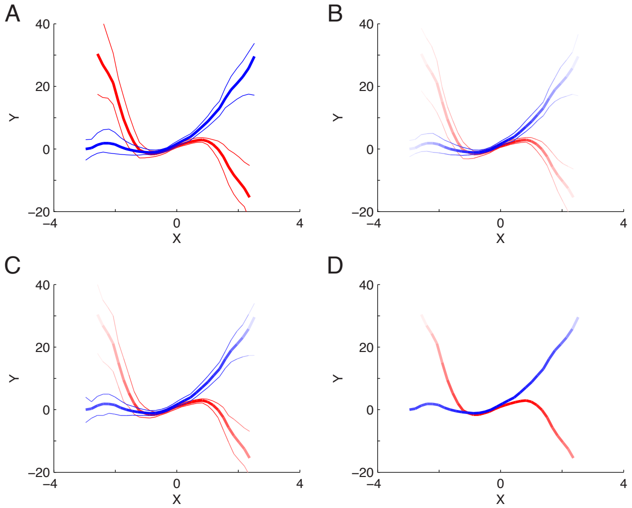

Figure 5: (A) Nonlinear regressions with bootstrapped confidence limits (95%) and no visual weighting. (B) Same as Panel A, but the visual weight for both the mean and the confidence limits are weighted according to $\hat{c}(X) = \sqrt{N(X)}$. (C) Same as Panel A, but the visual weight for both the mean and the confidence limits are weighted according to $\hat{c}(X) = -CI_{95}(X)$, where $CI_{95}(X)$ is the width of the 95% confidence interval at X. (D) Same as Panel C, but with no confidence limits plotted.