Regression Quantile Data View

Gib Bassett

March 17, 2023

1 LineLand

In LineLand[1] all data sets are univariate, $y_1, ..., y_n$.

In the olden days data sets were "viewed" by their mean and standard deviation (μ, σ). Situations in which this was a good-thing were derived. The good-thing situations came to be called normal. With the normal name the good-thing conditions were forgotten. People assumed that (μ, σ) was reasonable no matter what. They knew that the two numbers μ and σ could only give a blurry idea of n-sized data, especially with big n. But at least μ gave a sense of location, and σ a sense of dispersion. The mind's eye view of data was normal (μ, σ).

The (μ, σ) representation turned out to be a big mistake. The problem was not the blurriness produced by a few statistics to represent lots of data. The problem was that the same (μ, σ) could represent wildly different data. The mapping

$$(\mu, \sigma) \to Data$$

is one-to-"way-too-many". This was beautifully demonstrated by *Anscombe(*1973).

*Anscombe presented four very different data sets each with the same $(\mu, \sigma^2) = (0, 10)$; see Figure 1. DataSet1 was the normal one imagined by $(\mu, \sigma^2) = (0, 10)$. DataSet2 is very different: all but one of the observations is near 3.1 and one is far to the left at -10. But its summary is the same $(0, 10)$. The third data set features two clusters of data. The fourth data set has equally distributed data in $[-5, +5]$. All of these data differences result in the same summary statistics. The differences are invisible to (μ, σ).

[1] This report was motivated by discussion of the Anscombe data at the Andrew Gelman blog. Apologies to Edwin A. Abbot(1884).

[But it is not all bad news. For a given (μ, σ) (and n) the observations in a data set cannot be totally arbitrary, anywhere. There is a bound on how deviant you can be. The outlying observation shown in DataSet2 cannot be more than $3.17 = \sqrt{10} = \sqrt{n-1}$ standard deviations away from the mean. The comedian PAS joked that the (μ, σ) summary was selected *just so* all of the 101 residents in Woebegone would be within 10 standard deviations of the mean; *PASamuelson(*1968)].

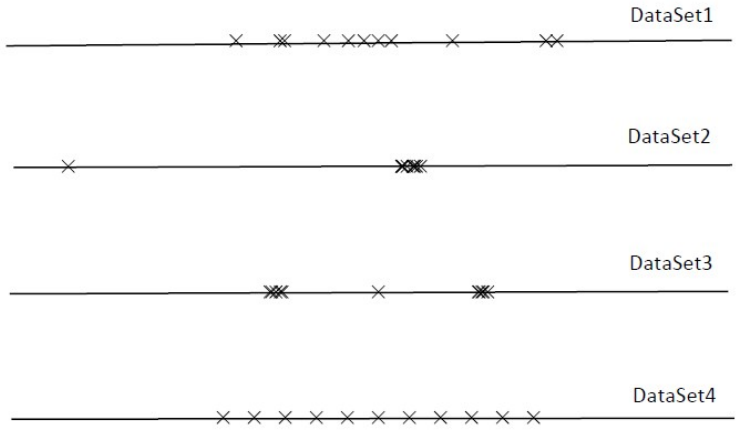

Figure 1: LineLand Data: $(\mu, \sigma^2) = (0, 10)$

Having presented the quartet, *Anscombe recommended caution in "the common use of library programs and various statistical packages...Most [of which] originated in the pre-visual era." Graphing the data would help, but it would not be easy in LineLand and it would require, as *Anscombe put it, a "cunning and fighting spirit".

1.1 The Quantile Data View

A cunning and fighting spirit was not enough for one-dimensional LineLand analysts to envision the two-dimensional Figure 1. Geometers reported that the graphs were in parallel one dimensional universes. Instead of graphs LineLand began to use the usual summary statistics along with a few selected quantiles. The extreme quantiles provided a sense of the bounds on the data, and the selected interior quantiles identified regions where the proportion of data was less than/greater than the quantile. This enhanced quantile summary accurately conveyed differences in the *Anscombe quartet. Analysts could better envision the data.

2 FlatLand

In (regression) FlatLand all the data sets are bivariate $(y_i, x_i), i = 1, ..., n$, and y_i depends on x_i including an intercept. The simplest and most common dependence is linear, which we focus on.

In the olden days data sets were represented by least squares regression output. Situations in which this was a good-thing were derived. That the use of least squares summaries depended on the good-thing situations was gradually forgotten as the good-thing situations came to be called normal. People knew that a few statistics could only provide a blurry idea of the data. But still: it was thought the summaries gave a rough idea of data location and dispersion. The mind's eye view was normal data with location, dispersion, and fit given by the least squares output.

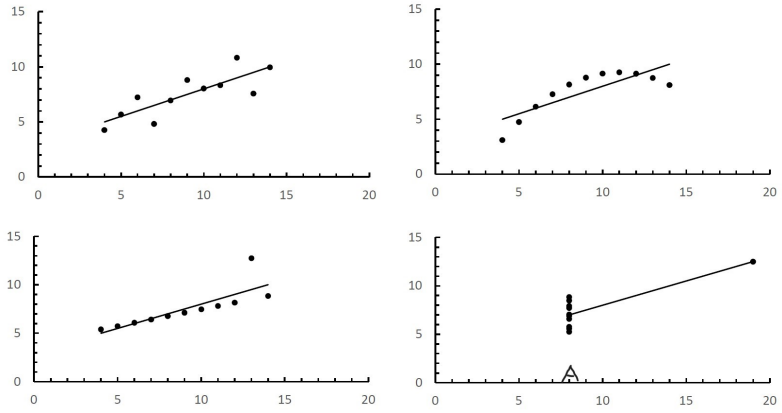

Figure 2: Anscombe Quartet

Similar to what happened in LineLand, Anscombe(1973) showed that the normal view could be very misleading; see Figure 2. The data sets are very different though their least squares summary statistics are identical (to two decimal places). DataSet1 shows the data in the normal case. DataSet2 shows data with $y(x)$ quadratic. DataSet3 shows all of the data colinear except for one observation lying to the north, tilting the least squares fit upwards. DataSet 4 shows the effect of an outlying x observation. The data sets are all very different but their least squares summary statistics are identical.

The Anscombe quartet was the inspiration for later updates, Chatterjee and Firat(2007), and Matejka and Fitzmaurice(2017). They showed how to build different data sets with the same summary statistics. Data appearing as a circle, a cross, and a dinosaur all had

the same summary statistics. You see a circle, you get a dinosaur.

What to do? Anscombe recommended complementing the usual summary statistics with graphical views. "[People use] a library [of] programs and various statistical packages. Most of these originated in the pre-visual era. The user is not showered with graphical displays. He can get them only with trouble, cunning and a fighting spirit. It's time that was changed."

2.1 The Quantile Data View

The residents of FlatLand were better able to see the two-dimensional quartet of Figure 2, but it was tricky. Standing at the origin of DataSet4 the FlatLander saw 11 points in the eastern sky, but not all at once, with perspective. To gain perspective the analyst had to walk on the x-axis toward point A with points rising in the east until at point A, 10 of the 11 eclipsed/merged into a single point directly overhead.

As in LineLand, a different way of envisioning data emerged. The regression version of quantiles, regression quantiles, began to be used as a complement to the usual statistics. The extreme regression quantiles identified the boundary of data: all of the y_i above and all below the respective extreme quantiles. A few of the middle τ quantiles identified lines with τ fraction of the y_i below and $1-\tau$ above.[2]

Figure 3 shows the quantile fits, .1,.5,.9, for the Anscombe data. For the FlatLander without a 3-D viewer the differences in the data sets were discerned via the differences in the quantile coefficients $\beta(.1), \beta(.5), \beta(.9)$; Figure 4.

An unintended consequence of the regression quantile summaries was a reimagining of relationships between y and x. In the olden days people only thought about exact relationships. For the linear model, $y_i = \alpha + \beta x_i$, exactly. Such data was contradicted by observations that had the same x but different ys; $x_1 = x_2$, but $y_1 \neq y_2$. So "y depending on x", exactly, was false. What to do? The answer: reinterpret $y(x)$ as random variable with a distribution of possible outcomes depending on x.

Initially, dependence was framed in terms of the expected value. This resolved the "different y, same x" problem, and it became so popular that the expected value part was forgotten. The conventional wisdom about $y(x)$ came to mean expected value dependence. People would say, "y does not depend on x because my data indicates $E(y|x) \approx 0$". The possibility that the *distribution* of y could nevertheless depend strongly on x was never

[2]For properties see Koenker(2005) and Koenker and Bassett(1978).

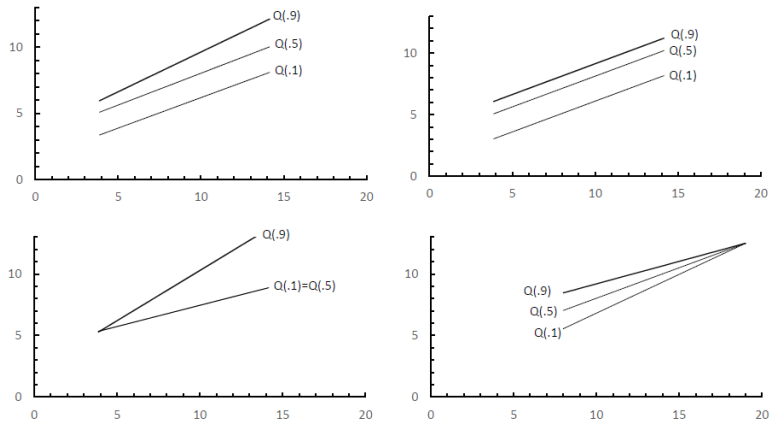

Figure 3: Anscombe Quantiles

considered.

	x-Coefficient				
	Q(.1)	Q(.5)	Q(.9)	OLS	Notes
Data1	.46	.48	.60	.50	Similar
Data2	.50	.50	.50	.50	Similar
Data3	.35	.35	.82	.50	Q(.9)!
Data4	.63	.50	.37	.50	Q(τ) declining

Figure 4: Anscombe Quantile Coefficients

As regression quantiles began to be used to summarize data, analysts wondered if dependence of y on x might be quantile dependent.[3] More than just a way to envision data, quantiles became a model feature. $E(y|x)$ was only one of the ways that y could depend on x. Models were reformulated to allow for and investigate dependence of the

[3]The happiness/income data set is a recent example. Happiness/income data showed expected happiness increasing with income up to about $\$100k$ where it flattened out, plateaued. The happiness plateau became the conventional wisdom. Among other things, the plateau resolved once and for all the St. Petersberg Paradox; Bassett(1987). However, on further review, it was found that the happiest people all had high incomes, no plateau? The conventional wisdom had forgotten that it was about the expected value. The distribution of happiness increased with income at high quantiles, but decreased at low quantiles. The highest happiness was at high incomes, but there were high income people who were sad! While *average* happiness plateaued, the distribution of happiness changed with income; see Koenker(2023).

distribution of y on x at different quantiles.

3 StringLand

In (regression) StringLand data sets are $(y_i, x_i), i = 1, ..., n$, where y_i is univariate and x_i a $(P-1)$-vector, $P \geq 1$, an integer. The variable y_i depends on $x_i = (x_{i1}, ..., x_{iP})$ and an intercept. The simplest and most common dependence is linear, which is our focus. $P = 1$ is LineLand, only an intercept. $P = 2$ is FlatLand.

The Anscombe lessons from LineLand and FlatLand apply to StringLand. Identical least squares summary output corresponds to wildly different data sets. What to do? Graphical views would be nice, but for analysts living in low dimensional subspaces they are not easy to envision.

Alternatively, least squares output can be complemented with a few selected regression quantiles. The regression quantiles provide an enhanced view of the data. They correspond to hyperplanes (for linear models) slicing through the cloud of data. The extreme quantiles identify the top and bottom of the data; all the y_i are above the bottom quantile hyperpane, and all the y_i are below the top quantile. The middle of the data is captured by selected quantile hyperplanes; the fraction of the data below is τ and the fraction above is $1 - \tau$.

The $\hat{\beta}(\tau)$ coefficients provide a sense of what the data looks like. For a given dependent variable, the $\hat{\beta}_p(.1), \hat{\beta}_p(.5), \hat{\beta}_p(.9)$, indicate how x_p is influencing the $.1, .5, .9$ quantiles of y. Similar coefficients correspond to similar effects–parallel quantile hyperplanes. Dissimilar coefficients mean quantile hyperplanes tilting this way and that, as in Figure 3.

The lesson of the Anscombe quartet is that important differences in data are not revealed by the usual statistics. The quantile view enhances data analysis by revealing features that would otherwise be missed.

References

[1] Abbot, Edwin A.(1884). *Flatland: A Romance of Many Dimensions*, Seeley&Co, 1884.

[2] Anscombe, F.J.(1973). Graphs in Statistical Analysis. *The American Statistician*, Vol. 27, No. 1. (Feb., 1973), pp. 17-21.

[3] Bassett, Gilbert (1987). The St. Petersburg Paradox and Bounded Utility, *History of Political Economy*, 1987, v. 19, n. 4. p. 517-523.

[4] Chatterjee, Sangit and Aykut Firat(2007). Generating Data with Identical Statistics: A Followup to the Anscombe Dataset. *The American Statistician*, Aug., 2007, Vol.61, No.3(Aug., 2007), pp.248-254. https ://www.jstor.org/stable/27643902.

[5] Koenker, Roger (2005). *Quantile Regression*. Cambridge University Press.

[6] Koenker, Roger (2023). *How Quantile Regression Rescued Happiness-For the Rich: An R Vinaigrette*. http://www.econ.uiuc.edu/ roger/research/vinaigrettes/Hap.pdf

[7] Koenker, Roger and Gilbert Bassett Jr. (1978). Regression Quantiles. *Econometrica*, Vol. 46, No. 1 (Jan., 1978), pp. 33-50

[8] Matejka, Justin and George Fitzmaurice(2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. *Autodesk Research*, Toronto Ontario Canada ACM 978-1-4503-4655-9/17/05.

[9] Samuelson, Paul A.(1968). How Deviant Can You Be? *Journal of the American Statistical Association*, Dec., 1968, Vol. 63, No. 324 pp. 1522-1525.