CHAPTER 1

Introduction

1.1 MEANS AND ENDS

Much of applied statistics may be viewed as an elaboration of the linear regression model and associated estimation methods of least squares. In beginning to describe these techniques, Mosteller and Tukey (1977), in their influential text, remark:

> What the regression curve does is give a grand summary for the averages of the distributions corresponding to the set of xs. We could go further and compute several different regression curves corresponding to the various percentage points of the distributions and thus get a more complete picture of the set. Ordinarily this is not done, and so regression often gives a rather incomplete picture. Just as the mean gives an incomplete picture of a single distribution, so the regression curve gives a correspondingly incomplete picture for a set of distributions.

My objective in the following pages is to describe explicitly how to "go further." Quantile regression is intended to offer a comprehensive strategy for completing the regression picture.

Why does least-squares estimation of the linear regression model so pervade applied statistics? What makes it such a successful tool? Three possible answers suggest themselves. One should not discount the obvious fact that the computational tractability of linear estimators is extremely appealing. Surely this was the initial impetus for their success. Second, if observational noise is normally distributed (i.e., Gaussian), least-squares methods are known to enjoy a certain optimality. But, as it was for Gauss himself, this answer often appears to be an *ex post* rationalization designed to replace the first response. More compelling is the relatively recent observation that least-squares methods provide a general approach to estimating conditional mean functions.

And yet, as Mosteller and Tukey suggest, the mean is rarely a satisfactory end in itself, even for statistical analysis of a single sample. Measures of spread, skewness, kurtosis, boxplots, histograms, and more sophisticated density estimation are all frequently employed to gain further insight. Can something similar be done in regression? A natural starting place for this would be to

supplement the conditional mean surfaces estimated by least squares with several estimated conditional quantile surfaces. In the chapters that follow, methods are described to accomplish this task. The basic ideas go back to the earliest work on regression by Boscovich in the mid-18th century to Edgeworth at the end of the 19th century.

1.2 THE FIRST REGRESSION: A HISTORICAL PRELUDE

It is ironic that the first faltering attempts to *do* regression are so closely tied to the notions of quantile regression. Indeed, as I have written on a previous occasion, the present enterprise might be viewed as an attempt to set statistics back 200 years, to the idyllic period before the discovery of least squares.

If least squares can be dated to 1805 by the publication of Legendre's work on the subject, then Boscovich's initial work on regression was half a century prior. The problem that interested Boscovich was the ellipticity of the earth. Newton and others had suggested that the earth's rotation could be expected to make it bulge at the equator with a corresponding flattening at the poles, making it an oblate spheroid, more like a grapefruit than a lemon. On the early history of regression and the contribution of Boscovich in particular, Stigler (1986) is the definitive introduction. Smith (1987) gives a detailed account of the development of geodesy, focusing attention on the efforts that culminated in the data appearing in Table 1.1.

To estimate the extent of this effect, the five measurements appearing in Table 1.1 had been made. Each represented a rather arduous direct measurement of the arc-length of 1° of latitude at five quite dispersed points – from Quito on the equator to a site in Lapland at 66°19′N. It was clear from these measurements that arc length was increasing as one moved toward the pole from the equator, thus qualitatively confirming Newton's conjecture. But how the five measurements should be combined to produce one estimate of the earth's ellipticity was unclear.

For short arcs, the approximation

$$y = a + b \sin^2 \lambda, \tag{1.1}$$

Table 1.1. *Boscovich ellipticity data*

Location	Latitude	$\sin^2$ (Latitude)	Arc Length
Quito	0° 0′	0	56,751
Cape of Good Hope	33° 18′	0.2987	57,037
Rome	42° 59′	0.4648	56,979
Paris	49° 23′	0.5762	57,074
Lapland	66° 19′	0.8386	57,422

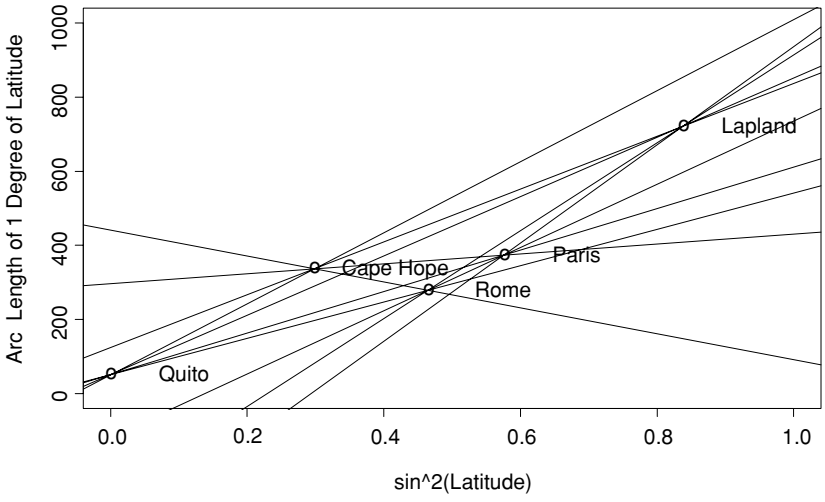

Figure 1.1. Boscovich ellipticity example. Boscovich computed all the pairwise slopes and initially reported a trimmed mean of the pairwise slopes as a point estimate of the earth's ellipticity. Arc length is measured as the excess over 56,700 toise per degree where one toise $\approx$ 6.39 feet, or 1.95 meters.

where y is the length of the arc and λ is the latitude, was known to be satisfactory. The parameter a could be interpreted as the length of a degree of arc at the equator and b as the exceedence of a degree of arc at the pole over its value at the equator. Ellipticity could then be computed as $1/\text{ellipticity} = \eta = 3a/b$. Boscovich, noting that any pair of observations could be used to compute an estimate of a and b, hence of η, began by computing all ten such estimates. These lines are illustrated in Figure 1.1. Some of these lines seemed quite implausible, especially perhaps the downward-sloping one through Rome and the Cape of Good Hope. Boscovich reported two final estimates: one based on averaging all ten distinct estimates of b, the other based on averaging all but two of the pairwise slopes with the smallest implied exceedence. In both cases the estimate of a was taken directly from the measured length of the arc at Quito. These gave ellipticities of 1/155 and 1/198, respectively. A modern variant on this idea is the median of pairwise slopes suggested by Theil (1950), which yields the somewhat lower estimate of 1/255.

It is a curiosity worth noting that the least-squares estimator of (a, b) may also be expressed as a weighted average of the pairwise slope estimates. Let h index the ten pairs, and write

$$b(h) = X(h)^{-1} y(h), \tag{1.2}$$

where, for the simple bivariate model and $h = (i, j)$,

$$X(h) = \begin{pmatrix} 1 & x_i \\ 1 & x_j \end{pmatrix} \quad y(h) = \begin{pmatrix} y_i \\ y_j \end{pmatrix}; \tag{1.3}$$

then we may write the least-squares estimator as

$$\hat{b} = \sum_h w(h)b(h), \tag{1.4}$$

where $w(h) = |X(h)|^2 / \sum_h |X(h)|^2$. As shown by Subrahmanyam (1972) and elaborated by Wu (1986), this representation of the least-squares estimator extends immediately to the general p-parameter linear regression model. In the bivariate example the weights are obviously proportional to the distance between each pair of design points, a fact that, in itself, portends the fragility of least squares to outliers in either x or y observations.

Boscovich's second attack on the ellipticity problem formulated only two years later brings us yet closer to quantile regression. In effect, he suggests estimating (a, b) in (1.1) by minimizing the sum of absolute errors subject to the constraint that the errors sum to zero. The constraint requires that the fitted line pass through the centroid of the observations, $(\bar{x}, \bar{y})$. Boscovich provided a geometric algorithm, which was remarkably simple, to compute the estimator. Having reduced the problem to regression through the origin with the aid of the constraint, we may imagine rotating a line through the new origin at $(\bar{x}, \bar{y})$ until the sum of absolute residuals is minimized. This may be viewed algebraically, as noted later by Laplace, as the computation of a *weighted median*. For each point we may compute

$$b_i = \frac{y_i - \bar{y}}{x_i - \bar{x}} \tag{1.5}$$

and associate with each slope the weight $w_i = |x_i - \bar{x}|$. Now let $b_{(i)}$ be the ordered slopes and $w_{(i)}$ the associated weights, and find the smallest j, say j^*, such that

$$\sum_{i=1}^{j} w_{(i)} > \frac{1}{2} \sum_{i=1}^{n} w_{(i)} \tag{1.6}$$

The Boscovich estimator, $\hat{\beta} = b_{(j^*)}$, was studied in detail by Laplace in 1789 and later in his monumental *Traite de Méchanique Céleste*. Boscovich's proposal, which Laplace later called the "method of situation," is a curious blend of mean and median ideas; in effect, the slope parameter b is estimated as a median, while the intercept parameter a is estimated as a mean.

This was clearly recognized by Edgeworth, who revived these ideas in 1888 after nearly a century of neglect. In his early work on index numbers and weighted averages, Edgeworth had emphasized that the putative optimality of the sample mean as an estimator of location was crucially dependent on the assumption that the observations came from a common normal distribution. If the observations were "discordant," say from normals with different variances, the median, he argued, could easily be superior to the mean. Indeed, anticipating the work of Tukey in the 1940s, Edgeworth compares the asymptotic variances of the median and mean for observations from scale mixtures of normals,

concluding that, for equally weighted mixtures with relative scale greater than 2.25, the median had smaller asymptotic variance than the mean.

Edgeworth's work on median methods for linear regression brings us directly to quantile regression. Edgeworth (1888) discards the Boscovich–Laplace constraint that the residuals sum to zero and proposes to minimize the sum of absolute residuals in both intercept and slope parameters, calling it a "double median" method and noting that it could be extended, in principle, to a "plural median" method. A geometric algorithm was given for the bivariate case, and a discussion of conditions under which one would prefer to minimize absolute error rather than the by-then well-established squared error is provided. Unfortunately, the geometric approach to computing Edgeworth's new median regression estimator was rather awkward, requiring, as he admitted later, "the attention of a mathematician; and in the case of many unknowns, some power of hypergeometrical conception." Only considerably later did the advent of linear programming provide a conceptually simple and efficient computational approach.

Once we have a median regression estimator it is natural to ask, "are there analogs for regression of the other quantiles?" The answer to this question is explored in the next section.

1.3 QUANTILES, RANKS, AND OPTIMIZATION

Any real-valued random variable X may be characterized by its (right-continuous) distribution function

$$F(x) = P(X \leq x), \tag{1.7}$$

whereas for any $0 < \tau < 1$,

$$F^{-1}(\tau) = \inf\{x : F(x) \geq \tau\} \tag{1.8}$$

is called the τth quantile of X. The median, $F^{-1}(1/2)$, plays the central role.

The quantiles arise from a simple optimization problem that is fundamental to all that follows. Consider a simple decision theoretic problem: a point estimate is required for a random variable with (posterior) distribution function F. If loss is described by the piecewise linear function illustrated in Figure 1.2

$$\rho_\tau(u) = u(\tau - I(u < 0)) \tag{1.9}$$

for some $\tau \in (0, 1)$, find $\hat{x}$ to minimize expected loss. This is a standard exercise in decision theory texts (e.g., Ferguson, 1967, p. 51). The earliest reference that I am aware of is Fox and Rubin (1964), who studied the admissibility of the quantile estimator under this loss function. We seek to minimize

$$E\rho_\tau(X - \hat{x}) = (\tau - 1)\int_{-\infty}^{\hat{x}} (x - \hat{x})dF(x) + \tau \int_{\hat{x}}^{\infty} (x - \hat{x})dF(x). \tag{1.10}$$

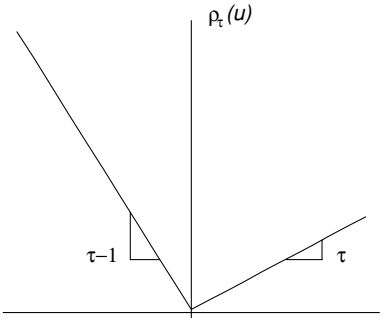

Figure 1.2. Quantile regression ρ function.

Differentiating with respect to $\hat{x}$, we have

$$0 = (1-\tau)\int_{-\infty}^{\hat{x}} dF(x) - \tau\int_{\hat{x}}^{\infty} dF(x) = F(\hat{x}) - \tau. \tag{1.11}$$

Since F is monotone, any element of $\{x : F(x) = \tau\}$ minimizes expected loss. When the solution is unique, $\hat{x} = F^{-1}(\tau)$; otherwise, we have an "interval of τth quantiles" from which the smallest element must be chosen – to adhere to the convention that the empirical quantile function be left-continuous.

It is natural that an optimal point estimator for asymmetric linear loss should lead us to the quantiles. In the symmetric case of absolute value loss it is well known to yield the median. When loss is linear and asymmetric, we prefer a point estimate more likely to leave us on the flatter of the two branches of marginal loss. Thus, for example, if an underestimate is *marginally* three times more costly than an overestimate, we will choose $\hat{x}$ so that $P(X \leq \hat{x})$ is three times greater than $P(X > \hat{x})$ to compensate. That is, we will choose $\hat{x}$ to be the 75th percentile of F.

When F is replaced by the empirical distribution function

$$F_n(x) = n^{-1}\sum_{i=1}^{n} I(X_i \leq x), \tag{1.12}$$

we may still choose $\hat{x}$ to minimize expected loss:

$$\int \rho_\tau(x-\hat{x})dF_n(x) = n^{-1}\sum_{i=1}^{n} \rho_\tau(x_i - \hat{x}) \tag{1.13}$$

and doing so now yields the τth *sample* quantile. When τn is an integer there is again some ambiguity in the solution, because we really have an interval of solutions, $\{x : F_n(x) = \tau\}$, but we shall see that this is of little practical consequence.

Much more important is the fact that we have expressed the problem of finding the τth sample quantile, a problem that might seem inherently tied to the notion of an ordering of the sample observations, as the solution to a simple

optimization problem. In effect we have replaced *sorting* by *optimizing*. This will prove to be the key idea in generalizing the quantiles to a much richer class of models in subsequent chapters. Before doing this, though, it is worth examining the simple case of the ordinary sample quantiles in a bit more detail.

The problem of finding the τth sample quantile, which may be written as

$$\min_{\xi \in \mathbf{R}} \sum_{i=1}^{n} \rho_\tau(y_i - \xi), \tag{1.14}$$

may be reformulated as a linear program by introducing $2n$ artificial, or "slack," variables $\{u_i, v_i : 1, \ldots, n\}$ to represent the positive and negative parts of the vector of residuals. This yields the new problem

$$\min_{(\xi,u,v)\in\mathbb{R}\times\mathbb{R}^{2n}_+} \left\{ \tau 1_n^\top u + (1-\tau) 1_n^\top v \,|\, 1_n \xi + u - v = y \right\}, \tag{1.15}$$

where 1_n denotes an n-vector of 1. Clearly, in (1.15) we are minimizing a linear function on a polyhedral constraint set consisting of the intersection of the $(2n+1)$-dimensional hyperplane determined by the linear equality constraints and the set $\mathbb{R} \times \mathbb{R}^{2n}_+$.

Figure 1.3 illustrates the most elementary possible version of the median linear programming problem. We have only one observation, at $y = 1$, and we wish to solve

$$\min_{(\xi,u,v)\in\mathbb{R}\times\mathbb{R}^{2}_+} \{u + v | \xi + u - v = y\}.$$

The constraint set is the triangular region representing the intersection of the plane $\{(\xi, u, v) | \xi + u - v = 1\}$ with the cone $\{(\xi, u, v) \in \mathbb{R}^3 | u \geq 0, v \geq 0\}$. The long edge of this triangle extends off into the deeper regions of the figure. The objective function is represented by a series of vertical planes perpendicular to the 45° line in the (u, v) (horizontal) plane. Moving toward the origin reduces $u + v$, thus improving the objective function. It is apparent that any feasible point (ξ, u, v) that has both u and v strictly positive can be improved by reducing v and increasing u to compensate. But with only one observation we can move further. Reducing u and increasing ξ to compensate – that is, moving along the interior edge of the constraint set – allows us to reduce the objective function to zero, setting $\xi = 1$, coming to rest at the upper-left corner of the constraint set. Now, if we try to imagine increasing the number of observations, we have contributions to the objective function from each observation like the one illustrated in Figure 1.3. Given a trial value of the parameter ξ, we can consider a feasible point that sets each u_i equal to the positive part of the residual $y_i - \xi$ and v_i equal to the negative part of the ith residual. But, as in Figure 1.3, such solutions can always be improved by moving ξ closer to one of the sample observations.

Many features of the solution are immediately apparent from these simple observations. To summarize, $\min\{u_i, v_i\}$ must be zero for all i, because otherwise the objective function may be reduced without violating the constraint

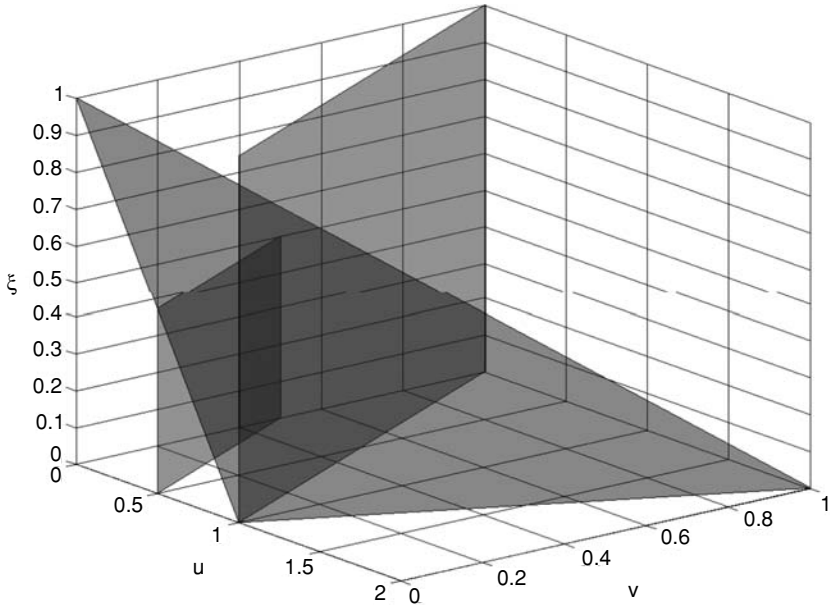

Figure 1.3. Computing the median with one observation. The figure illustrates the linear programming formulation of the median problem. The triangular region represents the constraint set; the vertical planes represent two different contours of the objective function, which decreases as ones moves toward the origin in the (u, v)-plane.

by shrinking such a pair toward zero. This is usually called complementary slackness in the terminology of linear programming. Indeed, for essentially the same reason we can restrict attention to "basic solutions" of the form $\xi = y_i$ for some observation i. Figure 1.4 depicts objective function (1.14) for three different random samples of varying sizes. The graph of the objective function is convex and piecewise linear with kinks at the observed y_is. When ξ passes through one of these y_is, the slope of the objective function changes by exactly 1 since a contribution of $\tau - 1$ is replaced by τ or vice versa.

Optimality holds at a point $\hat{\xi}$ if the objective function

$$R(\xi) = \sum_{i=1}^{n} \rho_\tau(y_i - \xi)$$

is increasing as one moves away from $\hat{\xi}$ to either the right or the left. This requires that the right and left derivatives of R are both nonnegative at the point $\hat{\xi}$. Thus,

$$R'(\xi+) \equiv \lim_{h \to 0} (R(\xi + h) - R(\xi))/h = \sum_{i=1}^{n} (I(y_i < \xi + 0) - \tau)$$

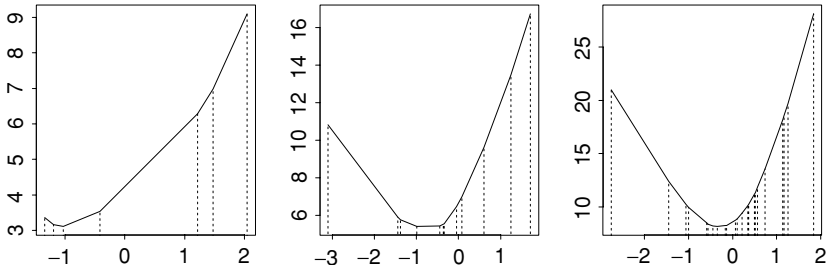

Figure 1.4. Quantile objective function with random data. The figure illustrates the objective function for the optimization problem defining the ordinary $\tau = 1/3$ quantile for three different random problems with y_is drawn from the standard normal distribution and sample sizes 7, 12, and 23. The vertical dotted lines indicate the position of the observations in each sample. Note that because 12 is divisible by 3, the objective function is flat at its minimum in the middle figure, and we have an interval of solutions between the fourth- and fifth-largest observations.

and

$$R'(\xi-) \equiv \lim_{h\to 0}(R(\xi - h) - R(\xi))/h = \sum_{i=1}^{n}(\tau - I(y_i < \xi - 0))$$

must both be nonnegative, and so $n\tau$ lies in the closed interval $[N^-, N^+]$, where N^+ denotes the number of y_i less than or equal to ξ and N^- denotes the number of y_i strictly less than ξ. When $n\tau$ is not an integer, there is a unique value of ξ that satisfies this condition. Barring ties in the y_is, this value corresponds to a unique order statistic. When there are ties, ξ is still unique, but there may be several y_i equal to ξ. If $n\tau$ *is* an integer then $\hat{\xi}_\tau$ lies between two adjacent order statistics. It is unique only when these order statistics coalesce at a single value. Usually, we can dismiss the occurrence of such ties as events of probability zero.

The duality connecting the sample quantiles and the ranks of the order statistics is further clarified through the formal duality of linear programming. While the primal problem, (1.15), may be viewed as generating the sample quantiles, the corresponding dual problem may be seen to generate the order statistics, or perhaps more precisely the *ranks* of the observations. This approach to ranks generalizes naturally to the linear model, yielding an elegant generalization of rank tests for the linear model.

1.4 PREVIEW OF QUANTILE REGRESSION

The observation developed in Section 1.3 that the quantiles may be expressed as the solution to a simple optimization problem leads, naturally, to more general methods of estimating models of conditional quantile functions. Least squares offers a template for this development. Knowing that the sample mean solves

the problem

$$\min_{\mu \in \mathbb{R}} \sum_{i=1}^{n} (y_i - \mu)^2 \tag{1.16}$$

suggests that, if we are willing to express the *conditional* mean of y given x as $\mu(x) = x^\top \beta$, then β may be estimated by solving

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - x_i^\top \beta \right)^2 . \tag{1.17}$$

Similarly, since the τth sample quantile, $\hat{\alpha}(\tau)$, solves

$$\min_{\alpha \in \mathbb{R}} \sum_{i=1}^{n} \rho_\tau (y_i - \alpha), \tag{1.18}$$

we are led to specifying the τth *conditional* quantile function as $Q_y(\tau|x) = x^\top \beta(\tau)$, and to consideration of $\hat{\beta}(\tau)$ solving

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho_\tau \left(y_i - x_i^\top \beta \right) . \tag{1.19}$$

This is the germ of the idea elaborated by Koenker and Bassett (1978).

Quantile regression problem (1.19) may be reformulated as a linear program as in (1.15):

$$\min_{(\beta, u, v) \in \mathbb{R}^p \times \mathbb{R}^{2n}_+} \left\{ \tau 1_n^\top u + (1 - \tau) 1_n^\top v | X\beta + u - v = y \right\}, \tag{1.20}$$

where X now denotes the usual n by p regression design matrix. Again, we have split the residual vector $y - X\beta$ into its positive and negative parts, and so we are minimizing a linear function on a polyhedral constraint set, and most of the important properties of the solutions, $\hat{\beta}(\tau)$, which we call "regression quantiles," again follow immediately from well-known properties of solutions of linear programs.

We can illustrate the regression quantiles in a very simple bivariate example by reconsidering the Boscovich data. In Figure 1.5 we illustrate all of the *distinct* regression quantile solutions for this data. Of the ten lines passing through pairs of points in Figure 1.1, quantile regression selects only four. Solving (1.19) for any τ in the interval $(0, 0.21)$ yields as a unique solution the line passing through Quito and Rome. At $\tau = 0.21$, the solution jumps, and throughout the interval $(0.21, 0.48)$ we have the solution characterized by the line passing through Quito and Paris. The process continues until we get to $\tau = 0.78$, where the solution through Lapland and the Cape of Good Hope prevails up to $\tau = 1$.

In contrast to the ordinary sample quantiles that are equally spaced on the interval [0,1], with each distinct order statistic occupying an interval of length exactly $1/n$, the lengths of the regression quantile solution intervals for $\tau \in [0, 1]$ are irregular and depend on the configuration of the design as well as the realized values of the response variable. *Pairs of points now play the role*