REGRESSION

"The invalid assumption that correlation implies cause is probably among the two or three most serious and common errors of human reasoning."
Stephen Jay Gould, Paleontologist

Regression refers to a subset of data modeling where we fit a simple model with a linear trend in one (or more) dimensions. Usually we also include a constant (intercept) term. Entire books have been written about regression and all the various methods, norms, and misfit-minimizations possible. A summary of some of these developments will be given below.

6.1 Line-Fitting Revisited

We will again consider the best-fitting line problem $y = a + bx$, this time with errors s_i in the observed y-values. We want to measure how well the model agrees with the data, and for this purpose we will use the χ^2 function, i.e.,

$$\chi^2(a,b) = \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{s_i} \right)^2 . \tag{6.1}$$

Minimizing χ^2 will give the best weighted least squares solution. Again, we set the partial derivatives to zero and obtain

$$\frac{\partial \chi^2}{\partial a} = 0 = -2 \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{s_i^2} \right),$$

$$\frac{\partial \chi^2}{\partial b} = 0 = -2 \sum_{i=1}^{n} \left(\frac{y_i - a - bx_i}{s_i^2} \right) x_i . \tag{6.2}$$

Let us define the following terms (unless noted, all sums go from $i = 1$ to n):

$$S = \sum \frac{1}{s_i^2}, \quad S_x = \sum \frac{x_i}{s_i^2}, \quad S_y = \sum \frac{y_i}{s_i^2}, \quad S_{xx} = \sum \frac{x_i^2}{s_i^2}, \quad S_{xy} = \sum \frac{x_i y_i}{s_i^2}. \tag{6.3}$$

REGRESSION

Then, (6.2) reduces to

$$\begin{aligned} aS + bS_x &= S_y, \\ aS_x + bS_{xx} &= S_{xy}. \end{aligned} \tag{6.4}$$

Introducing

$$\Delta = SS_{xx} - S_x^2 \tag{6.5}$$

we find

$$\begin{aligned} a &= \frac{S_{xx}S_y - S_xS_{xy}}{\Delta}, \\ b &= \frac{SS_{xy} - S_xS_y}{\Delta}. \end{aligned} \tag{6.6}$$

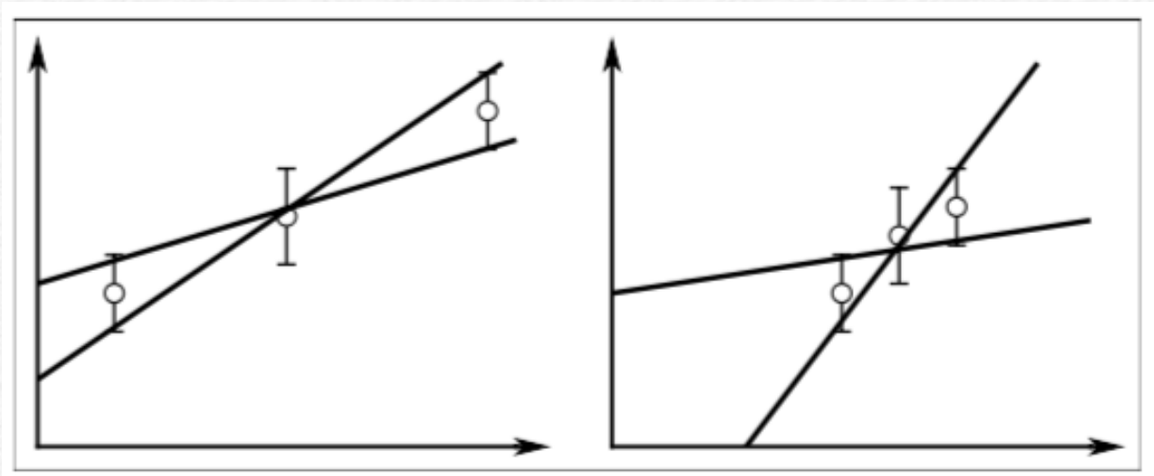

Figure 6.1: The uncertainty in the line fit depends to a large extent on the distribution of the *x*-positions as well as the uncertainties in the *y*-values.

All this is swell but we must also estimate the uncertainties in *a* and *b*. For the same s_i we may get large differences in the uncertainties in *a* and *b* (e.g., Figure 6.1). As shown in Chapter 1, consideration of the propagation of errors (e.g., 2.37) shows that the variance σ_f^2 in the value of any function is

$$\sigma_f^2 = \sum \left(\frac{\partial f}{\partial y_i} \sigma_i \right)^2, \tag{6.7}$$

where we now consider *f* a function of all the *n* independent parameters y_i. For our model, *f* is either *a* or *b* so the partial derivatives become

$$\begin{aligned} \frac{\partial a}{\partial y_i} &= \frac{S_{xx} - S_x x_i}{s_i^2 \Delta}, \\ \frac{\partial b}{\partial y_i} &= \frac{S x_i - S_x}{s_i^2 \Delta}. \end{aligned} \tag{6.8}$$

REGRESSION

Inserting in turn these terms into (6.7) now gives

$$\begin{aligned} s_a^2 &= \sum s_i^2 \left[\frac{S_{xx} - S_x x_i}{s_i^2 \Delta}\right]^2 = \sum \frac{S_{xx}^2 - 2S_{xx}S_x x_i + S_x^2 x_i^2}{s_i^2 \Delta^2} \\ &= \frac{S_{xx}^2}{\Delta^2}\sum \frac{1}{s_i^2} - \frac{2S_{xx}S_x}{\Delta^2}\sum \frac{x_i}{s_i^2} + \frac{S_x^2}{\Delta^2}\sum \frac{x_i^2}{s_i^2} = \frac{S_{xx}^2 S}{\Delta^2} - \frac{2S_{xx}S_x^2}{\Delta^2} + \frac{S_{xx}S_x^2}{\Delta^2} \\ &= \frac{S_{xx}(S_{xx}S - S_x^2)}{\Delta^2} = \frac{S_{xx}}{\Delta} \end{aligned} \tag{6.9}$$

and

$$\begin{aligned} s_b^2 &= \sum s_i^2 \left[\frac{S x_i - S_x}{s_i^2 \Delta}\right]^2 = \sum \frac{S^2 x_i^2 - 2S\, S_x x_i + S_x^2}{s_i^2 \Delta^2} \\ &= \frac{S^2}{\Delta^2}\sum \frac{x_i^2}{s_i^2} - \frac{2S\, S_x}{\Delta^2}\sum \frac{x_i}{s_i^2} + \frac{S_x^2}{\Delta^2}\sum \frac{1}{s_i^2} = \frac{S^2 S_{xx}}{\Delta^2} - \frac{2S\, S_x^2}{\Delta^2} + \frac{S S_x^2}{\Delta^2} \\ &= \frac{S(S_{xx}S - S_x^2)}{\Delta^2} = \frac{S}{\Delta}. \end{aligned} \tag{6.10}$$

Similarly, we can find the covariance s_{ab} from

$$s_{ab}^2 = \sum s_i^2 \left(\frac{\partial a}{\partial y_i}\right)\left(\frac{\partial b}{\partial y_i}\right) = -\frac{S_x}{\Delta}. \tag{6.11}$$

Thus, the correlation between a and b becomes

$$r_{ab} = \frac{-S_x}{\sqrt{S S_{xx}}}. \tag{6.12}$$

It is therefore useful to shift the origin to $\bar{x}$ so that $r_{ab} = 0$, leaving our estimates for slope and intercept uncorrelated. Finally, we must check if the fit is significant. We determine critical χ^2_α for $n - 2$ degrees of freedom and test if our computed χ^2 exceeds the critical limit. If it does not, then we may say the fit is *significant* at the α level of confidence.

REGRESSION

6.1.1 Confidence interval on regression

The formalism in the previous section allowed us to derive solutions for slope and intercept given via (6.6). Let us for a moment consider the case where there are no individual uncertainties σ_i associated with the data. We write the least squares fit as $\hat{y} = a + bx$, and by substituting $a = \bar{y} - b\bar{x}$ we obtain

$$\hat{y} = \bar{y} + b(x - \bar{x}). \tag{6.13}$$

Here, both the mean y-value ($\bar{y}$) and slope (b) are subject to error and these in turn affect $\hat{y}$. For some chosen location x_0 the prediction for the regression would be

$$\hat{y}_0 = \bar{y} + b(x_0 - \bar{x}). \tag{6.14}$$

In this formulation, with a local origin at ($\bar{x}$, $\bar{y}$), the correlation between $\bar{y}$ and b is zero. This fact allows us to compute the expected variance V of the *mean predicted value* thus:

$$V(\hat{y}_0) = V(\bar{y}) + V(b)(x - \bar{x})^2 = \frac{s^2}{n} + \frac{s^2(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}. \tag{6.15}$$

Here, s is our sample estimate of the standard deviation of the regression residuals,

$$s^2 = \frac{\sum(y_i - \bar{y})^2}{n - 2}. \tag{6.16}$$

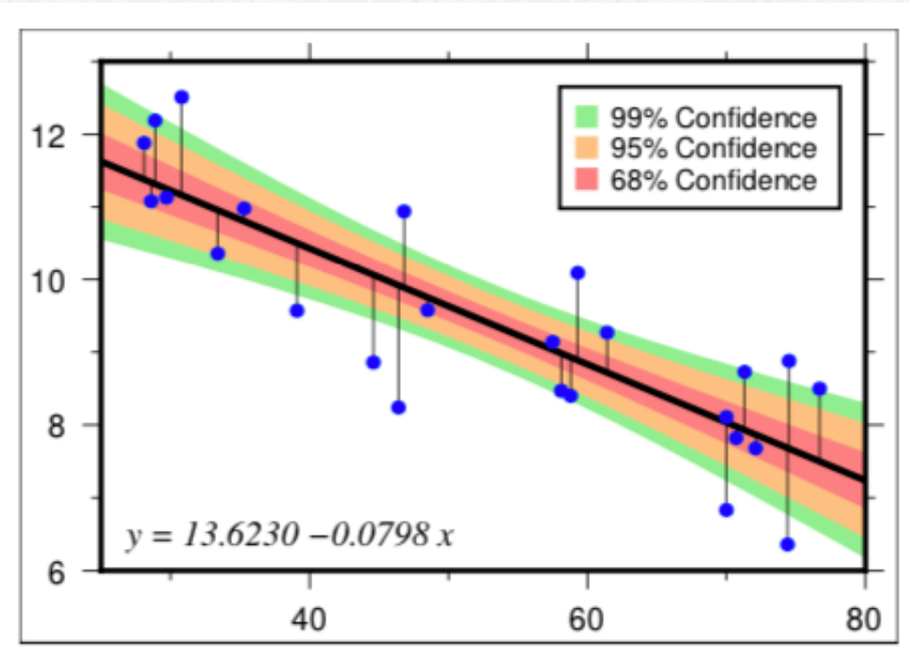

Figure 6.2: Solid line shows the least squares regression fit to the data points (blue circles), with color bands reflecting different confidence levels. Short vertical lines are the residual errors which are squared and summed in (6.16).

Because $\bar{y}$ and b are uncorrelated we obtain the variance of an individual observation by adding the independent variance of y about the mean, which is s^2, and find the regression variance for the individual prediction at $x = x_0$ to be

$$s_0^2 = s^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_i - \bar{x})^2}\right]. \tag{6.17}$$

REGRESSION

To obtain confidence intervals on the linear regression we simply scale (6.17) by a critical Student's *t*-value for the degrees of freedom ν and chosen confidence level (Table A.2), i.e.,

$$\hat{y}_0 \pm t(\nu, \alpha/2) s \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \tag{6.18}$$

For larger data sets the Student's *t* values approach the normal distribution critical values $|z_{\alpha/2}|$. Figure 6.2 illustrates how confidence bands on a least-squares regression fit takes on a parabolic shape around the best-fit line.

6.2 Orthogonal Regression

6.2.1 Major axis

It is often the case that the uncertainties in our (*x*, *y*) data affect both coordinates. Examples where this is the case include situations where both *x* and *y* are observed quantities (and hence are known to have errors). It is also applicable when *y* is a function of *x*, but *x* (e.g., distance or time) itself has uncertainties. In these cases, orthogonal regression is the correct way to determine linear relationships between *x* and *y* (Figure 6.3). We will use the least squares principle and minimize the sum of the squared *perpendicular* distances d_i^2 from the data points (x_i, y_i) to the regression line.

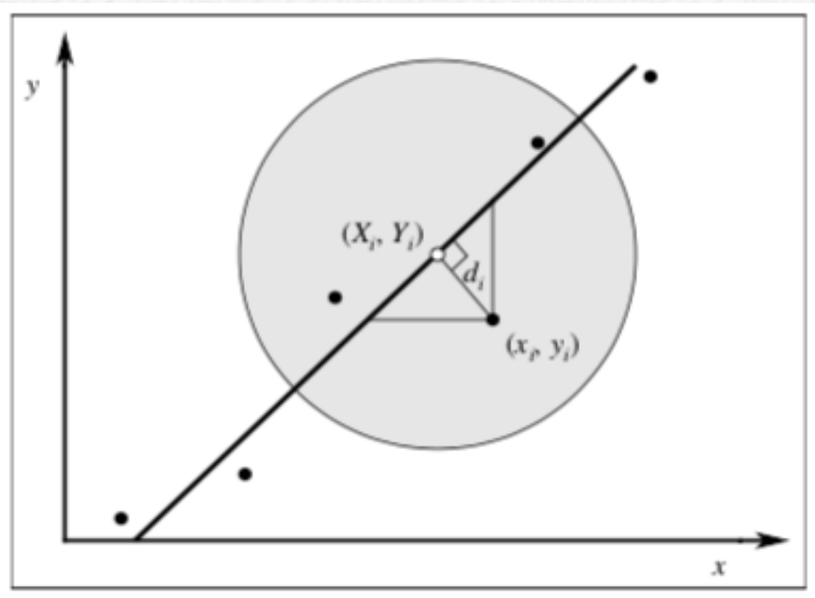

Figure 6.3: The misfit is measured in the direction perpendicular to the line. Note that (x_i, y_i) denote our data points (black circles) and (X_i, Y_i) are the coordinates of their orthogonal projections (white circles; only one is shown here) onto the regression line.

The function we want to minimize is

$$E = \sum_{i=1}^{n} \left[(X_i - x_i)^2 + (Y_i - y_i)^2 \right], \tag{6.19}$$

where lowercase (x_i, y_i) again are our observations and uppercase (X_i, Y_i) are the "adjusted" coordinates we want to find. These are, of course, required to lie on a straight line described by

$$Y_i = a + bX_i, \tag{6.20}$$

REGRESSION

or equivalently

$$f_i = a + bX_i - Y_i = 0. \tag{6.21}$$

Thus, we cannot simply find *any* set of (X_i, Y_i) as they also have to lie on a straight line. The problem of minimizing the function (6.19) under the specified constraints (6.21) can be solved by a method known as *Lagrange's multipliers*. This method says we should form a new function *F* by adding the original function (6.19) and all the constraints (6.21), with each constraint scaled by an unknown (Lagrange) multiplier λ_i. Since (6.21) is actually *n* constraints, we find

$$F = E + \lambda_1 f_1 + \lambda_2 f_2 + \ldots + \lambda_n f_n = E + \sum_{i=1}^{n} \lambda_i f_i. \tag{6.22}$$

We may now set the partial derivatives of *F* to zero and solve the resulting set of equations:

$$\frac{\partial F}{\partial X_i} = \frac{\partial F}{\partial Y_i} = \frac{\partial F}{\partial a} = \frac{\partial F}{\partial b} = 0, \tag{6.23}$$

or, when viewed separately (with all sums over *i* = 1 to *n* unless explicitly stated),

$$\frac{\partial F}{\partial X_i} = \sum_j^n \frac{\partial}{\partial X_i}(X_j - x_j)^2 + \sum_j^n \frac{\partial}{\partial X_i}(\lambda_j b X_j) = 0 \Rightarrow 2(X_i - x_i) + b\lambda_i = 0, \tag{6.24}$$

$$\frac{\partial F}{\partial Y_i} = \sum_j^n \frac{\partial}{\partial Y_i}(Y_j - y_j)^2 - \sum_j^n \frac{\partial}{\partial Y_i}(\lambda_j Y_j) = 0 \Rightarrow 2(Y_i - y_i) - \lambda_i = 0, \tag{6.25}$$

$$\frac{\partial F}{\partial a} = \sum \frac{\partial}{\partial a}(\lambda_i a) = 0 \Rightarrow \sum \lambda_i = 0, \tag{6.26}$$

$$\frac{\partial F}{\partial b} = \sum \frac{\partial}{\partial b}(\lambda_i b X_i) = 0 \Rightarrow \sum \lambda_i X_i = 0. \tag{6.27}$$

Since each *i* represents a separate equation, we find

$$2(X_i - x_i) = -b\lambda_i \Rightarrow X_i = x_i - b\lambda_i/2, \tag{6.28}$$

$$2(Y_i - y_i) = \lambda_i \Rightarrow Y_i = y_i + \lambda_i/2. \tag{6.29}$$

REGRESSION

Substituting these expressions for X_i and Y_i into (6.20), we obtain

$$y_i + \lambda_i/2 = a + b\,(x_i - b\lambda_i/2) = a + bx_i - b^2\lambda_i/2 \tag{6.30}$$

or

$$\lambda_i = \frac{2}{1+b^2}\,(a + bx_i - y_i)\,. \tag{6.31}$$

Now, (6.26), (6.27), and (6.31) gives us $n + 2$ equations in $n + 2$ unknowns (all the λ_i plus a and b). Combining and (6.26) gives

$$\sum \frac{1}{1+b^2}\,(a + bx_i - y_i) = 0 \tag{6.32}$$

and (6.27) using (6.28) gives

$$\sum \lambda_i x_i - b\lambda_i^2/2 = 0, \tag{6.33}$$

into which we substitute (6.31) and find

$$\sum \frac{1}{1+b^2}\,(ax_i + bx_i^2 - y_i x_i) - \sum \frac{b}{(1+b^2)^2}\,(a + bx_i - y_i)^2 = 0. \tag{6.34}$$

These two equations (6.32) and (6.34) relate the parameters a and b to the given data values x_i and y_i. We find the solution by solving the equations simultaneously. Noting that the denominator in (6.32) cannot be zero, we find

$$\sum (a + bx_i - y_i) = 0 \Rightarrow na + b\sum x_i = \sum y_i \tag{6.35}$$

or

$$a = \bar{y} - b\bar{x}, \tag{6.36}$$

where $\bar{x}$ and $\bar{y}$ are the mean data values. This expression for the intercept can now be substituted into (6.34) so we may solve for the slope. We multiply through by $(1 + b^2)^2$ and obtain

$$\sum (1+b^2)\left(\bar{y}x_i - b\bar{x}x_i + bx_i^2 - x_i y_i\right) - b\sum (\bar{y} - b\bar{x} + bx_i - y_i)^2 = 0 \tag{6.37}$$

REGRESSION

which simplify to

$$(1+b^2)\sum x_i(\bar{y}-y_i+b(x_i-\bar{x}))-b\sum(b(x_i-\bar{x})-(y_i-\bar{y}))^2=0, \tag{6.38}$$

We now introduce the residuals $u_i = x_i - \bar{x}$ and $v_i = y_i - \bar{y}$. Then

$$(1+b^2)\sum(u_i+\bar{x})(bu_i-v_i)-b\sum(bu_i-v_i)^2=0,$$

$$(1+b^2)\sum\left(bu_i^2-u_iv_i+\bar{x}bu_i-\bar{x}v_i\right)-b\sum\left(b^2u_i^2-2bu_iv_i+v_i^2\right)=0,$$

$$\sum\left(bu_i^2+b^3u_i^2-u_iv_i-b^2u_iv_i-b^3u_i^2+2b^2u_iv_i-bv_i^2\right)=0,$$

$$\sum\left(b^2u_iv_i+b(u_i^2-v_i^2)-u_iv_i\right)=0.$$

where we have used the properties $\sum u_i = \sum v_i = 0$. These steps finally give the solution for the slope as

$$b=\frac{\sum v_i^2-\sum u_i^2\pm\sqrt{(\sum u_i^2-\sum v_i^2)^2+4(\sum u_iv_i)^2}}{2\sum u_iv_i}. \tag{6.39}$$

This equation gives two solutions for the slope and we choose the one that minimizes E. The other solution maximizes E and makes an angle of 90 degrees with the optimal solution.

6.2.2 Reduced major axis (RMA) regression

In this alternative formulation of misfit we minimize the sum of the *areas* of the rectangles defined by Δx and Δy (Figure 6.4). Hence, the function to minimize is

$$E=\sum(X_i-x_i)(Y_i-y_i). \tag{6.40}$$

The constraints remain the same, i.e. $Y_i = a + bX_i$.

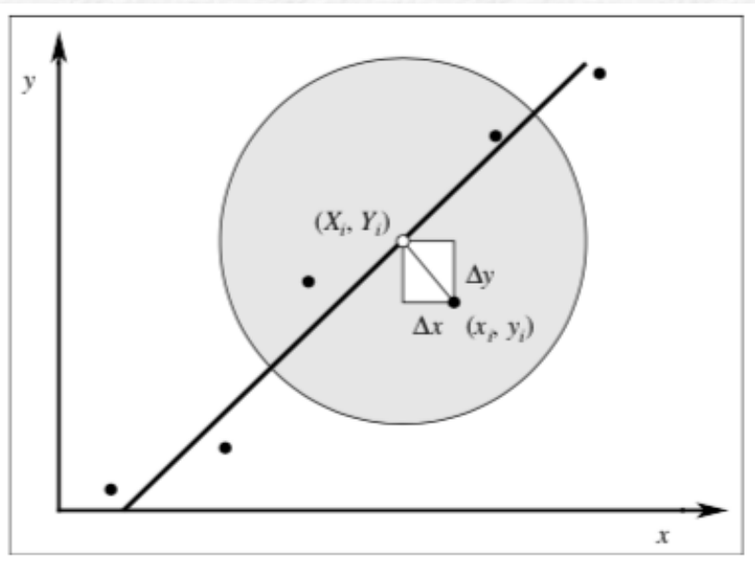

Figure 6.4: RMA regression minimizes the sum of the *areas* of the (white) rectangles defined by the data points (black circles) and their orthogonal projection points (white circles; only one is shown here) on the regression line.

REGRESSION

The Lagrange's multiplier method leads to a system of equations similar to those discussed in the previous section (6.22), but now we find the 2*n* + 2 equations

$$\frac{\partial F}{\partial X_i} = \frac{\partial}{\partial X_i}\sum_j^n (X_j - x_j)(Y_j - y_j) + \frac{\partial}{\partial X_i}\sum_j^n \lambda_j b X_j = 0 \Rightarrow Y_i - y_i + b\lambda_i = 0 \quad i = 1, n, \tag{6.41}$$

$$\frac{\partial F}{\partial Y_i} = \frac{\partial}{\partial Y_i}\sum_j^n (X_j - x_j)(Y_j - y_j) - \frac{\partial}{\partial Y_i}\sum_j^n \lambda_j Y_j = 0 \Rightarrow X_i - x_i - \lambda_i = 0 \quad i = 1, n, \tag{6.42}$$

$$\frac{\partial F}{\partial a} = \frac{\partial}{\partial a}\sum \lambda_i a = 0 \Rightarrow \sum \lambda_i = 0, \tag{6.43}$$

$$\frac{\partial F}{\partial b} = \frac{\partial}{\partial b}\sum \lambda_i b X_i = 0 \Rightarrow \sum \lambda_i X_i = 0. \tag{6.44}$$

We find

$$X_i - x_i - \lambda_i = 0 \Rightarrow X_i = x_i + \lambda_i, \tag{6.45}$$

$$Y_i - y_i + b\lambda_i = 0 \Rightarrow Y_i = y_i - b\lambda_i. \tag{6.46}$$

Substituting these values into the equation for the line (i.e., $Y_i = a + bX_i$) gives

$$y_i - b\lambda_i = a + b(x_i + \lambda_i) \tag{6.47}$$

or

$$\lambda_i = \frac{y_i - a - bx_i}{2b}. \tag{6.48}$$

Since $\sum \lambda_i = 0$, we find again

$$a = \bar{y} - b\bar{x}. \tag{6.49}$$

Substituting λ_i, X_i and *a* into (6.44) gives

$$\sum \left(x_i + \frac{y_i - \bar{y} + b\bar{x} - bx_i}{2b}\right)\left(\frac{y_i - \bar{y} + b\bar{x} - bx_i}{2b}\right) = 0.$$

REGRESSION

We let $u_i = x_i - \bar{x}$ and $v_i = y_i - \bar{y}$ as before and obtain

$$\sum (b(u_i + \bar{x}) + v_i - bu_i)(v_i - bu_i) = 0,$$

$$\sum (2b\bar{x} + bu_i + v_i)(v_i - bu_i) = 0,$$

$$\sum (2b\bar{x}v_i - 2b^2\bar{x}u_i + bu_iv_i - b^2u_i^2 + v_i^2 - bu_iv_i) = 0,$$

$$\sum v_i^2 - b^2 \sum u_i^2 = 0.$$

where we again have used the property that sums of u_i and v_i are zero. Finally, we obtain

$$b = \sqrt{\sum v_i^2 / \sum u_i^2} = \frac{s_y}{s_x}, \qquad (6.50)$$

i.e., the best slope equals the ratio of the *y* and *x* observations' separate standard deviations.

6.3 Robust Regression

In simple regression one assumes a relation of the type

$$y_i = a + bx_i + \varepsilon_i, \qquad (6.51)$$

in which x_i is called the *explanatory variable* or *regressor*, and y_i is the *response variable*. Again, we seek to estimate *a* and *b* (intercept and slope) from the data (x_i , y_i). It is commonly assumed that the deviations ε_i are normally distributed. Fortunately, in simple regression the observations (x_i, y_i) are 2-D so they can be plotted. It is always a good idea to do that first to see if any unusual features are present and to make sure the data are roughly linear. Applying a regression estimator to the data (x_i, y_i) will result in the two regression coefficients $\hat{a}$ and $\hat{b}$. They are not the *true* parameters *a* and *b*, but our "best" *estimates* of them.

REGRESSION

We can insert those into (6.51) and find the predicted estimate as

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \quad i = 1, n. \tag{6.52}$$

The residual is then the difference between the observed and estimated values, yielding

$$e_i = y_i - \hat{y}_i, \quad i = 1, n. \tag{6.53}$$

Note that there is a difference between e_i (the misfit) and ε_i (the deviation), because $\varepsilon_i = y_i - a - bx_i$ are evaluated with the true unknown a, b. We can compute e_i, but not ε_i.

The most popular regression estimator dates back to the early 1800's (to our old friends Gauss and Legendre) and is called the "least-squares" (LS) method since it seeks to minimize

$$E = \sum_{i=1}^{n} e_i^2. \tag{6.54}$$

The rationale was to make the residuals very small. Gauss preferred the least-squares criterion to other objective functions because in this way the regression coefficients could be computed explicitly from the data (no computers back in the day, at least not mechanical or electronic ones). Later, Gauss introduced the normal, or Gaussian, distribution for which least-squares is optimal.

More recently, people have realized that real data often do not satisfy the Gaussian assumption and this "failure to comply" may have a dramatic effect on the LS results. In Figure 6.5 we have five points that lie almost on a straight line. Here, the *LS* line fits the data very well. However, let us see what happens if we get a wrong value for y_4 because of a recording or copying error:

The bad point y_4 is called an *outlier in the y-direction*, and it has a dramatic effect on the *LS* line which now is tilted away from the trend of the remaining data. Such outliers have received the most attention because most experiments are set up to expect errors in y only. However, in observational studies it often happens that outliers occur in the x_i data as well.

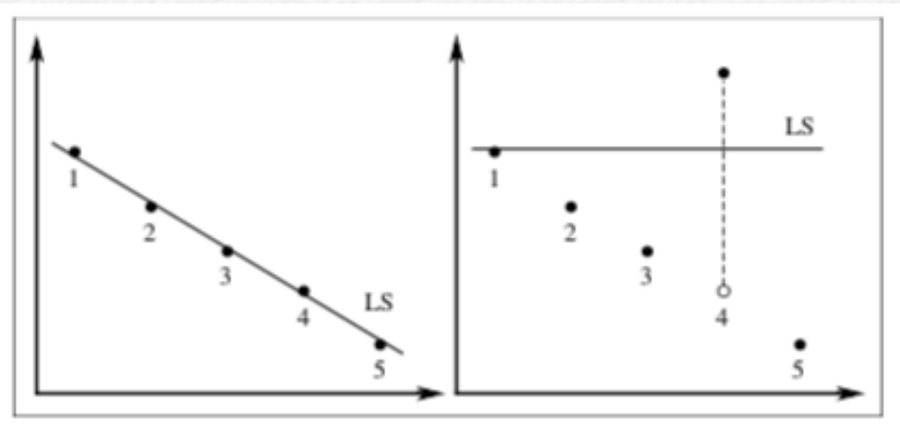

Figure 6.5: Pitfalls of least-squares regression, part I. An outlying point in the y-direction will affect the regression line considerably.

REGRESSION

Figure 6.6 illustrates the effect of an outlier in the *x*-direction. It has an even more dramatic effect on *LS* since it now is almost perpendicular to the actual trend. Because this single point has such a large influence we denote it as a *leverage* point. This is because the residual e_i (measured in the *y*-direction) is enormous with regard to the original *LS* fit. The second *LS* fit reduces this enormous error at the expense of increasing the errors at all other points. In general, we call the *k*'th point a leverage point if x_k lies far from the bulk of the x_i. Note that this definition does not take y_i into account. For instance, Figure 6.7 shows a "good" leverage point since it lies on the linear trend set by the majority of the data. Thus, a leverage point only refers to its *potential* for influencing the coefficients $\hat{a}$, $\hat{b}$. When a point (x_i, y_i) deviates from the linear relation of the majority it is called a *regression outlier*, taking into account both x_i and y_i simultaneously. As such, it is a vertical outlier or a bad leverage point. It is often stated that regression outliers can be detected by looking at the *LS* residuals. This is not always true. The bad leverage point 1 has tilted the line so much that its residual is very small. Consequently, a rule that says "delete points with highest *LS* residual" would first find points number 2 and 5, thereby deleting the good points first.

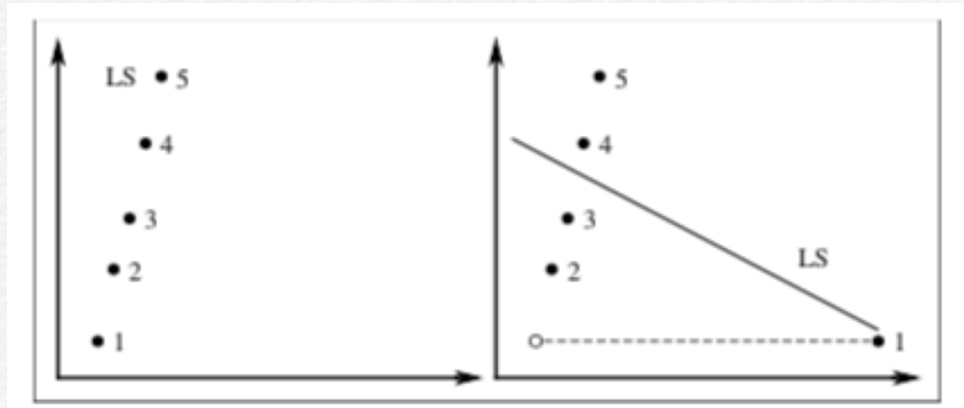

Figure 6.6: Pitfalls of least-squares regression, part II. Here, an outlying point in the *x*-direction can have a huge effect on the regression line

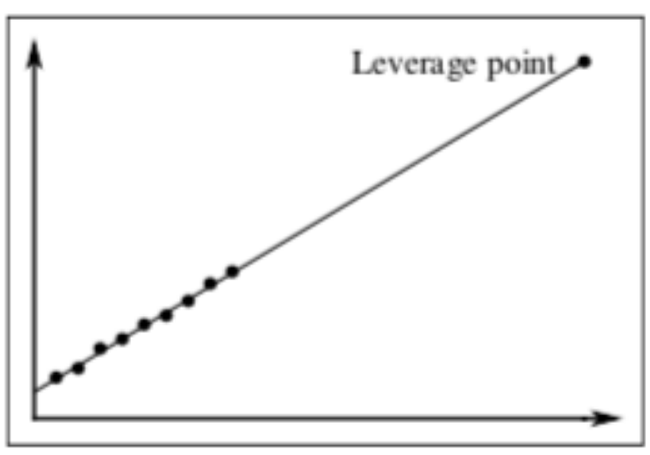

Figure 6.7: The effect of leverage points in regression can be enormous, whether the data point is a valid observation or a bad outlier.

Of course, in simple $x - y$ regression we have the benefit of being able to plot the data so this is not often a problem, except when the number of data sets and points are large.

From the simple examples we have just seen, we find that the breakdown point for *LS* regression is merely 1/*n* since one point is enough to ruin the day — analogous to the breakdown point for the mean, which was also based on *LS*.

A first step toward a more robust regression was taken more than 100 years ago when Edgeworth suggested that one could instead minimize

$$E = \sum_{i=1}^{n} |e_i|, \qquad (6.55)$$

REGRESSION

which we will call L_1 regression. Unfortunately, while L_1 regression is robust with respect to outliers in *y*, it offers no protection toward bad leverage points. Thus, the breakdown point is still only 1/*n*.

While there are many methods that offer a higher breakdown point than L_1 and L_2, we will concentrate our presentation on one particular method. Again, let us look at the *LS* formulation:

$$\underset{\hat{a},\hat{b}}{\text{minimize}} \quad E = \sum_{i=1}^{n} e_i^2.$$

At first glance, you would think that a better name for *LS* would be least *sum* of squares. Apparently, few people objected to removing the word "sum" as if the only sensible thing to do with *n* numbers would be to add them. Adding the e_i^2 is the same as using their mean (dividing by *n* does not affect the minimization). Why not replace the mean (i.e., the sum) by a median, which we know is very robust? This yields the "least median of squares" (LMS) criterion:

$$\underset{\hat{a},\hat{b}}{\text{minimize}} \quad \text{median} \ e_i^2. \tag{6.56}$$

It turns out the LMS fit is very robust with respect to outliers in *y* as well as in *x*. Its breakdown point is 50%, which is the most we can ask for. If more than half your data are bad then you cannot tell which part is good unless you have additional information. Figure 6.8 shows what we get using this method on the data that made the LS technique fail so badly.

The LMS line also has an intuitive geometric interpretation: it lies at the center of the narrowest strip that covers half of the data points. By half of the points we mean *n*/2 + 1, and the thickness of the strip is measured in the vertical direction (i.e., Figure 6.9) .

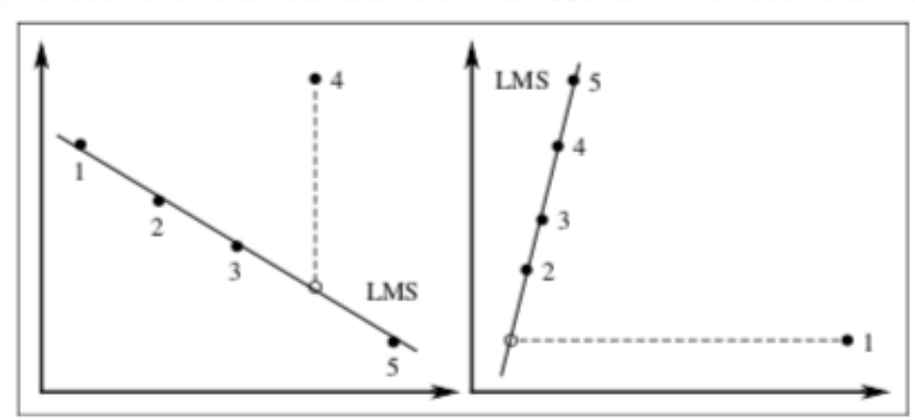

Figure 6.8: Robust regression, such as LMS, is very tolerant of outlying points in both the *x* and *y* directions.

REGRESSION

An example of LMS regression comes from astronomy. Astronomers often look for a linear relationship between the logarithm of the light intensity and the logarithm of the surface temperature of stars. A scatter plot of observed quantities may look like Figure 6.10. Here, the LMS line defines what is known as the *main sequence*; the four outlying stars turned out to be red giants that do not follow the general trend. The *LS* fit produces a rather worthless compromise solution. Here, the outliers are not so much errors as *contamination from a different population*.

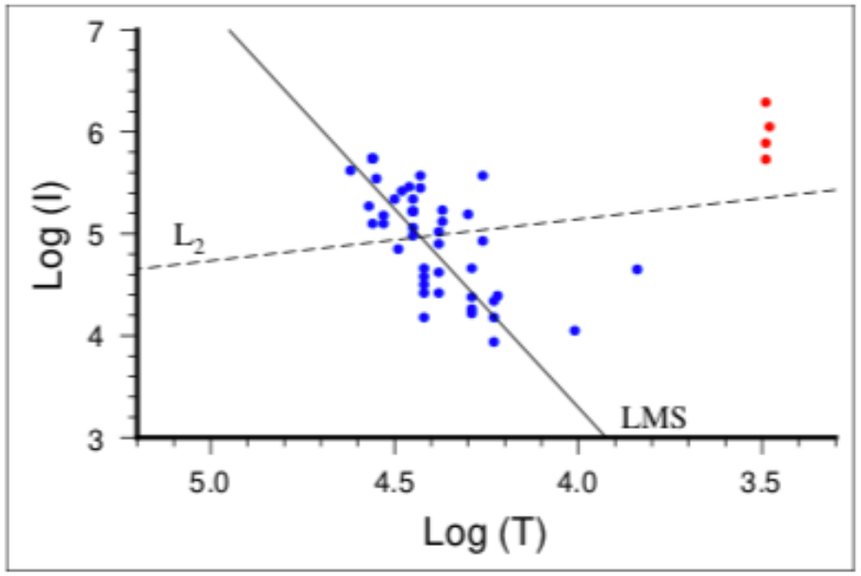

Figure 6.10: Example of robust regression in astronomy. We see a Hertzsprung-Russell diagram of the star cluster CYG OB1 with the least squares (dashed line) and LMS (solid line) fit. The red giants are distorting the LS fit. Data taken from Rousseeuw, P. J., and A. M. Leroy (1987), *Robust regression and outlier detection*, 329 pp., John Wiley and Sons, New York.

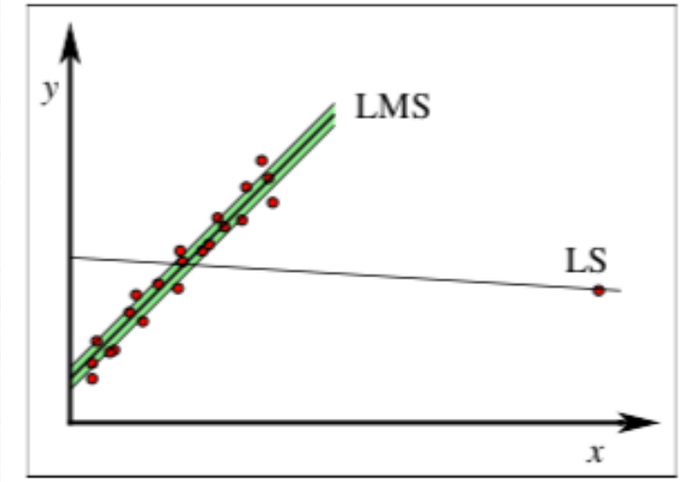

Figure 6.9: Geometrical meaning of the LMS regression: the narrowest strip that covers half the data points.

6.3.1 How to estimate LMS regression

We can rewrite the minimization criterion as follows:

$$\underset{\hat{a},\hat{b}}{\text{minimize}}\{\text{median } e_i^2\} = \underset{\hat{a},\hat{b}}{\text{minimize}}\{\text{median } ((y_i - \hat{b}x_i) - \hat{a})^2\} \tag{6.57}$$

in the form

$$\underset{\hat{b}}{\text{minimize}} \left\{ \underset{\hat{a}}{\text{minimize median}} \; ((y_i - \hat{b}x_i) - \hat{a})^2 \right\} . \tag{6.58}$$

REGRESSION

We will treat the two minimizations here separately. The innermost minimization is the easy part, because for any given $\hat{b}$ it becomes essentially a 1-D problem, i.e., we want to find the value for $\hat{a}$ that minimizes the median

$$\underset{\hat{a}}{\text{minimize}} \left\{ \text{median } (u_i - \hat{a})^2 \right\}, \tag{6.59}$$

Where u_i is calculated as $u_i = y_i - \hat{b}\, x_i$ (remember, we assumed that $\hat{b}$ was given). This minimization problem is the same one we found earlier to give a good estimate of the mode. Thus, this operation finds the mode of the u_i data set. We therefore need to find the $\hat{b}$ for which

$$E(\hat{b}) = \text{median } [(y_i - \hat{b}x_i) - \hat{a}]^2 \tag{6.60}$$

is minimal. This is simply the minimization of a 1-D function $E(\hat{b})$ which is continuous but not everywhere differentiable.

To find this minimum we make the observation that the slope in the $x - y$ plane must be in the ±90° range (when expressed as an angle β, with $b = \tan\beta$). We then simply perform a search for the optimal angle. Starting with β = −90 °, we form the resulting u_i and solve the 1-D minimization problem for $\hat{a}$, i.e., finding the LMS mode estimate $\hat{a}$. We now increment the angle β by $d\beta$ to, say, −89 °, and repeat the process. At each step we keep track of what $E(\beta)$ is and repeat these steps for all angles through β = 90 °.

Having found the slope $\hat{b}$ that gave the smallest misfit we may improve on this estimate by using a smaller step size $d\beta$ in this region to pinpoint the best choice for $\hat{b}$. A plot of $E(\beta)$, shown in Figure 6.11, is very useful since it may tell us how unique the LMS regressions are: if more than one minimum is found they may indicate a possible ambiguity as there may be two or more lines that fit the data equally well.

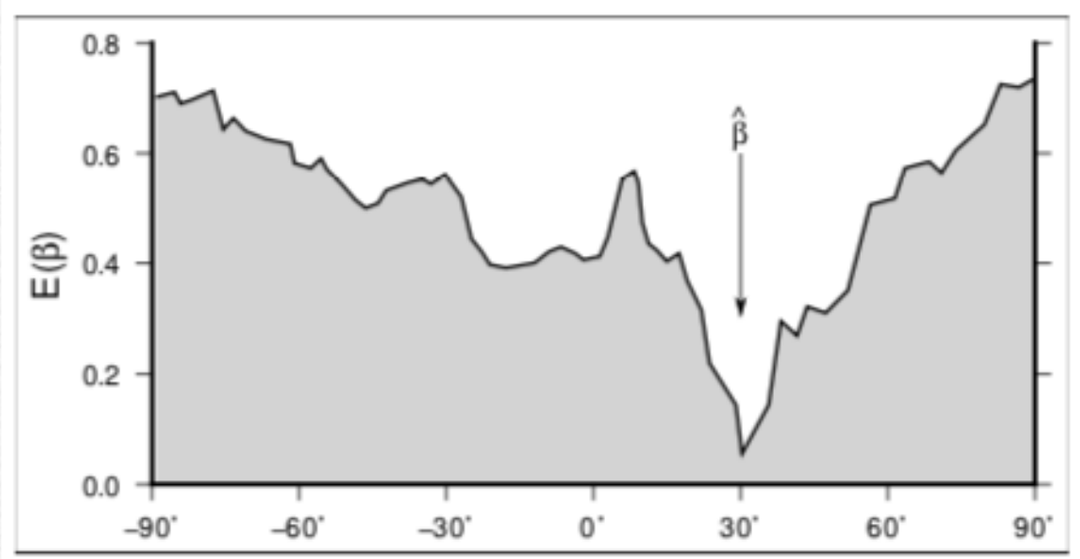

Figure 6.11: Determining the best regression slope, $\hat{b} = \tan\hat{\beta}$. The misfit function is not smooth but usually has a minimum for the optimal slope. It is also likely to reveal several local minima that could trick a simpler search

REGRESSION

We will elaborate on the breakdown point for simple regression and illustrate it with a simple experiment. Consider a data set that contains 100 good data points that exhibit a strong linear relation, computed from

$$y_i = 1.0x_i + 2.0 + \varepsilon_i \quad 1 \le x_i \le 4, \tag{6.61}$$

with ε_i normally distributed, with $\mu = 0$ and $\sigma = 0.2$.

Any regression technique, including L_2, will of course recover estimates of the slope and intercept that are very close to the true values 1 and 2 (Figure 6.12a). Then, we will start to contaminate the data by replacing "good" points with bad ones, the latter coming from a bivariate normal distribution with $\mu = (7, 2)$ and $\sigma_r = 0.5$. We systematically substitute one bad point for a good point and recompute the regression parameters after each step. What we find is that the LS estimate goes bad right away (Figure 6.12b). The bad points are basically bad leverage points, which we know the LS process cannot handle. We keep track of the slope estimate after each substitution and graph the results in Figure 6.13.

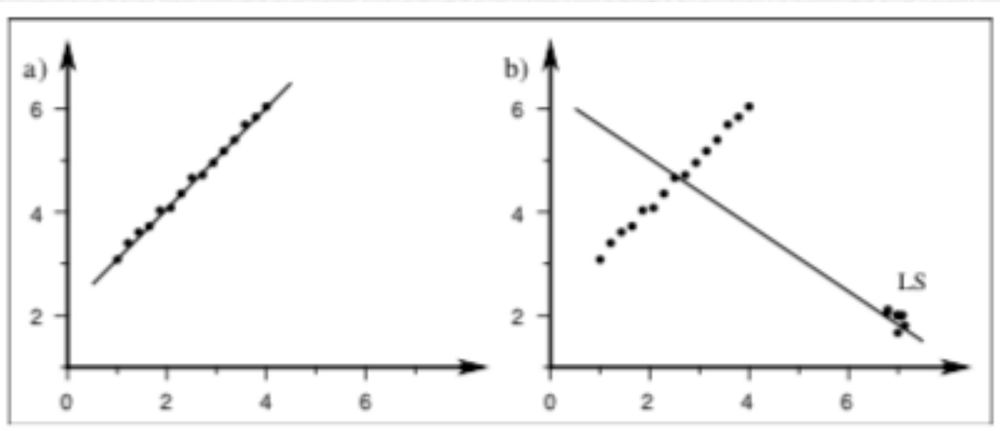

Figure 6.12: (a) Best LS fit to synthetic data set computed from a linear model with Gaussian noise. (b) Synthetic data set computed from a linear model with Gaussian noise, but now contaminated by points from another (bivariate) distribution centered on (7,2). The LS line is pulled way off by these bad leverage points. We can then plot the slope value as a function of the percentage of contamination.

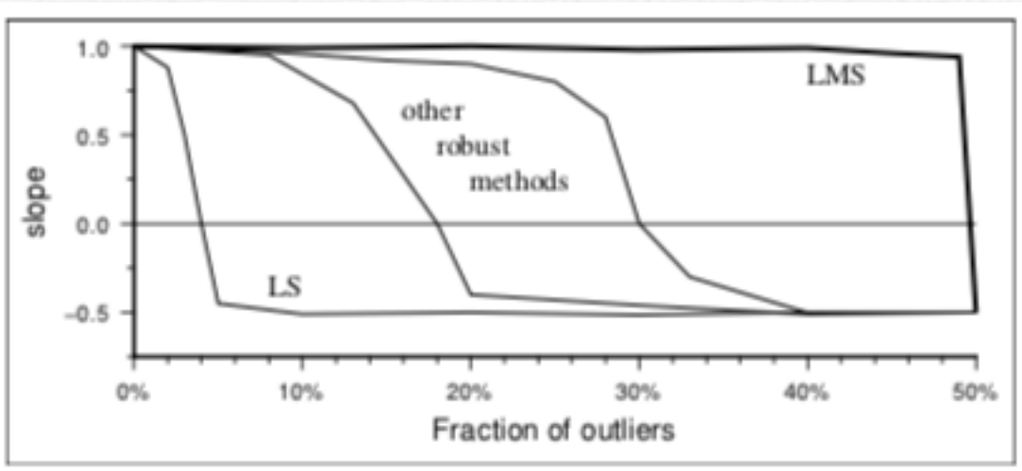

Figure 6.13: Breakdown plot for several regressors. The LMS only breaks down when 50% of the data are outlying. LS breaks down immediately because every point matters.

We call the percentage at which the slope starts to deviate significantly from the true value the *breakdown point*. The clear winner of this test is the LMS regression which keeps finding the correct trend until half the points have been replaced.

REGRESSION

We should add that while LMS always have a breakdown point of 50%, it is found that the effect of the outliers often depends on the quality of the good data. In cases where the good data exhibit a strong correlation the outliers do less damage than in a case where there is little or no correlation (Figure 6.14). Of course, when the correlation among the good data is minimal it is probably not very useful to insist that the data really exhibit a linear trend in the first place.

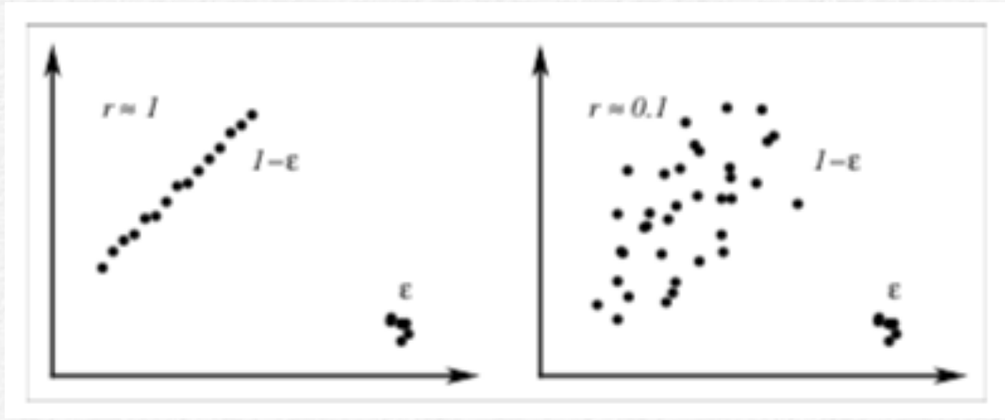

Figure 6.14: Two data sets with the same degree of contamination. However, one exhibit a much stronger correlation between the "good" points than the other.

6.3.2 How to find LMS 1-D Location (single mode)

When we discussed estimates of central location we briefly mentioned that the value $\hat{x}$ that minimized

$$\underset{\hat{x}}{\text{minimize}} \quad \{\text{median } (x_i - \hat{x})^2\} \tag{6.62}$$

was called the LMS location and that it was a good approximation to the *mode*, but how do we determine the LMS estimate? It turns out that it is rather simple. The following recipe will do:

1. Sort the data into ascending order.
2. Determine the shortest half of the sample, i.e., find the value for i that yields the smallest of the differences $(x_{h+i} - x_i)$ with $h = n/2+1$.
3. The LMS estimate is the midpoint of the shortest half:

$$\text{LMS} = \frac{1}{2}(x_{h+i} + x_i). \tag{6.63}$$

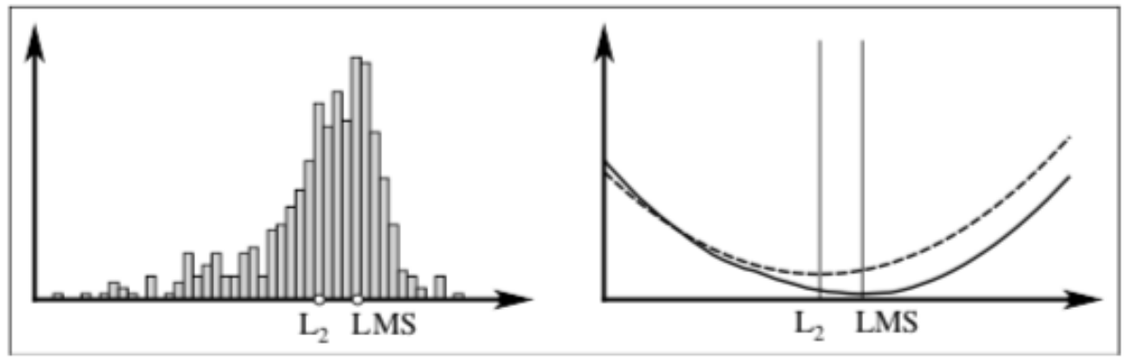

Figure 6.15: LMS defines the mode while L2 defines the mean; both are the locations where their respective objective functions are minimized.

You can empirically try this out by letting $\hat{x}$ take on all values within the data range, compute the median of all $(x_i - \hat{x})^2$ values, and graph the curve and find its minimum (Figure 6.15).

REGRESSION

6.3.3 Making LMS "analytical" — finding outliers

There is one problem with using the robust LMS parameters: the method is not *analytical* so it does not lend itself easily to standardized statistical tests. We will look into how we can overcome this obstacle.

The main problem comes from the fact that the outliers cause L_2 estimates to become unreliable. We will avoid this problem altogether by using the best from both worlds (L_2 *and* LMS): We will use robust LMS techniques to find the best parameters in the regression model and then use the robust residuals to detect outliers. Finally, we recompute weighted L_2 parameters with outliers given zero weights and other points given unit weights. These L_2 estimates now represent only the "good" portion of the data and confidence limits and statistical tests may be based on the behavior of these good values. We call this technique "Re-weighted Least Squares" (RLS).

First, we need a robust scale estimate for the residuals that will make them nondimensional. It is customary to choose the preliminary scale estimate

$$s^0 = 1.4826\left(1+\frac{5}{n-2}\right)\sqrt{\text{median } e_i^2}, \tag{6.64}$$

where $e_i = y_i - \hat{y}_i$ are the residuals (i.e., misfits) at each point. With this scale estimate we can evaluate the normalized residuals as

$$z_i = e_i/s^0. \tag{6.65}$$

Now use these numbers to design the weights:

$$w_i = \begin{cases} 1, & |z_i| \le 2.5 \\ 0, & \text{otherwise} \end{cases}. \tag{6.66}$$

The final LMS regression scale estimate is then given by

$$s^* = \sqrt{\sum w_i e_i^2/(\sum w_i - 2)}. \tag{6.67}$$

The RLS regression parameters are therefore obtained by minimizing the weighted, squared residuals:

$$\min E = \sum_{i=1}^{n} w_i e_i^2. \tag{6.68}$$

REGRESSION

This is simply the L_2 solution when only the good data are used. As shown when we discussed the weighted L_2 regression problem, this technique will provide confidence intervals on both the slope and intercept and it also allows us to use the χ^2-test to check whether the RLS fit is significant or not.

The strength of the linear relationship can again be measured by the (Pearson) correlation coefficient. The LMS estimate of correlation is now given by

$$r = \sqrt{1 - \left(\frac{\text{median } |e_i|}{\text{MAD } y_i} \right)^2} \tag{6.69}$$

with

$$\text{MAD } y_i = \text{median } |y_i - \tilde{y}|. \tag{6.70}$$

Compare this to the L_2 case, where

$$r = \frac{s_{xy}}{s_x s_y} = \sqrt{1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}} = \sqrt{1 - \left(\frac{\bar{e}}{s_y} \right)^2}. \tag{6.71}$$

This comparison shows that the robust estimates for “average” and “scale” have replaced the traditional least-squares estimates in the expression for correlation.