

# Fitting a curve to data: least squares regression

Brad Biggerstaff  
Centers for Disease Control and Prevention  
National Center for Infectious Diseases  
Division of Vector-Borne Infectious Diseases  
P.O. Box 2087, Fort Collins, Colorado 80522-2087, USA  
(970) 221-6473 / bbiggerstaff@cdc.gov

May 1, 2000

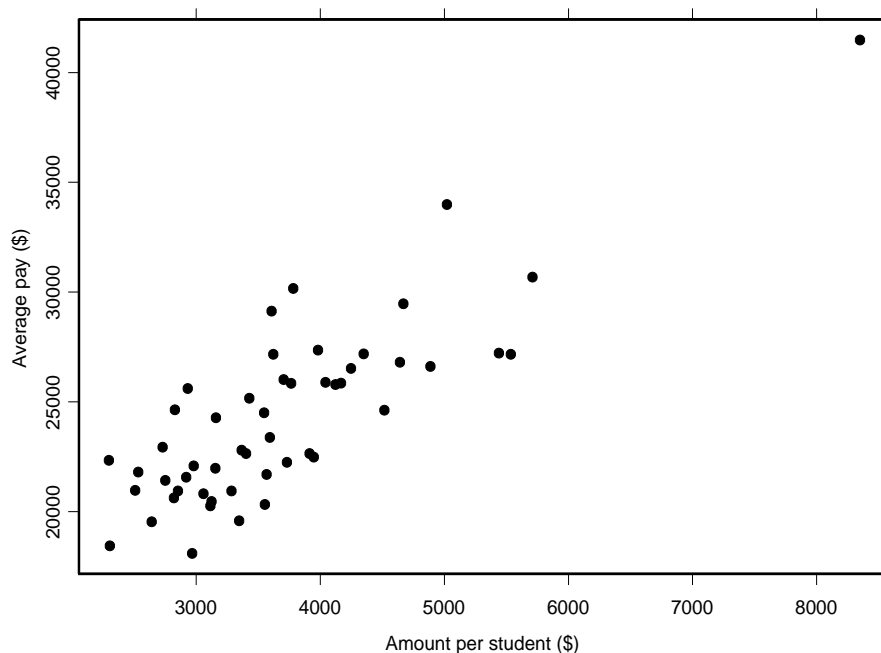
## Abstract

A sometimes useful way to summarize a set of points  $(x_i, y_i)$ , for  $i = 1, 2, \dots, n$ , is to derive a curve—often a line—relating the  $x$ s to the  $y$ s. In this note I describe the method of the sum of least squared deviations or more simply *the method of least squares* for curve fitting. Along the way I also introduce *least absolute deviations* regression. A worked example of linear least squares regression is given using 1985 data relating the average expenditure per public school student by state and the average pay for the state's teachers. A more difficult, non-linear example of least squares methodology is given, along with an S-Plus routine for performing the computations. General comments and some references for further reading close the paper.

## 1 Introduction

Let  $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$  be a set of points in the plane. For example, consider the 51 points plotted in Figure 1, where the horizontal axis marks the average amount each state and the District of Columbia spent on each student in public school in 1985 (the  $x$ s) and the vertical axis marks the average pay for the teachers in the state for the same year (the  $y$ s). Taken off the internet from the Data & Story Library (DASL) in the on-line Statistics Library (StatLib, <http://lib.stat.cmu.edu/>), the data were originally quoted from the National Education Association and reported on November 7, 1985, in the *Albuquerque Tribune*. The data are reproduced in Table 2 in the appendix. One

Figure 1: Average teacher salary by average amount spent per student for each state, 1985



can see a general pattern fairly easily: increasing expenditure on students is basically associated with increasing expenditure on teachers. Based on this information, one may ask some sensible questions. A couple are: What would be the average pay for teachers if a state were to spend, on average, \$3,500 per student? How much does the average teacher salary rise with \$1,000 increase in average expenditure per student?

A useful way to begin to answer questions like these is to summarize the obvious relationship seen in the graph with a line which relates  $y =$  average teacher pay to  $x =$  average amount spent on a student:

$$y = \beta_0 + \beta_1 x, \tag{1}$$

where the *parameters*  $\beta_0$  and  $\beta_1$ , when replaced by appropriate numbers, specify the particular line to represent the points. Now, there are many ways one might obtain numerical values for  $\beta_0$  and  $\beta_1$  for any particular problem. Perhaps the easiest is just to grab a ruler and a pencil, arrange the ruler to go through the points as best one can by eye, then compute a value for the slope  $\beta_1$  by the change in the two  $y$  values for two fixed  $x$  values, and then compute a value for the intercept  $\beta_0$  by extrapolating the line all the way down to  $x = 0$ . (Or, once the slope is known, to use the formula in Equation (1) and some algebra to solve for  $\beta_0$  for some fixed pair of points on the line.) There are other, more mathematical ways, however, to go about “fitting” a line, a term used simply to mean finding numerical values for  $\beta_0$  and  $\beta_1$  for a particular set of points. In the next section of this essay, I introduce two different methods of fitting curves to data points, one more intuitive but also more difficult, the other a more historically successful way often called the *method of least squares*. A term often used for this kind of curve fitting is *regression*.

Before that, I want to point out that what I have said so far applies more generally than just fitting a line to points, in that the same ideas can be used for more complicated curves than a line. One may see, for example, a parabolic shape in the graph of the points, and one may then want fit to a parabola

$$y = \beta_0 + \beta_1 x + \beta_2 x^2. \quad (2)$$

Or, for that matter, one may decide to fit any polynomial of order  $p$  (when  $p < n$ ):

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_p x^p = \sum_{k=0}^p \beta_k x^k.$$

In fact, the same method works in principle for many functions, for example even the more complicated function

$$y = e^{-e^{\beta_0 + \beta_1 x}}, \quad (3)$$

an example for which we visit below in Section 4.

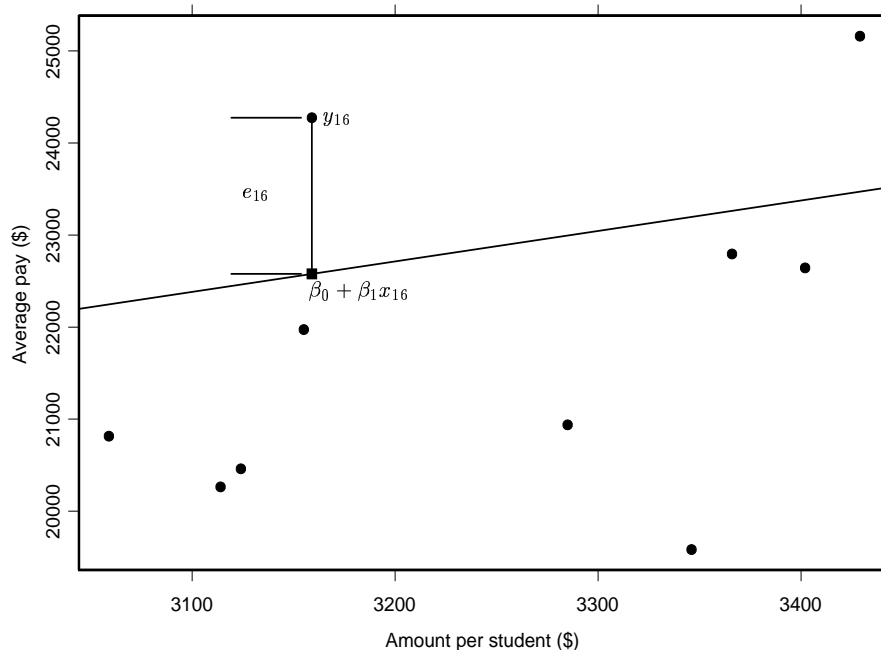
## 2 Two methods of fitting

The first thing to acknowledge is that we cannot expect, with a line, to go through all the points in Figure 1, though we might fortunately have our summary line go through some of the points. Because of this, it will be useful to represent the points  $y_i$ , for  $i = 1, 2, \dots, n$ , as

$$y_i = \beta_0 + \beta_1 x_i + e_i, \quad (4)$$

where for each  $i$  the term  $e_i$  represents the *error* between our line and the actual value of  $y_i$  along the  $y$ -axis. Figure 2 shows this decomposition graphically for the 16th point (for Indiana) in the teacher pay/student expenditure dataset introduced in the first section.

Figure 2: Graphic display of the breakdown in Equation (4)



## 2.1 Least absolute deviation

Now, if we can somehow fit our line so that the values for the errors  $e_i$  are all near 0, then we know our line will represent the points well. A simple idea keeping this in mind is to try to find values for  $\beta_0$  and  $\beta_1$  which minimize all the distances from the data points to the line—this is basically what we do when we fit a line “by eye.” The distances from the line for each point are just  $|e_i| = |y_i - (\beta_0 + \beta_1 x_i)|$ , taking note of Equation (4) and Figure 2. To minimize each of these distances individually is hard, because the right minimum for one particular point is of course 0, but this can make the distances for others very large. A way around this is to try to find values for  $\beta_0$  and  $\beta_1$  which minimize the *average* of these distances,

$$A(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n |e_i| = \frac{1}{n} \sum_{i=1}^n |y_i - (\beta_0 + \beta_1 x_i)|, \quad (5)$$

which is equivalent to minimizing the *sum* of the distances. This method is often called the method of *least absolute deviations* or *least mean absolute deviation* (“mean” is just another word for “average”). Unfortunately, this problem turns out to be extremely hard to solve, in general requiring complex computational algorithms and fast computers (or lots of time). Historically, then, this method—though perhaps the most obvious one—was all but abandoned for the next idea.

## 2.2 Least squares

Equation (5) arose because it seemed the natural thing to do, but because the mathematical problem turned out to be too hard, some clever mathematicians in the 1700s used a then-new tool called “calculus” to arrive at a different solution. Using similar reasoning—wanting somehow to minimize the errors  $e_i$ —rather than trying to minimize the average of the distances  $A(\beta_0, \beta_1)$  in Equation (5), use the *squared* distances  $e_i^2$  and try to find values for  $\beta_0$  and  $\beta_1$  which minimize the average of these:

$$S(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n e_i^2 = \frac{1}{n} \sum_{i=1}^n [y_i - (\beta_0 + \beta_1 x_i)]^2. \quad (6)$$

Geometrically, this is not as intuitive, but, as I now show, we can readily obtain a solution to this minimization problem. (More on this seeming sell-out is discussed in Section 5.)

Viewing the data points  $(x_i, y_i)$ , for  $i = 1, 2, \dots, n$ , as fixed values, the function  $S(\beta_0, \beta_1)$  in Equation (6) is fairly easy to differentiate as a function of the  $\beta$ s. So, for each  $\beta$  in turn,

$$\begin{aligned} \frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta_0} [y_i - (\beta_0 + \beta_1 x_i)]^2 \\ &= \frac{1}{n} \sum_{i=1}^n 2[y_i - (\beta_0 + \beta_1 x_i)](-1) \\ &= -\frac{2}{n} \left( \sum_{i=1}^n y_i - n\beta_0 + \beta_1 \sum_{i=1}^n x_i \right) \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta_1} [y_i - (\beta_0 + \beta_1 x_i)]^2 \\ &= \frac{1}{n} \sum_{i=1}^n 2[y_i - (\beta_0 + \beta_1 x_i)](-x_i) \\ &= -\frac{2}{n} \left( \sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \right). \end{aligned}$$

To find values of  $\beta_0$  and  $\beta_1$  which minimize  $S(\beta_0, \beta_1)$ , set these two expressions equal to 0 and solve the resulting system (two equations with two unknowns) for  $\beta_0$  and  $\beta_1$ :

$$\begin{cases} \sum_i y_i &= n\beta_0 + \beta_1 \sum_i x_i \\ \sum_i x_i y_i &= \beta_0 \sum_i x_i + \beta_1 \sum_i x_i^2 \end{cases} \quad (7)$$

Using some algebra, from the first equation in (7) we get

$$\hat{\beta}_0 = \frac{1}{n} \left( \sum_i y_i - \beta_1 \sum_i x_i \right),$$

where I have used the convention of putting a  $\hat{\phantom{\beta}}$  over the solution value for the parameter  $\beta_0$ . Plugging  $\hat{\beta}_0$  in for  $\beta_0$  in the second equation in (7), solve for  $\beta_1$  by the sequence of manipulations

$$\begin{aligned} n \sum_i x_i y_i &= \left( \sum_i y_i - \beta_1 \sum_i x_i \right) \sum_i x_i + n \beta_1 \sum_i x_i^2 \\ n \sum_i x_i y_i - \left( \sum_i x_i \right) \left( \sum_i y_i \right) &= \beta_1 \left[ n \sum_i x_i^2 - \left( \sum_i x_i \right)^2 \right] \\ \hat{\beta}_1 &= \frac{n \sum_i x_i y_i - \left( \sum_i x_i \right) \left( \sum_i y_i \right)}{n \sum_i x_i^2 - \left( \sum_i x_i \right)^2}. \end{aligned}$$

The two expressions

$$\begin{cases} \hat{\beta}_0 &= \frac{1}{n} \left( \sum_i y_i - \hat{\beta}_1 \sum_i x_i \right) \\ \hat{\beta}_1 &= \frac{n \sum_i x_i y_i - \left( \sum_i x_i \right) \left( \sum_i y_i \right)}{n \sum_i x_i^2 - \left( \sum_i x_i \right)^2} \end{cases}$$

therefore constitute the solution to the minimization problem, that is, these values for  $\beta_0$  and  $\beta_1$  minimize the sum of the squared deviations  $S(\beta_0, \beta_1)$ . Further, they are easy to compute: all one needs is the sums of the  $y_i$ s and  $x_i$ s, and the sum of the squared  $x_i$ s.

A representation for the solution  $\hat{\beta}_0$  that is useful for interpretation is to use the average or mean values  $\bar{y} = \frac{1}{n} \sum_i y_i$  and  $\bar{x} = \frac{1}{n} \sum_i x_i$ :

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

We gain from this the immediate observation that the fitted line passes exactly through the mean values of the  $x_i$ s and  $y_i$ s:  $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$ . Further, algebraic manipulations of  $\hat{\beta}_1$  using  $\bar{x}$  and  $\bar{y}$  yield

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}.$$

The interpretation of this representation is somewhat more complicated, so I refrain from discussion it here.

(Note that this process, as I have computed so far, does not actually guarantee that these values for the  $\hat{\beta}$ s minimize  $S(\beta_0, \beta_1)$ ; it only assures us that these values give a local extremum. Finding second derivatives to show positive concavity, and showing somehow that the extremum is global is really required to prove that the solutions given are, in fact, the global minimizing values for  $S(\beta_0, \beta_1)$ . Fortunately for us all, someone else did that a long time ago, so I will leave that part out.)

This procedure, that of finding values of the parameters for the curve which minimize the sum of the squared deviations—the method of least squares—is in principle no more difficult than this. However, if the resulting system of equations is hard to solve for the parameter values, formulas for the solutions may be impossible to derive. In such cases one can employ computers to find for the solutions, and I give an example of this in Section 4.

### 3 Least squares fit for the Teacher Pay data

Let's now put all that work to use. Computations for the teacher pay data given in Table 2 in the appendix and graphed in Figure 1 yield the summaries in Table 1. Using these results, the values for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are

Table 1: Summary statistics for Teacher Pay data

$n$	$\sum_i x_i$	$\sum_i y_i$	$\sum_i x_i^2$	$\sum_i x_i y_i$	$\bar{x}$	$\bar{y}$
51	188,527	1,242,167	752,536,385	4,775,791,992	3,696.61	24,356.22

$$\begin{aligned}\hat{\beta}_1 &= \frac{51 \cdot 4775791992 - 188527 \cdot 1262167}{51 \cdot 752536385 - (188527)^2} \\ &= 9383373583/2836925906 \\ &= 3.307585\end{aligned}$$

$$\begin{aligned}\hat{\beta}_0 &= 24356.22 - 3.307585 \cdot 3696.61 \\ &= 12129.37.\end{aligned}$$

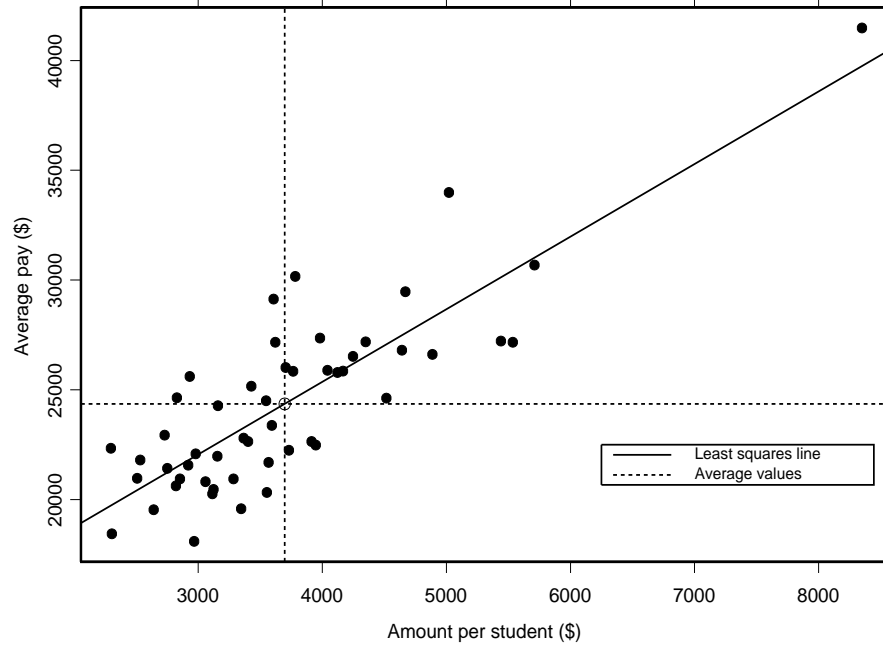
The least squares fitted line  $y = 12129.37 + 3.31x$  is shown as a solid line with the original data in Figure 3; this line is also called the fitted *regression line* of  $y$  on  $x$ . Also included in the graph are dashed lines representing the average values for each variable, and we see that the fitted line does, indeed, pass through the point  $(\bar{x}, \bar{y}) = (3,696.61, 24,356.22)$ , as we showed above that it should.

In the Introduction, I asked two questions relating to the teacher pay data, and I also suggested that fitting the line would help us get at the answers. The first question was, “What would the average pay for teachers be if a state were to spend on average \$3,500 per student?” We can answer this by using the fitted line, evaluating it at  $x = 3,500$ , giving  $12,129.37 + 3.31(3,500) = \$23,705.92$ . So a state which would spend an average of \$3,500 per student would pay its teachers, on average, \$23,705.92. (Recall that this is for 1985.)

The fitted line can also help us to answer the other question, “How much does the average teacher salary rise with \$1,000 increase in average expenditure per student?” Recalling that the slope of a line gives the unit change in  $y$  for a unit change in  $x$ , multiplying the slope  $\hat{\beta}_1 = 3.307585$  of the fitted line by 1,000 gives the associated change in  $y$  of 3,307.58. So a \$1,000 increase in average expenditure per student results in a \$3,307.58 change in average teacher pay.

### 4 A more difficult least squares problem

I said above that the least squares method works, in principle, for curves other than a line. Look at the data in Figure 4.

Figure 3: Least squares fit  $y = 12,129.37 + 3.31x$  to the Teacher Pay data

The function of interest, the one I want to fit to these points, is

$$y = e^{-e^{\beta_0 + \beta_1 x}}, \quad (8)$$

and the associated sum of squared deviations function for this curve is

$$S(\beta_0, \beta_1) = \frac{1}{n} \sum_{i=1}^n (y_i - e^{-e^{\beta_0 + \beta_1 x_i}})^2.$$

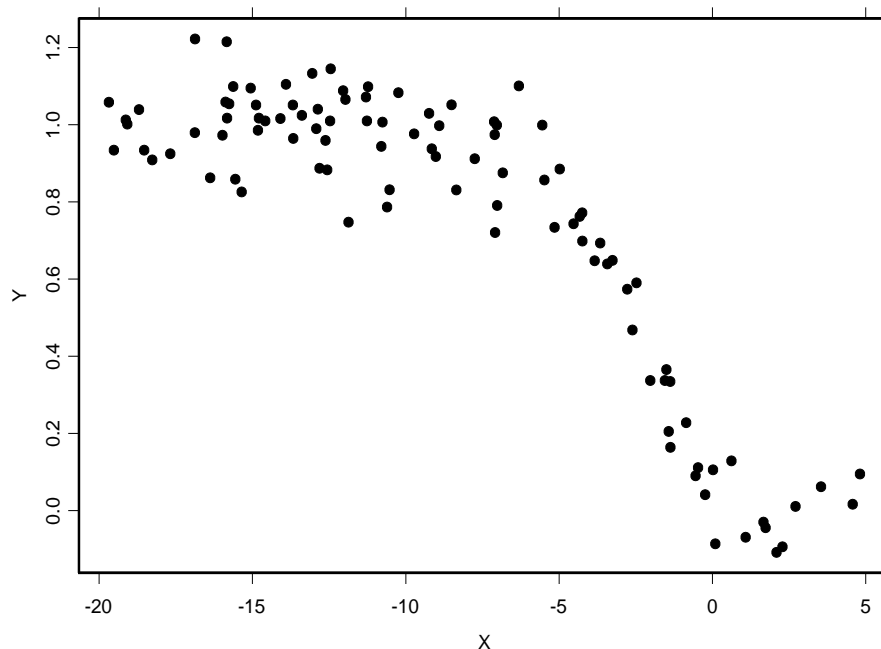
To find  $\beta_0$  and  $\beta_1$  which minimize  $S(\beta_0, \beta_1)$ , take derivatives with respect to the  $\beta$ s (a useful trick for avoiding getting buried in the chain rule when taking derivatives like this is given in the appendix):

$$\begin{aligned} \frac{\partial}{\partial \beta_0} S(\beta_0, \beta_1) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta_0} (y_i - e^{-e^{\beta_0 + \beta_1 x_i}})^2 \\ &= -\frac{2}{n} \sum_{i=1}^n (y_i - e^{-e^{\beta_0 + \beta_1 x_i}}) e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}} \end{aligned}$$

and

$$\begin{aligned} \frac{\partial}{\partial \beta_1} S(\beta_0, \beta_1) &= \frac{1}{n} \sum_{i=1}^n \frac{\partial}{\partial \beta_1} (y_i - e^{-e^{\beta_0 + \beta_1 x_i}})^2 \\ &= -\frac{2}{n} \sum_{i=1}^n (y_i - e^{-e^{\beta_0 + \beta_1 x_i}}) x_i e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}}. \end{aligned}$$

Figure 4: Double exponential data



Setting these expressions equal to 0 results in the system

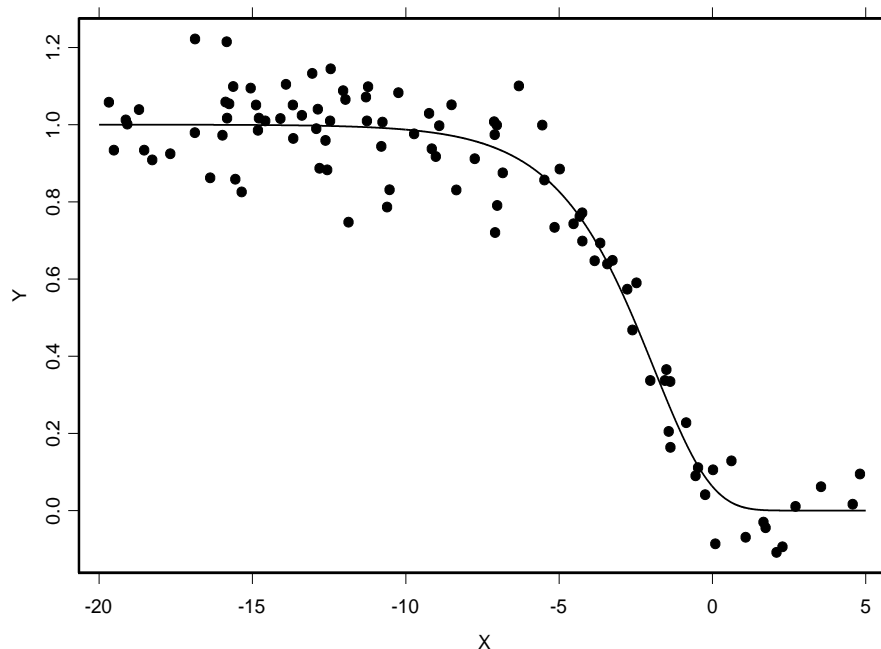
$$\begin{cases} \sum_i y_i e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}} = \sum_i e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}} \\ \sum_i x_i y_i e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}} = \sum_i x_i e^{\beta_0 + \beta_1 x_i - e^{\beta_0 + \beta_1 x_i}} \end{cases} \quad (9)$$

Clearly this is a pair of ugly equations, and it seems apparent that we have no hope of solving these explicitly (coming up with explicit formulas) for  $\beta_0$  and  $\beta_1$ . Because there are two equations with two unknowns, one expects that there is a solution, and fortunately computers are the perfect tools to help us. (I do need to note that even with computers, the computational task is not always straightforward, so care must be taken when implementing software to perform such computations that the resulting “solutions” are indeed correct.)

Fortunately, there is a variety of commercial and free software packages and routines which solve such equations, or one can program them oneself if needed. I used a commercial software package called S-Plus to arrive at a solution to the above system of equations. The values for the parameters with the present set of points (Figure 4) are  $\hat{\beta}_0 = 1.02154$  and  $\hat{\beta}_1 = 0.5360564$ , the associated curve being plotted in Figure 5.

I must admit that I cheated a bit. I generated the data points for this example in Figure 4 by taking some values for  $x$  between  $-20$  and  $5$  and plugging them into the function in Equation (8) with parameters set to  $\beta_0 = 1$  and  $\beta_1 = 0.5$  to get some  $y$  values. Then I “wiggled” these  $y$  values by adding some random error (normally distributed with

Figure 5: Double exponential fit



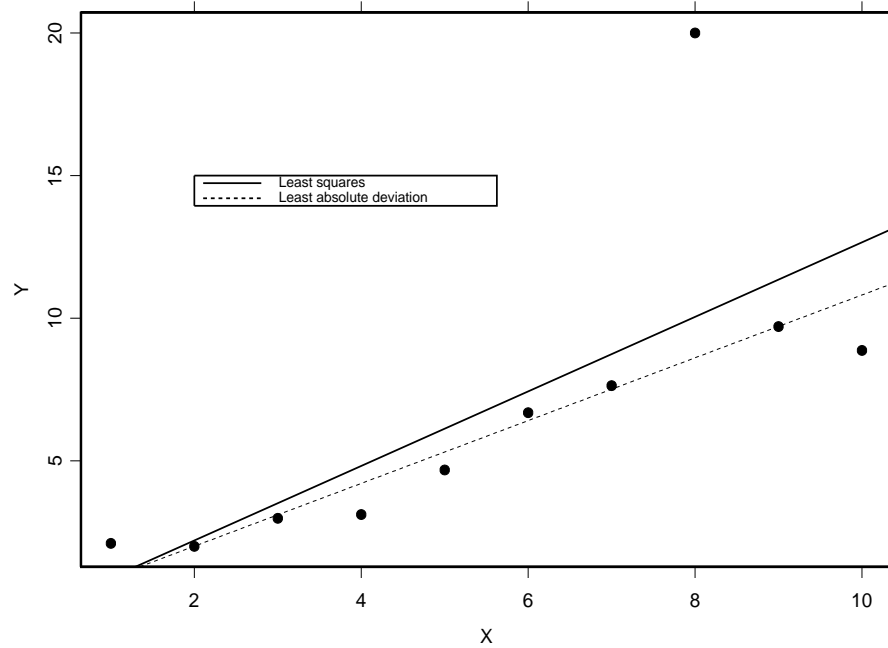
mean 0 and variance 0.01). This is why the values for  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are close to 1 and 0.5—and it is a good thing they are! I did this simply to illustrate the method. The S-Plus program used to find the values for the  $\hat{\beta}$ s is given in the appendix.

## 5 Comments

The method of least squares is not perfect. In fact, when it was first introduced, it met with sharp criticism, and the same issues voiced then apply now. The main difficulty is that because the errors  $e_i$  are squared in the function  $S(\beta_0, \beta_1)$ , points which disagree with the general pattern seen in the remaining points are given too much emphasis, in that their  $e_i$  values are large compared with the others'. The result of this is that the fitted curve is “pulled” toward these outlying points. This is not the case for least absolute deviations (LAD) curve fitting, which is more intuitive anyhow. An artificial example of this phenomenon is given in Figure 6, where the solid line is the least squares (LS) fit, and the dashed line is the LAD fit. The LS-fitted line is clearly drawn toward the outlying point, while the LAD-fitted line is relatively unaffected by this point and tracks the general, linear shape of other points reasonably well.

Other than the fact the method of least squares produces a reasonably easily computable answer, the parameter *estimates*—the values derived for  $\hat{\beta}$ s from a dataset for a particular problem—have some nice “statistical” properties under some commonly made assumptions. Those properties are beyond this essay, but in general they let one make

Figure 6: Least squares vs. Least absolute deviation with outlying value



reasonable statements about how precise the estimates are to the “truth.” Some references on the regression, including the least squares method, are given in the References section at the end of the paper. Many, many more are available.

A final note to make concerns the use of the least squares over least absolute deviation today, in this age of fast computers and readily available software. The statistical properties mentioned above for estimates from least squares are still true, but there is an ever-increasing variety of other, computer intensive ways to investigate and get a handle on the statistical properties of more complicated, less well-behaves estimates like those from least absolute deviation. The intuitive appeal and inherent validity of some methods either previously abandoned or not even attempted are being revisited using more modern, computer based approaches to statistical analyses, allowing one to get for problems previously impossible to solve.

## 6 Appendix

### 6.1 Teacher Pay Data

These data can be found at <http://lib.stat.cmu.edu/DASL/Stories/teacherpay.html> as part of the Data & Story Library (DASL) collection in the Statistics Library (StatLib). See this site for complete reference.

Table 2: Teacher Pay data

State	Average Teacher Pay	Average Per Student	State	Average Teacher Pay	Average Per Student
ME	19583	3346	NC	22795	3366
NH	20263	3114	SC	21570	2920
VT	20325	3554	GA	22080	2980
MA	26800	4642	FL	22250	3731
RI	29470	4669	KY	20940	2853
CT	26610	4888	TE	21800	2533
NY	30678	5710	AL	22934	2729
NJ	27170	5536	MS	18443	2305
PA	25853	4168	AR	19538	2642
OH	24500	3547	LA	20460	3124
IN	24274	3159	OK	21419	2752
IL	27170	3621	TX	25160	3429
MI	30168	3782	MT	22482	3947
WI	26525	4247	ID	20969	2509
MN	27360	3982	WY	27224	5440
IA	21690	3568	CO	25892	4042
MO	21974	3155	NM	22644	3402
ND	20816	3059	AZ	24640	2829
SD	18095	2967	UT	22341	2297
NB	20939	3285	NV	25610	2932
KA	22644	3914	WA	26015	3705
DE	24624	4517	OR	25788	4123
MD	27186	4349	CA	29132	3608
DC	33990	5020	AK	41480	8349
VA	23382	3594	HA	25845	3766
WV	20627	2821			

## 6.2 S-Plus function for double exponential example

```
"rocky.ls"<-
  function(x, y, b0 = 1, b1 = 1, trace = F)
  {
    iter1 <- function(b0, b1, x, y)
    {
      sum((y - exp(-exp(b0 + b1 * x))) *
          (exp(2 * b0 + b1 * x - exp(b0 + b1 * x))))
    }
    iter2 <- function(b1, b0, x, y)
    {
      sum((y - exp(-exp(b0 + b1 * x))) *
          (x * exp(b0 + b1 * x - exp(b0 + b1 * x))))
    }
    max.diff <- 1
    b1.new <- b1
    b0.new <- b0
    while(max.diff > 0.0001) {
      b1.old <- b1.new
      b0.old <- b0.new
      b0.new <- uniroot(iter1, c(-10, 10), b1 = b1.old, x = x,
                        y = y)$root
      b1.new <- uniroot(iter2, c(-10, 10), b0 = b0.new, x = x,
                        y = y)$root
      if(trace)
        max.diff <- max(abs(b0.old - b0.new), abs(b1.old - b1.new))
    }
    cbind(b0 = b0.new, b1 = b1.new)
  }
}
```

## 7 References

- Applied regression analysis, linear models, and related methods*, John Fox, Sage Publications, Inc., Thousand Oaks, Calif., 1997.
- Regression analysis: concepts and applications*, Franklin A. Graybill and Hariharan K. Iyer, Duxbury Press, Belmont, Calif., 1994.
- Regression analysis by example, 3rd Ed.* Samprit Chatterjee, Ali S. Hadi, and Bertram Price, John Wiley & Sons, Inc., New York, 2000.