Least-Absolute-Value Regression

by Michael Kupferschmid

Introduction

Often there is an approximately linear relationship between two measured quantities x and y, and we want to find the model function $\hat{y}(x) = \beta_1 + \beta_2 x$ that best fits the data (x_k, y_k), where $k = 1 \ldots m$. As discussed in [1, §2.3] and [2, §8.6.4], one way of doing this is by finding β_1 and β_2 to minimize the sum of the absolute values of the deviations $d_k = y_k - \hat{y}(x_k)$ between the model function and the observations. This way of determining the model function is called **least-absolute-value regression**. It might be that some of the data points are more certainly known or more important than others, and if so they should be weighted more heavily, so we will associate with each data point (x_k, y_k) a weight w_k and

$$\underset{\beta_1,\beta_2}{\text{minimize}} \sum_{k=1}^{m} w_k |d_k|.$$

Linear Programming Formulation

To eliminate the absolute value, we can let $d_k = u_k - v_k$, where $u_k \geq 0$, $v_k \geq 0$, and one or the other (depending on the sign of d_k) is zero. Then we can write $|d_k| = u_k + v_k$, and the optimization problem becomes the linear program

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v},\beta_1,\beta_2}{\text{minimize}} \quad & \sum_{k=1}^{m} w_k (u_k + v_k) \\ \text{subject to} \quad & u_k - v_k = y_k - (\beta_1 + \beta_2 x_k) \\ & \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0} \\ & \beta_1, \beta_2 \quad \text{free} \end{aligned}$$

To get standard form we can write each free variable as the difference between two nonnegative variables so that $\beta_1 = \beta_1^+ - p$ and $\beta_2 = \beta_2^+ - p$, where $\beta_1^+ \geq 0$, $\beta_2^+ \geq 0$, and $p \geq 0$. Then the linear program becomes

$$\begin{aligned} \underset{\mathbf{u},\mathbf{v},\beta_1^+,\beta_2^+,p}{\text{minimize}} \quad & w_1 u_1 + w_1 v_1 + \cdots + w_m u_m + w_m v_m \\ \text{subject to} \quad & \beta_1^+ + x_1 \beta_2^+ - (1 + x_1) p + u_1 - v_1 = y_1 \\ & \beta_1^+ + x_2 \beta_2^+ - (1 + x_2) p + u_2 - v_2 = y_2 \\ & \vdots \\ & \beta_1^+ + x_m \beta_2^+ - (1 + x_m) p + u_m - v_m = y_m \\ & \beta_1^+ \geq 0, \beta_2^+ \geq 0, p \geq 0, \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0} \end{aligned}$$

This linear program has the tableau shown at the top of the next page.

	β_1^+	β_2^+	p	u_1	v_1	u_2	v_2	$\cdots$	u_m	v_m
0	0	0	0	w_1	w_1	w_2	w_2	$\cdots$	w_m	w_m
y_1	1	x_1	$-(1+x_1)$	1	-1	0	0		0	0
y_2	1	x_2	$-(1+x_2)$	0	0	1	-1		0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$					$\ddots$		
y_m	1	x_m	$-(1+x_m)$	0	0	0	0		1	-1

Behavior of the Regression Model

It is instructive to consider some small examples of least-absolute-value regression in action. First suppose that the data points to be fitted with a straight line are (0,1), (1,3), and (2,5) and that they are all weighted the same. Then this is the linear program.

	β_1^+	β_2^+	p	u_1	v_1	u_2	v_2	u_3	v_3
0	0	0	0	1	1	1	1	1	1
1	1	0	-1	1	-1	0	0	0	0
3	1	1	-2	0	0	1	-1	0	0
5	1	2	-3	0	0	0	0	1	-1

Solving by the simplex method yields this unique optimal-form tableau.

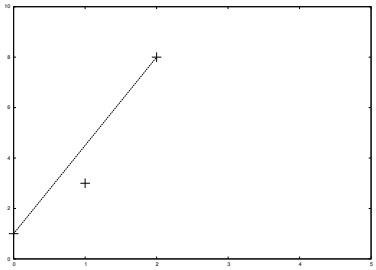
	β_1^+	β_2^+	p	u_1	v_1	u_2	v_2	u_3	v_3
0	0	0	0	$\frac{1}{2}$	$1\frac{1}{2}$	2	0	$\frac{1}{2}$	$1\frac{1}{2}$
1	1	0	-1	1	-1	0	0	0	0
2	0	1	-1	$-\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$-\frac{1}{2}$
0	0	0	0	$\frac{1}{2}$	$-\frac{1}{2}$	-1	1	$\frac{1}{2}$	$-\frac{1}{2}$

Thus $\beta_1 = 1$ and $\beta_2 = 2$, so the model function is $\hat{y}(x) = 1 + 2x$, which fits the data precisely (the objective value is zero). Next consider what happens when the final point is moved off that line, to (2,8). Here are the new initial and final tableaus.

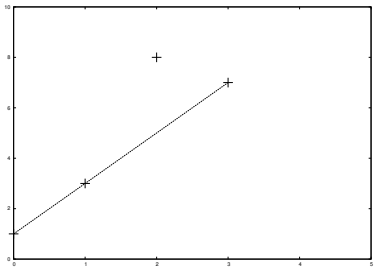
	β_1^+	β_2^+	p	u_1	v_1	u_2	v_2	u_3	v_3
0	0	0	0	1	1	1	1	1	1
1	1	0	-1	1	-1	0	0	0	0
3	1	1	-2	0	0	1	-1	0	0
8	1	2	-3	0	0	0	0	1	-1

	β_1^+	β_2^+	p	u_1	v_1	u_2	v_2	u_3	v_3
$-1\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$1\frac{1}{2}$	2	0	$\frac{1}{2}$	$1\frac{1}{2}$
1	1	0	-1	1	-1	0	0	0	0
$3\frac{1}{2}$	0	1	-1	$-\frac{1}{2}$	$\frac{1}{2}$	0	0	$\frac{1}{2}$	$-\frac{1}{2}$
$1\frac{1}{2}$	0	0	0	$\frac{1}{2}$	$-\frac{1}{2}$	-1	1	$\frac{1}{2}$	$-\frac{1}{2}$

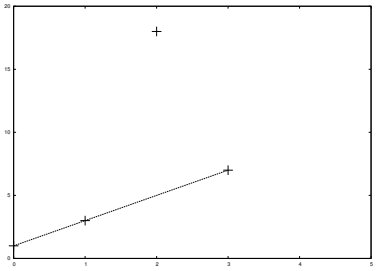
The unique optimal solution now gives $\hat{y}(x) = 1+3.5x$, which goes through the first and third points but completely ignores the second (this is in contrast to the least-squares regression line, which is $\bar{y}(x) = 0.5 + 3.5x$ and does not pass through any of the data points).



Here $\hat{y}(1) = 4.5$ so $d_2 = 1.5$ and that is the optimal objective value. Next suppose we add the point (3,7).



Now three of the data points lie on the regression line but the fourth, which we can now call an "outlier," is ignored. What is remarkable about this result is that the regression line remains unchanged *no matter how far away we move the outlier.* Here is the picture with the third point moved to (3,18). The vertical scale has changed to accommodate the moved point.

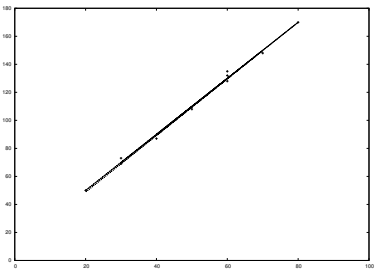


Because of the linearity of the regression line and the fact that we are minimizing the sum of the distances from it to the points, there is no way to adjust the line that results in a lower error, no matter how far away the outlier is (this is analogous to the median of several numbers remaining unchanged as the highest or lowest value is made more extreme). Decreasing the deviation of the outlier always increases the sum of the other deviations by more. It is this property that makes least-absolute-value regression useful for rejecting outliers.

Comparison to Least-Squares Regression

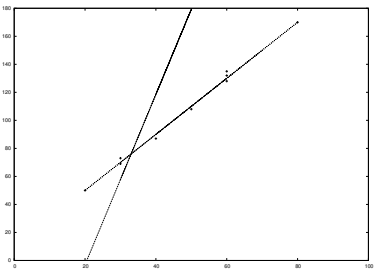
Now consider the two larger datasets given below. The left table is from the Westwood Company example in [3, §3], and the right table is the same data with a digit inversion in one of the observations (marked $\star$). This typographical error introduces a very pronounced outlier, which is off the vertical scale in the rightmost graph.

x_k	y_k
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	148
60	132



$$\bar{y}(x) = 10 + 2x$$
$$\hat{y}(x) = 8.4 + 2.02x$$

x_k	y_k
30	73
20	50
60	128
80	170
40	87
50	108
60	135
30	69
70	$\star$841
60	132



$$\bar{y}(x) = -124.5 + 6.076x$$
$$\hat{y}(x) = 10 + 2x$$

Except for the outlier, each graph shows the tabular data as points and both the least-squares ($\bar{y}$) and least-absolute-values ($\hat{y}$) regression lines. In the left graph, where there are no significant outliers, the model functions are similar and both regression lines provide a good fit to the data. The outlier in the right dataset pulls the least-squares regression line far from the remainder of the data, while the least-absolute-values regression line is affected very little (in fact, the outlier changes the set of data points that matter to the linear program in a way that just happens to yield the same model function that least-squares regression found for the original data).

Trustworthiness of Regression Models

Typically the values of y_k include some random errors resulting from limits on the precision of the measurements or our neglect of factors other than x that also slightly influence the value of y. Because of these errors, our linear model function will seldom pass exactly through all of the data points. How confident can we be in the predictive value of the regression, if it does not match the data perfectly? If more data or different measurements were used, we would probably get slightly different values for β_1 and β_2. Just how different might they be?

Often it is reasonable to assume that the errors in the measured values of the y_k are normally distributed with zero mean. In that case their variance can be estimated, using the data and the regression function, as

$$\hat{\sigma}^2 = \frac{1}{m-2} \sum_{k=1}^{m} (y_k - (\beta_1 + \beta_2 x_k))^2$$

In the case of least-squares regression, β_1 and β_2 are functions of the data determined by these formulas.

$$\beta_2 = \frac{\sum x_k y_k - \frac{1}{m}(\sum x_k)(\sum y_k)}{\sum x_k^2 - \frac{1}{m}(\sum x_k)^2}$$

$$\beta_1 = \frac{1}{m}\Big(\sum y_k - \beta_2 \sum x_k\Big)$$

Assuming that the errors in the y_k are normally distributed with mean zero and variance $\hat{\sigma}^2$, techniques of mathematical statistics can be used to derive the probability distributions of these functions of the random variable y. It is then possible to make probability statements about the regression coefficients β_1 and β_2, such as giving 90% confidence intervals on their values.

In the case of least-absolute-values regression β_1 and β_2 are also functions of the data, but those functions are *not* given by formulas. Instead, the relationship between the random variable y and the regression coefficients is determined by the solution of a linear programming problem. The techniques of mathematical statistics that work so well and easily in the case of least-squares regression fail us here. Instead, we must resort to simulation experiments and determine the probability distributions of β_1 and β_2 experimentally.

Simulation Experiments

To determine the probability distributions of β_1 and β_2 experimentally, I wrote `simulation.f`, which peturbs the y_k in the Westwood data by adding pseudorandom noise uniformly distributed on the interval $[-r, r]$, where r is the **noise amplitude**. For each of 900 vectors **y** perturbed in this way, the program computes the least-squares and least-absolute-values regression, and saves the resulting values of β_2 (the slope of the regression line). Then it prints histograms (sample probability densities) of the two β_2 datasets. Some results are shown below and on the next page.

```
L2, r= 0.0                              L1, r= 0.0
    -1.50                                   -1.50
    -0.50                                   -0.50
     0.50                                    0.50
     1.50 ..............................     1.50 ..............................
     2.50                                    2.50
     3.50                                    3.50
     4.50                                    4.50
     5.50                                    5.50
```

When the noise amplitude is zero, all of the regressions yield the same slope, 2.00 for least squares and 2.02 for least absolute values as we found above. As the noise in the data

increases, both least-squares and least-absolute-values regression yield β_2 probability distributions that increase in variance. The least-squares density becomes skewed, and at $r = 10$ it is also biased (in that its peak occurs in the wrong histogram bin). The least-absolute-values density remains unbiased and roughly symmetric, but it spreads out more.

```
L2, r= 3.0                          L1, r= 3.0
    -1.50                               -1.50
    -0.50                               -0.50
     0.50 .                              0.50 .
     1.50 ............................   1.50 ........................
     2.50                                2.50 ..
     3.50                                3.50
     4.50                                4.50
     5.50                                5.50


L2, r= 5.0                          L1, r= 5.0
    -1.50                               -1.50
    -0.50                               -0.50
     0.50 .........                      0.50 ..
     1.50 ....................          1.50 ......................
     2.50 .                              2.50 .....
     3.50                                3.50
     4.50                                4.50
     5.50                                5.50


L2, r=10.0                          L1, r=10.0
    -1.50                               -1.50
    -0.50 .                             -0.50 .
     0.50 ..............                 0.50 .....
     1.50 ..........                     1.50 ..............
     2.50 ....                           2.50 ......
     3.50                                3.50 ..
     4.50                                4.50 .
     5.50                                5.50
```

Open Questions

The conventional analysis of least-squares regression assumes that measurement errors in the y_k are normally distributed, and this yields variations in the β_2 estimates that follow Student's t distribution, but here we have simulated uniformly distributed measurement errors. How should the β_2 that we find using least-squares regression be distributed then, according to mathematical statistics? It would be reassuring to find that the densities on the left above match those predictions.

It's not possible to derive in closed form the probability distribution of β_2 when it is estimated using least-absolute-values regression, but the bottom right histogram above is suggestive of a double-exponential density. It would be interesting to see how that changes if the errors in the y_k are assumed to be normally, rather than uniformly, distributed.

These results are based on a single stream of pseudorandom numbers, so to confirm the particular observations made above it would be prudent to repeat the calculations using some different starting seeds. If the double-exponential appearance of the least-absolute-values β_2 density persists, other and larger problems could be studied in search of an empirical analytic model for these uncertainties.

When $r = 10$ some of the least-absolute-value β_2 estimates are quite different from the estimate of 2.02 we get when $r = 0$. What does the linear program look like in those cases? Is the optimal basic sequence different from what it is when $r = 0$? If not, does making the noise amplitude even bigger cause a change in the basic sequence, and what does the histogram of β_2 values look like then? Letting $r = 50$ yields the following histograms.

```
L2, r=50.0
    -1.50 ....
    -0.50 ...
     0.50 ..
     1.50 ..
     2.50 .
     3.50 .
     4.50 .
     5.50 .
```

```
L1, r=50.0
    -1.50 ...
    -0.50 ....
     0.50 ....
     1.50 ...
     2.50 ..
     3.50 .
     4.50 ..
     5.50 .
```

These histograms display the same range of β_2 values that we used before, but now some observations fall outside that range so not all 30 of the histogram dots show. Now the right-hand density looks quite different from the one for $r = 10$, having for one thing a second hump around $\beta_2 = 5$. Does that correspond to a different optimal basis in the linear program? Errors of ± 50 in data values that are barely bigger than that would probably make the data useless in practice, but it is interesting anyway to ask how changes in the optimal basic sequence of the linear program affect the β_2 density plot.

References

[1] **Ecker, Joseph G.** and **Kupferschmid, Michael**, *Introduction to Operations Research*, Third Edition, Krieger Publishing, 2004.

[2] **Kupferschmid, Michael**, *Introduction to Mathematical Programming,* First Edition, 2023.

[3] **Neter, John** and **Wasserman, William**, *Applied Linear Statistical Models: Regression, Analysis of Variance, and Experimental Designs,* Irwin, 1974.