Some Surprising Results in Linear Regression

Wagner L. Truppel

January 19, 2004

Review

Let $S_1 = \{x_i, y_i\}$, $1 \leq i \leq m$ and $S_2 = \{x'_j, y'_j\}$, $1 \leq j \leq n$ be two sets of points. Performing a least-squares regression on S_1 alone yields:

$$\begin{aligned}
y &= a_1\, x + b_1 \\
\bar{y}_i &= a_1\, x_i + b_1, \quad 1 \leq i \leq m \\
R_1 &= \sum_{i=1}^{m} (\bar{y}_i - y_i)^2 = \sum_{i=1}^{m} \Big[(a_1\, x_i + b_1) - y_i\Big]^2 \\
\Delta_1 &= m \sum_{i=1}^{m} x_i^2 - \Big(\sum_{i=1}^{m} x_i\Big)^2 \\
\Delta_1\, a_1 &= m \sum_{i=1}^{m} x_i\, y_i - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} y_i \\
\Delta_1\, b_1 &= \sum_{i=1}^{m} y_i \sum_{i=1}^{m} x_i^2 - \sum_{i=1}^{m} x_i \sum_{i=1}^{m} x_i\, y_i\,.
\end{aligned}$$

Let's review how these results are obtained. In the process, we will come across two very useful identities. We need to find the line which minimizes the *residual*, that is, the line $y = a_1\, x + b_1$ for which $R_1 = \sum_{i=1}^{m} \Big[(a_1\, x_i + b_1) - y_i\Big]^2$ has the smallest possible value. The value $a_1\, x_i + b_1$, which we defined above as $\bar{y}_i$, is the y-value for x_i when the point with that x-coordinate is 'forced' to lie on the regressed line.

As a function of a_1 and b_1, R_1 will achieve its minimum or maximum value when its first partial derivatives with respect to a_1 and b_1 both vanish. Whether it achieves a minimum or a maximum value, however, will depend on its second derivatives. The first and second

partial derivatives of R_1 with respect to a_1 and b_1 are:

$$\begin{aligned}
\frac{\partial R_1}{\partial a_1} &= 2\sum_{i=1}^{m}\Big[(a_1\,x_i + b_1) - y_i\Big]x_i &&= 2\Big(a_1\sum_{i=1}^{m}x_i^2 + b_1\sum_{i=1}^{m}x_i - \sum_{i=1}^{m}x_i\,y_i\Big)\\
\frac{\partial R_1}{\partial b_1} &= 2\sum_{i=1}^{m}\Big[(a_1\,x_i + b_1) - y_i\Big] &&= 2\Big(a_1\sum_{i=1}^{m}x_i + m\,b_1 - \sum_{i=1}^{m}y_i\Big)\\
\frac{\partial^2 R_1}{\partial a_1^2} &= 2\sum_{i=1}^{m}x_i^2\,, \quad \frac{\partial^2 R_1}{\partial b_1^2} = 2\,m\,, \qquad \text{and} \quad \frac{\partial^2 R_1}{\partial a_1 \partial b_1} = \frac{\partial^2 R_1}{\partial b_1 \partial a_1} = 2\sum_{i=1}^{m}x_i\,.
\end{aligned}$$

Note that $\partial^2 R_1/\partial a_1^2$ and $\partial^2 R_1/\partial b_1^2$ are both positive, while the mixed second derivative may be positive or negative (or even zero). R_1 will be a minimum provided the matrix

$$\begin{pmatrix} \dfrac{\partial^2 R_1}{\partial a_1^2} & \dfrac{\partial^2 R_1}{\partial a_1 b_1} \\[2ex] \dfrac{\partial^2 R_1}{\partial b_1 a_1} & \dfrac{\partial^2 R_1}{\partial b_1^2} \end{pmatrix} = 2\begin{pmatrix} \displaystyle\sum_{i=1}^{m}x_i^2 & \displaystyle\sum_{i=1}^{m}x_i \\[2ex] \displaystyle\sum_{i=1}^{m}x_i & m \end{pmatrix}$$

has positive eigenvalues. It's not difficult to show that this condition is satisfied if and only if $m\displaystyle\sum_{i=1}^{m}x_i^2 \geq (\sum_{i=1}^{m}x_i)^2$. But this is always true, as the following argument shows:

$$\sum_{i=1}^{m}(x_i-\mu)^2 \;\geq\; 0 \quad\Rightarrow\quad \sum_{i=1}^{m}x_i^2 \;\geq\; m\,\mu^2 \quad\Rightarrow\quad m\sum_{i=1}^{m}x_i^2 \;\geq\; \Big(\sum_{i=1}^{m}x_i\Big)^2\,,$$

where $\mu = \dfrac{1}{m}\displaystyle\sum_{i=1}^{m}x_i$ is the mean of the x values. Thus, we have proved that least-squares regression does indeed *always* minimize the residual. The actual optimal values of a_1 and b_1 are then obtained by setting the first partial derivatives to zero and solving the corresponding set of linear equations for the unknown parameters, resulting in the expressions quoted earlier. Note the intermediate results

$$\begin{aligned}
a_1\sum_{i=1}^{m}x_i^2 + b_1\sum_{i=1}^{m}x_i &= \sum_{i=1}^{m}x_i\,y_i\\
a_1\sum_{i=1}^{m}x_i + m\,b_1 &= \sum_{i=1}^{m}y_i\,.
\end{aligned}$$

These two identities will come to be essential in proving several of the results which are to follow.

Two surprising results

Performing a least-squares regression on S_2 alone yields similar results:

$$\begin{aligned}
y &= a_2\,x + b_2 \\
\bar{y}'_j &= a_2\,x'_j + b_2, \quad 1 \le j \le n \\
R_2 &= \sum_{j=1}^{n} (\bar{y}'_j - y'_j)^2 = \sum_{j=1}^{n} \Big[(a_2\,x'_j + b_2) - y'_j\Big]^2 \\
\Delta_2 &= n \sum_{j=1}^{n} {x'_j}^2 - \Big(\sum_{j=1}^{n} x'_j\Big)^2 \\
\Delta_2\,a_2 &= n \sum_{j=1}^{n} x'_j\,y'_j - \sum_{j=1}^{n} x'_j \sum_{j=1}^{n} y'_j \\
\Delta_2\,b_2 &= \sum_{j=1}^{n} y'_j \sum_{j=1}^{n} {x'_j}^2 - \sum_{j=1}^{n} x'_j \sum_{j=1}^{n} x'_j\,y'_j\,.
\end{aligned}$$

Now, suppose we combine all points and perform a least-squares regression on the combined set $S_3 = S_1 \cup S_2$. We then get:

$$\begin{aligned}
y &= a_3\,x + b_3 \\
R_3 &= \sum_{\text{all}} \Big[(a_3\,x_k + b_3) - y_k\Big]^2 \\
\Delta_3 &= (m+n) \sum_{\text{all}} x_k^2 - \Big(\sum_{\text{all}} x_k\Big)^2 \\
\Delta_3\,a_3 &= (m+n) \sum_{\text{all}} x_k\,y_k - \sum_{\text{all}} x_k \sum_{\text{all}} y_k \\
\Delta_3\,b_3 &= \sum_{\text{all}} y_k \sum_{\text{all}} x_k^2 - \sum_{\text{all}} x_k \sum_{\text{all}} x_k\,y_k\,.
\end{aligned}$$

Alternatively, suppose we replace $S_1 = \{x_i, y_i\}$ with the set $\bar{S}_1 = \{x_i, \bar{y}_i\}$ and $S_2 = \{x'_j, y'_j\}$ with the set $\bar{S}_2 = \{x'_j, \bar{y}'_j\}$, that is, we force the y-values for each x to lie on the corresponding regressed lines. Then, we can perform a least-squares regression on the

combined set $\bar{S}_3 = \bar{S}_1 \cup \bar{S}_2$:

$$\begin{aligned} y &= \bar{a}_3\, x + \bar{b}_3 \\ \bar{R}_3 &= \sum_{i=1}^{m} \Big[(\bar{a}_3\, x_i + \bar{b}_3) - \bar{y}_i\Big]^2 + \sum_{j=1}^{n} \Big[(\bar{a}_3\, x'_j + \bar{b}_3) - \bar{y}'_j\Big]^2 \\ &= \sum_{i=1}^{m} \Big[(\bar{a}_3 - a_1)\, x_i + (\bar{b}_3 - b_1)\Big]^2 + \sum_{j=1}^{n} \Big[(\bar{a}_3 - a_2)\, x'_j + (\bar{b}_3 - b_2)\Big]^2 . \end{aligned}$$

Our first surprising result is that $\bar{a}_3 = a_3$ and $\bar{b}_3 = b_3$. That is, the slopes and intercepts are the same when we combine all points prior to performing the regression and when we first force each point in each set to fall on the corresponding regressed line, prior to combining these two new sets for a combined regression.

The proof is simple:

$$\begin{aligned} 0 = \tfrac{1}{2}\frac{\partial \bar{R}_3}{\partial \bar{a}_3} &= \sum_{i=1}^{m} \Big[(\bar{a}_3 - a_1)\, x_i + (\bar{b}_3 - b_1)\Big] x_i + \sum_{j=1}^{n} \Big[(\bar{a}_3 - a_2)\, x'_j + (\bar{b}_3 - b_2)\Big] x'_j \\ &= (\bar{a}_3 - a_1) \sum_{i=1}^{m} x_i^2 + (\bar{b}_3 - b_1) \sum_{i=1}^{m} x_i + (\bar{a}_3 - a_2) \sum_{j=1}^{n} {x'_j}^2 + (\bar{b}_3 - b_2) \sum_{j=1}^{n} x'_j \\ &= \Big(\sum_{i=1}^{m} x_i^2 + \sum_{j=1}^{n} {x'_j}^2\Big)\, \bar{a}_3 + \Big(\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} x'_j\Big)\, \bar{b}_3 \\ &\quad - \Big(a_1 \sum_{i=1}^{m} x_i^2 + b_1 \sum_{i=1}^{m} x_i\Big) - \Big(a_2 \sum_{j=1}^{n} {x'_j}^2 + b_2 \sum_{j=1}^{n} x'_j\Big) . \end{aligned}$$

The last two parenthesized terms are equal, respectively, to $\sum_{i=1}^{m} x_i\, y_i$ and $\sum_{j=1}^{n} x'_j\, y'_j$, thanks to the identities we derived earlier. Thus,

$$\begin{aligned} 0 = \tfrac{1}{2}\frac{\partial \bar{R}_3}{\partial \bar{a}_3} &= \Big(\sum_{i=1}^{m} x_i^2 + \sum_{j=1}^{n} {x'_j}^2\Big)\, \bar{a}_3 + \Big(\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} x'_j\Big)\, \bar{b}_3 - \Big(\sum_{i=1}^{m} x_i\, y_i + \sum_{j=1}^{n} x'_j\, y'_j\Big) \\ &= \bar{a}_3 \sum_{\text{all}} x_k^2 + \bar{b}_3 \sum_{\text{all}} x_k - \sum_{\text{all}} x_k\, y_k . \end{aligned}$$

Likewise,

$$\begin{aligned} 0 = \tfrac{1}{2}\frac{\partial \bar{R}_3}{\partial \bar{b}_3} &= \Big(\sum_{i=1}^{m} x_i + \sum_{j=1}^{n} x'_j\Big)\, \bar{a}_3 + (m+n)\, \bar{b}_3 - \Big(\sum_{i=1}^{m} y_i + \sum_{j=1}^{n} y'_j\Big) \\ &= \bar{a}_3 \sum_{\text{all}} x_k + (m+n)\bar{b}_3 - \sum_{\text{all}} y_k . \end{aligned}$$

But these are the exact same equations used to derive a_3 and b_3. Therefore, $\bar{a}_3 = a_3$ and $\bar{b}_3 = b_3$. Note, however, that $\bar{R}_3$ does *not* equal R_3:

$$\begin{aligned} R_3 &= \sum_{i=1}^{m} \Big[(a_3\, x_i + b_3) - y_i\Big]^2 + \sum_{j=1}^{n} \Big[(a_3\, x'_j + b_3) - y'_j\Big]^2 \\ \bar{R}_3 &= \sum_{i=1}^{m} \Big[(a_3\, x_i + b_3) - \bar{y}_i\Big]^2 + \sum_{j=1}^{n} \Big[(a_3\, x'_j + b_3) - \bar{y}'_j\Big]^2 . \end{aligned}$$

Adding and subtracting y_i inside the first bracket of $\bar{R}_3$, y'_j inside the second bracket, and expanding the result yields:

$$\begin{aligned} \bar{R}_3 &= \sum_{i=1}^{m} \Big[(a_3\, x_i + b_3 - y_i) - (\bar{y}_i - y_i)\Big]^2 + \sum_{j=1}^{n} \Big[(a_3\, x'_j + b_3 - y'_j) - (\bar{y}'_j - y'_j)\Big]^2 \\ &= R_3 + R_1 + R_2 + 2\sum_{i=1}^{m} (a_1\, x_i + b_1 - y_i)\, y_i + 2\sum_{j=1}^{n} (a_2\, x'_j + b_2 - y'_j)\, y'_j \,, \end{aligned}$$

where we used once again the identities derived previously. Thus,

$$\begin{aligned} \bar{R}_3 &= R_1 + R_2 + R_3 \\ &\quad +2\Big(a_1 \sum_{i=1}^{m} x_i\, y_i + b_1 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} y_i^2\Big) + 2\Big(a_2 \sum_{j=1}^{n} x'_j\, y'_j + b_2 \sum_{j=1}^{n} y'_j - \sum_{j=1}^{n} {y'_j}^2\Big) . \end{aligned}$$

As it turns out, however, the first parenthesized term equals $-R_1$ and the second, $-R_2$. Here's the proof for R_1:

$$\begin{aligned} R_1 &= \sum_{i=1}^{m} \Big[(a_1\, x_i + b_1) - y_i\Big]^2 \\ &= a_1 \underbrace{\Big(a_1 \sum_{i=1}^{m} x_i^2 + b_1 \sum_{i=1}^{m} x_i - \sum_{i=1}^{m} x_i\, y_i\Big)}_{\text{zero}} + b_1 \underbrace{\Big(a_1 \sum_{i=1}^{m} x_i + m\, b_1 - \sum_{i=1}^{m} y_i\Big)}_{\text{zero}} \\ &\quad - \Big(a_1 \sum_{i=1}^{m} x_i\, y_i + b_1 \sum_{i=1}^{m} y_i - \sum_{i=1}^{m} y_i^2\Big) . \end{aligned}$$

The proof for R_2 is similar. We then obtain our second surprising result,

$$\begin{aligned} \bar{R}_3 &= R_3 - (R_1 + R_2) \quad \text{or, in a more suggestive form,} \\ R_3 &= R_1 + R_2 + \bar{R}_3 . \end{aligned}$$

Since all residuals are intrinsically non-negative quantities, we may conclude the following inequalities:

$$\begin{aligned} R_3 &\geq R_1 \\ R_3 &\geq R_2 \\ R_3 &\geq \bar{R}_3 \\ R_3 &\geq R_1 + R_2 \\ R_3 &\geq R_1 + \bar{R}_3 \\ R_3 &\geq R_2 + \bar{R}_3 \,. \end{aligned}$$

In particular, the residual of the regression of all points combined (R_3) is always larger than the residual of all points combined after they've been forced to fall on their corresponding intermediate regression lines ($\bar{R}_3$).

Online regression

Suppose we've observed the set S_1 and have computed a least-squares regressed line. A natural question to ask is how that line's slope and intercept, as well as its associated residual, would change if we were to add a new point to the set, or remove one point from it.

So, let a_m, b_m, and R_m be the line's slope, intercept, and residual for the set of m points. Similarly, let $a_{m\pm1}$, $b_{m\pm1}$, and $R_{m\pm1}$ be the corresponding quantities when we have added to the set (upper sign) or removed from it (lower sign) one point, namely, $\{\hat{x}, \hat{y}\}$. It's an easy exercise to show that:

$$\begin{aligned} \Delta_{m\pm1} &= \Delta_m \pm m\,\hat{x}^2 \pm \sum_{i=1}^{m} x_i^2 \mp 2\,\hat{x} \sum_{i=1}^{m} x_i \\ \Delta_{m\pm1}\, a_{m\pm1} &= \Delta_m\, a_m \pm m\,\hat{x}\,\hat{y} \pm \sum_{i=1}^{m} x_i y_i \mp \hat{y} \sum_{i=1}^{m} x_i \mp \hat{x} \sum_{i=1}^{m} y_i \\ \Delta_{m\pm1}\, b_{m\pm1} &= \Delta_m\, b_m \pm \hat{y} \sum_{i=1}^{m} x_i^2 \pm \hat{x}^2 \sum_{i=1}^{m} y_i \mp \hat{x} \sum_{i=1}^{m} x_i\, y_i \mp \hat{x}\,\hat{y} \sum_{i=1}^{m} x_i \,. \end{aligned}$$

These expressions allow us to compute the new slope and intercept from the old ones, without having to redo the entire regression all over again.

The change in the residual is slightly trickier to obtain. Recall the identity

$$R_m = \sum_{i=1}^{m} y_i^2 - a_m \sum_{i=1}^{m} x_i\, y_i - b_m \sum_{i=1}^{m} y_i\,,$$

from which we may write the analogous result

$$R_{m\pm1} = \Big(\sum_{i=1}^{m} y_i^2 \pm \hat{y}^2\Big) - a_{m\pm1}\Big(\sum_{i=1}^{m} x_i\, y_i \pm \hat{x}\,\hat{y}\Big) - b_{m\pm1}\Big(\sum_{i=1}^{m} y_i \pm \hat{y}\Big)\,.$$

Subtracting the two expressions, we deduce that

$$R_{m\pm1} = R_m \pm \hat{y}^2 \mp a_{m\pm1}\hat{x}\,\hat{y} \mp b_{m\pm1}\hat{y} - (a_{m\pm1} - a_m)\sum_{i=1}^{m} x_i y_i - (b_{m\pm1} - b_m)\sum_{i=1}^{m} y_i\,.$$

Summary

We've proved four distinct results:

- $\bar{a}_3 = a_3$ and $\bar{b}_3 = b_3$,
- $R_3 = R_1 + R_2 + \bar{R}_3$,
- Online version of the regression line's slope and intercept:

$$\begin{aligned}
\Delta_{m\pm1} &= \Delta_m \pm m\,\hat{x}^2 \pm \sum_{i=1}^{m} x_i^2 \mp 2\,\hat{x}\sum_{i=1}^{m} x_i \\
\Delta_{m\pm1}\,a_{m\pm1} &= \Delta_m\,a_m \pm m\,\hat{x}\,\hat{y} \pm \sum_{i=1}^{m} x_i y_i \mp \hat{y}\sum_{i=1}^{m} x_i \mp \hat{x}\sum_{i=1}^{m} y_i \\
\Delta_{m\pm1}\,b_{m\pm1} &= \Delta_m\,b_m \pm \hat{y}\sum_{i=1}^{m} x_i^2 \pm \hat{x}^2\sum_{i=1}^{m} y_i \mp \hat{x}\sum_{i=1}^{m} x_i\,y_i \mp \hat{x}\,\hat{y}\sum_{i=1}^{m} x_i\,,
\end{aligned}$$

- Online version of the residual:

$$R_{m\pm1} = R_m \pm \hat{y}^2 \mp a_{m\pm1}\hat{x}\,\hat{y} \mp b_{m\pm1}\hat{y} - (a_{m\pm1} - a_m)\sum_{i=1}^{m} x_i y_i - (b_{m\pm1} - b_m)\sum_{i=1}^{m} y_i\,.$$

■