Linear Regression with variances on the dependent variable

Wagner L. Truppel

December 2016

This is just a brief collection of results used in my *WTLinearRegression* library.

Definitions

Let $S_n = \{(x_i, y_i, \sigma_i)\}$, $1 \leq i \leq n$, be a collection of *observations*, where σ_i^2 is the variance associated with y_i or, equivalently, σ_i is its standard deviation. We're interested in performing a linear regression on that collection. Start by giving a name to the sum of inverse variances, since that sum appears all over the place,

$$\text{sivar}_n \equiv \sum_{i=1}^{n} \frac{1}{\sigma_i^2} \,.$$

A least-squares linear regression on S_n amounts to minimizing the quantity

$$\textbf{mean squared residual error} \equiv \langle \text{resSE} \rangle_n \equiv \frac{1}{\text{sivar}_n} \sum_{i=1}^{n} \Big[\frac{(a_n x_i + b_n) - y_i}{\sigma_i} \Big]^2$$

with respect to all possible choices of a_n and b_n. Other useful measures of error are

$$\textbf{mean squared regression error} \equiv \langle \text{regSE} \rangle_n \equiv \frac{1}{\text{sivar}_n} \sum_{i=1}^{n} \Big[\frac{(a_n x_i + b_n) - \langle y \rangle_n}{\sigma_i} \Big]^2$$

$$\text{and } \textbf{mean total squared error} \equiv \langle \text{totSE} \rangle_n \equiv \frac{1}{\text{sivar}_n} \sum_{i=1}^{n} \Big(\frac{y_i - \langle y \rangle_n}{\sigma_i} \Big)^2$$

where

$$\langle y \rangle_n \equiv \frac{1}{\text{sivar}_n} \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} \,.$$

The usual dance of finding the partial derivatives of $\langle \text{resSE} \rangle_n$ with respect to a_n and b_n and setting them both to zero yields the system of linear equations

$$\Big(\sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2} \Big) a_n \; + \; \Big(\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \Big) b_n \; = \; \sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}$$

$$\Big(\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2} \Big) a_n \; + \; \Big(\sum_{i=1}^{n} \frac{1}{\sigma_i^2} \Big) b_n \; = \; \sum_{i=1}^{n} \frac{y_i}{\sigma_i^2} \, .$$

Dividing every term by sivar_n, we find

$$\begin{aligned} \langle x^2 \rangle_n \, a_n \; + \; \langle x \rangle_n \, b_n \; &= \; \langle xy \rangle_n \\ \langle x \rangle_n \, a_n \; + \qquad b_n \; &= \; \langle y \rangle_n \end{aligned}$$

with obvious definitions for the various weighted means. A solution, when it exists, gives us a line with *slope* a_n and *y-intercept* b_n where

$$\begin{aligned} \Delta_n \; &= \; \langle x^2 \rangle_n - \langle x \rangle_n^2 \\ \Delta_n \, a_n \; &= \; \langle xy \rangle_n - \langle x \rangle_n \, \langle y \rangle_n \\ \Delta_n \, b_n \; &= \; \langle x^2 \rangle_n \, \langle y \rangle_n - \langle x \rangle_n \, \langle xy \rangle_n \, . \end{aligned}$$

Note that Δ_n can be rewritten as

$$\Delta_n = \frac{1}{\text{sivar}_n} \sum_{i=1}^{n} \Big(\frac{x_i - \langle x \rangle_n}{\sigma_i} \Big)^2$$

so, being the sum of non-negative numbers, cannot itself have a negative value. Moreover, it's zero *only* when $x_i = \langle x \rangle_n$ for all i. In other words, $\Delta_n \geq 0$ and equals zero only when all x_i have the same value, namely, $\langle x \rangle_n$.

Thus, provided that $\Delta_n \neq 0$, we obtain a finite slope and a finite y-intercept for our regression line. If, however, the x_i values are all the same, then either we have a vertical line at $x = \langle x \rangle_n$ (if the y_i values have some spread), or we have a situation where all the observations are identical ($x_i = \langle x \rangle_n$ and $y_i = \langle y \rangle_n$, for all i), in which case there is no regression line.[1]

We can summarize the conditions under which we can compute a regression line as follows:

- line with a finite slope a_n and a finite y-intercept b_n, if the set of x_i values has a **non-zero** variance, *i.e.*, $\Delta_n \neq 0$.

[1] We could say that the regression line degenerates to a point.

- vertical line at the x-intercept $\langle x\rangle_n$, if the set of x_i values has **zero** variance and the set of y_i values has a **non-zero** variance, *i.e.*, $\Delta_n = 0$ and $\langle \text{totSE}\rangle_n \neq 0$.
- undefined (or degenerate) line if both the set of x_i values and the set of y_i values have **zero** variance, *i.e.*, $\Delta_n = 0$ and $\langle \text{totSE}\rangle_n = 0$.

Once we have a regression line, how do we compute a measure of its fitness to the data? The standard practice is to compute

$$r_n^2 \equiv 1 - \frac{\langle \text{resSE}\rangle_n}{\langle \text{totSE}\rangle_n} \,.$$

The closer this value is to 1, the closer the regression line fits the data. From the considerations above, it's not obvious when r_n^2 will have a well-defined value. The table below makes that determination clear:

Δ_n	$\langle \text{totSE}\rangle_n$	line	slope, y-intercept	x-intercept	$\langle \text{resSE}\rangle_n$	r_n^2
$= 0$	$= 0$	degenerate	undefined	undefined	undefined	undefined
$= 0$	> 0	vertical	undefined	$\langle x\rangle_n$	undefined	undefined
> 0	$= 0$	horizontal	$a = 0$, $b = \langle y\rangle_n$	undefined	$= 0$	undefined
> 0	> 0	typical	finite $a \neq 0$, b	$-b/a$	≥ 0	≤ 1

In the cases where the line is exactly vertical or exactly horizontal, we could alternatively *define* r^2 to be 1 — despite the fact that we wouldn't be able to compute its value from its definition — since those lines are perfect fits for the data in question.

Confidence intervals

The standard variances of the estimates for the slope a_n and y-intercept b_n are given by[2]

$$(\sigma_a^2)_n = \frac{1}{(\text{sivar}_n - 2)} \frac{\displaystyle\sum_{i=1}^{n} \frac{(a_n x_i + b_n - y_i)^2}{\sigma_i^2}}{\displaystyle\sum_{i=1}^{n} \frac{(x_i - \langle x\rangle_n)^2}{\sigma_i^2}} = \frac{1}{(\text{sivar}_n - 2)} \frac{\langle \text{resSE}\rangle_n}{\Delta_n}$$

$$(\sigma_b^2)_n = (\sigma_a^2)_n \, \langle x^2\rangle_n \,.$$

Technically, we'd have to multiply $(\sigma_a)_n$ and $(\sigma_b)_n$ by a factor coming from a *Student's t* distribution with $n-2$ degrees of freedom. In practice, for any number of observations above 20 or so, that factor for a 95% confidence interval is about 2, so we can say that the slope a_n and y-intercept b_n have a width of about $4(\sigma_a)_n$ and $4(\sigma_b)_n$, respectively, centered on their estimated point values.

[2]Adapted from `https://en.m.wikipedia.org/wiki/Simple_linear_regression`.

Online regression

Now suppose that we have a collection of n observations, for which we have already computed all the quantities mentioned in the previous section, and we then add a new observation or remove a previous observation. The goal is to be able to compute the quantities from the previous section — after $n \pm 1$ observations — from the quantities after n observations. It's clear from the previous sections that we'll need to maintain the various weighted sums

- $\displaystyle\sum_{i=1}^{n} \frac{1}{\sigma_i^2}$
- $\displaystyle\sum_{i=1}^{n} \frac{x_i}{\sigma_i^2}, \sum_{i=1}^{n} \frac{x_i^2}{\sigma_i^2}$
- $\displaystyle\sum_{i=1}^{n} \frac{y_i}{\sigma_i^2}, \sum_{i=1}^{n} \frac{y_i^2}{\sigma_i^2}$
- $\displaystyle\sum_{i=1}^{n} \frac{x_i y_i}{\sigma_i^2}$

as these will let us compute the various Δs, the estimated slope and its variance, and the estimated y-intercept and its variance. But what about the mean errors and r^2?

The mean total squared error $\langle \mathbf{totSE} \rangle_n$

$$\langle \mathrm{totSE} \rangle_n \equiv \frac{1}{\mathrm{sivar}_n} \sum_{i=1}^{n} \Big(\frac{y_i - \langle y \rangle_n}{\sigma_i} \Big)^2 = \langle y^2 \rangle_n - \langle y \rangle_n^2 \,.$$

The mean squared residual error $\langle \mathbf{resSE} \rangle_n$

In order to express $\langle \mathrm{resSE} \rangle_n$ in terms of the various weighted sums, we need first to get rid of the dependency on a_n^2, b_n^2, and $a_n b_n$ resulting from expanding $[(a_n x_i + b_n) - y_i]^2$,

$$\langle \mathrm{resSE} \rangle_n = \langle x^2 \rangle_n \, a_n^2 + b_n^2 + \langle y^2 \rangle_n + 2 \langle x \rangle_n \, a_n b_n - 2 \langle xy \rangle_n \, a_n - 2 \langle y \rangle_n \, b_n \,.$$

We can do that by using the two linear equations resulting from setting the partial derivatives to zero,

$$\begin{aligned} \langle x^2 \rangle_n \, a_n \; + \; \langle x \rangle_n \, b_n \; &= \; \langle xy \rangle_n \\ \langle x \rangle_n \, a_n \; + \qquad b_n \; &= \; \langle y \rangle_n \,, \end{aligned}$$

multiplying the first by a_n and the second by b_n, then adding them up to obtain

$$\langle x^2\rangle_n\, a_n^2 + 2\langle x\rangle_n\, a_n b_n + b_n^2 = \langle xy\rangle_n\, a_n + \langle y\rangle_n\, b_n$$

so

$$\langle \mathrm{resSE}\rangle_n = \langle y^2\rangle_n - \langle xy\rangle_n\, a_n - \langle y\rangle_n\, b_n\,.$$

The mean squared regression error $\langle$regSE$\rangle_n$

Similarly,

$$\begin{aligned}
\langle \mathrm{regSE}\rangle_n &\equiv \frac{1}{\mathrm{sivar}_n} \sum_{i=1}^{n} \Big[\, \frac{(a_n x_i + b_n) - \langle y\rangle_n}{\sigma_i} \,\Big]^2 \\
&= \langle x^2\rangle_n\, a_n^2 + 2\langle x\rangle_n\, a_n b_n + b_n^2 + \langle y\rangle_n^2 - 2\langle y\rangle_n \Big(a_n\, \langle x\rangle_n + b_n\Big) \\
&= \Big(\langle xy\rangle_n - 2\langle x\rangle_n\, \langle y\rangle_n\Big)\, a_n + \langle y\rangle_n \Big(\langle y\rangle_n - b_n\Big)
\end{aligned}$$

■