Model fitting with non-normally distributed residuals

August 25, 2019

I recently saw a commentator on the internet attempt to show the link between increased IQ and increased earnings. While I agreed with his thesis, he used a simple linear regression, of which one of the main assumptions is that the residuals are normally distributed. As income is Pareto distributed, this assumption is clearly false. So I set out to derive a formula which would apply for models where the residuals are Pareto distributed. This has almost certainly been done by someone before, but it was fun anyway.

The classical Pareto distribution for a variable y is controlled by two parameters: the minimum value, y_0 and the slope parameter, α, and is given by:

$$\mathcal{P}(y) = \frac{\alpha y_0^{\alpha}}{y^{\alpha+1}} \tag{1}$$

We now extend this model and declare that α is now a linear function of a second variable x, i.e. $\alpha = m * x + b$ where m and b are constants. The equation then becomes:

$$\mathcal{P}(y|x, m, b) = \frac{(mx+b)y_0^{mx+b}}{y^{mx+b+1}} \tag{2}$$

Then, given a set of n observations of y and x: $\{y_i\}$, $\{x_i\}$ for $i \in \{1, ..., n\}$, we seek the maximum likelihood estimator of m and b. The likelihood $\mathcal{L}$ of the n observations is given by:

$$\mathcal{L}(m, b) = \prod_{i=1}^{n} \mathcal{P}(y_i|x_i, m, b) = \prod_{i=1}^{n} (mx_i + b) \frac{y_0^{mx_i+b}}{y^{mx_i+b+1}} \tag{3}$$

We maximise this by maximizing the log-likelihood instead:

$$\sum_{i=1}^{n} \log(mx_i + b) + (mx_i + b)\log(y_0) - (mx_i + b + 1)\log(y_i) \tag{4}$$

To do this, we compute the derivative of the likelihood w.r.t. m and b and search for a point where they are both zero.

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{n} \frac{1}{mx_i + b} + \log(y_0) - \log(y_i) \tag{5}$$

$$\frac{\partial \mathcal{L}}{\partial m} = \sum_{i=1}^{n} \frac{x_i}{mx_i + b} + x_i \log(y_0) - x_i \log(y_i) \tag{6}$$

I could not solve these equations analytically, and so had to settle for using the gradients to perform gradient descent on the training data. Note that extending this process to situations where α is an arbitrary function of x is trivial.