Lecture 4 Comments

Dorian Goldman

February 17, 2017

I wanted to give some updated answers to questions i received during lecture that I don't feel I answered optimally.

Question 1: What is np.logspace()?

Answer: np.logspace() creates numbers which are equally spaced on a *log scale*. In general when we are trying to find the optimal parameter size λ, we want to consider numbers which range over *different orders of magnitude*. For example, when minimizing

$$\mathcal{L}_\lambda(\beta) := \frac{1}{N}\sum_{i=1}^{N}(y_i - \beta^T \cdot x_i)^2 + \lambda\|\beta\|_{L^p}, \tag{0.1}$$

we generally experiment with scales which range λ over values such as $1, 10, 100$, etc - it provides a larger range of values from which to determine what is the best value (see the digit recognition example where $C = 1, 10, 100$ in lecture notes).
`https://docs.scipy.org/doc/numpy/reference/generated/numpy.logspace.html`

Question 2: Aren't the p values always the same?

Focusing on the 1d example, I represented the formula for the error in the coefficient estimate for the slope:

$$\mathrm{SE}(\beta_1)^2 \sim \frac{1}{N}\frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (x_i ? \bar{x})^2}$$

This formula *is only for the slope, and only in one dimension.*

In general if we are looking at multiple features (ie. $X \in \mathbb{R}^{n \times p}$), the error in our regression coefficients is normally distributed given by

$$\mathcal{N}(0, (X^T X)^{-1}\sigma^2)$$

Let's pick a simple example where

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}.$$

This is equivalent to solving $y = c + bx$ (a 1 d problem with an intercept - you can interpret one 'feature' as just 1s).
(ie we just have an intercept term). Then the variance above is computed as

$$(\mathbf{X}^T\mathbf{X})^{-1} = \frac{1}{n\sum x_i^2 - (\sum x_i)^2}\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix} \tag{0.2}$$

Here I've just used the 2d formula for inverting the matrix $X^T X$ (you can look that up).
Now we see from this that (using $y = c + bx$) so that $\beta = (a, b)$),

$$\sqrt{\widehat{\mathrm{Var}}(b)} = \sqrt{[\hat{\sigma}^2(\mathbf{X'X})^{-1}]_{22}} = \sqrt{\frac{n\hat{\sigma}^2}{n\sum x_i^2 - (\sum x_i)^2}}.$$

$$\sqrt{\widehat{\mathrm{Var}}(c)} = \sqrt{[\hat{\sigma}^2(\mathbf{X'X})^{-1}]_{11}} = \sqrt{\frac{\sum_i x_i^2\hat{\sigma}^2}{n\sum x_i^2 - (\sum x_i)^2}}.$$

So we see that even in this simple example, the variances differ. The denominator can be seen as the *spread* of the data, while the numerator is the spread of the response variable. You can interpret this as, the more range you have in your x values, the more confident you are in the coefficient estimate. Think about an example where y had a non-zero variance, but x_i was the same for every i - you would have no confidence.
In general for a matrix of data $\mathbf{X}$, the above matrix is more complicated, but this shows the point. The p values are calculated via

$$t_c = \frac{c - 0}{\sqrt{\text{Var}(c)}} \tag{0.3}$$

$$b_c = \frac{b - 0}{\sqrt{\text{Var}(b)}} \tag{0.4}$$

which we see from this example, depends crucially on the distribution of the feature. This number represents the z-score, or the number of standard deviations away from the mean your observation is. The p-value is just just the area under the curve of a unit variance, mean zero normal distribution (denoted $\Phi(t)$), starting at this point, $\Phi(t_c)$ and $\Phi(t_b)$.

Question: Doesn't regularization eliminate the need to even think about p values or cross validation? The question was asked in the context of "aren't we only looking at p values to throw away bad coefficients?".

Answer: In general, *regularization will give you the optimal coefficients for your model, but this doesn't answer the question of whether or not any given variable has a 'significant' impact on your model a priori.*

Consider Figure 0.1. This is an example of a plot of the top coefficients in a classification model (we will cover it next week) with the error bars representing the spread from the variance of the coefficients.
On the other hand, figure 0.2 shows the error when we range over 5 fold cross validation (sorry the figure scales appear different for the y axis). Both models have been optimized. Notice that the variable "Male" seems quite important but also has a large error range over cross validation, while "Associate-Academic" appears quite high (although not as much) but has a large spread, hence large p-value.

While it is most often the case that the top variables will have more 'confidence' (ie. smaller variance, p values), it's still important to understand the uncertainty in measuring the coefficient.

People may not just want to know what the best model is, but rather, how any one given variable impacts the outcome, and with what confidence.

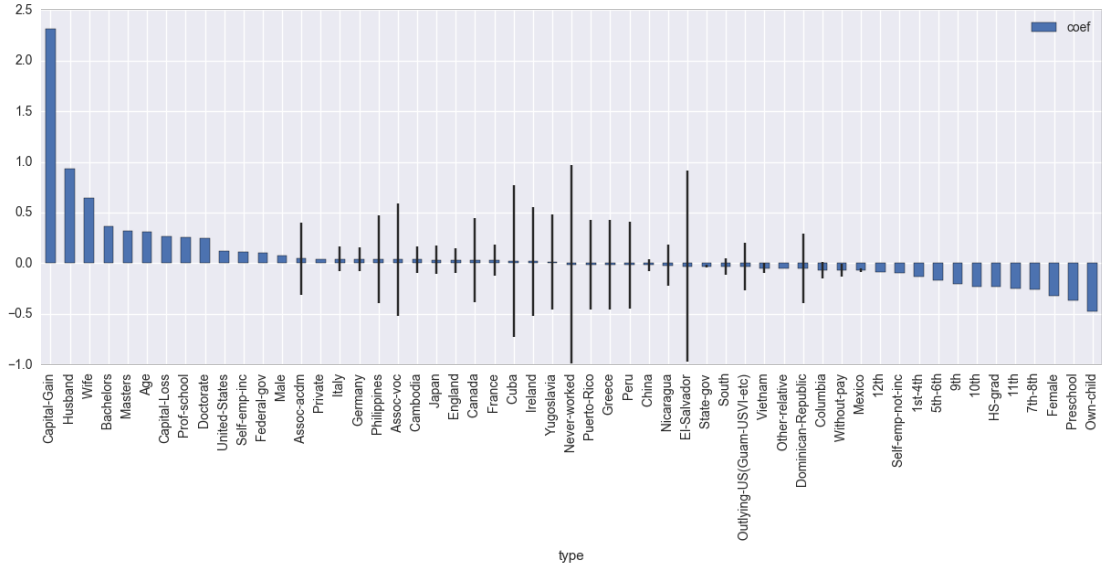

Figure 0.1: Coefficient values and corresponding error ranges from standard error of coefficients.

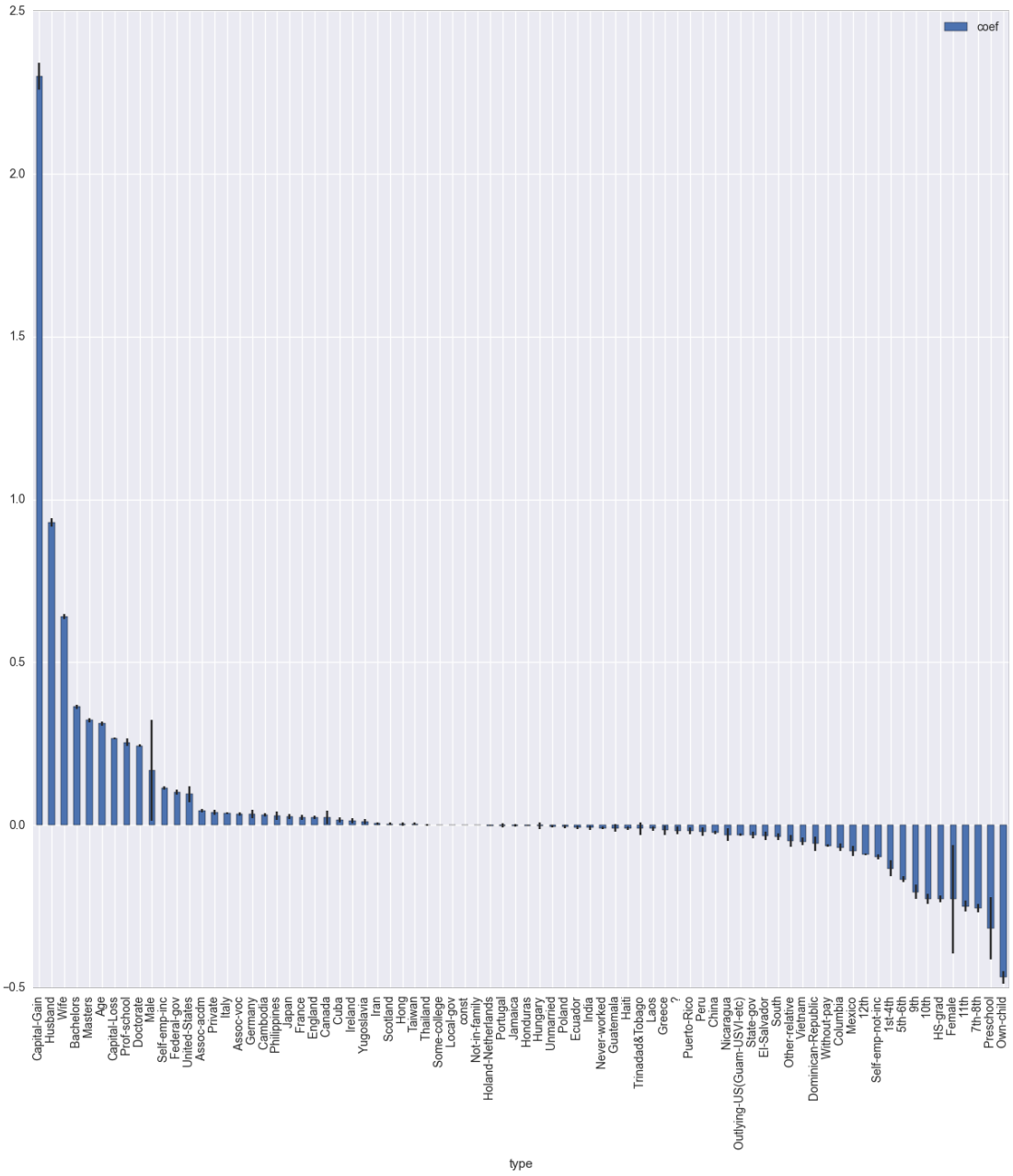

Figure 0.2: Coefficient values and corresponding error ranges from cross validation averages