Probabilistic programming

Vid Stropnik, 63200434

Introduction

As part of this assignment, we were given data about the amount of resources 50 successful startups have invested in different sectors. Based on this information, we want to understand the most prosperous investment strategy. Furthermore, through the use of probabilistic programming, our goal is to understand potential differences in spending between the three states, wherein the data was sampled. All conclusions were achieved using the *Stan* probabilistic programming language and the work is reproducible and available on Github.

I. Setup and assumptions

For this analysis, we make some core assumptions: the data are sampled without error, the net spends on individual sectors are assumed to be non-multicolinear. Another assumption at the heart of the linear model is that there is some multilinear dependence between the sector net spend and a company's profit. More formally, this relation is assumed as

$$\beta_r * \mathcal{R} + \beta_m * \mathcal{M} + \beta_a * \mathcal{A} + \mathrm{n}, \tag{1}$$

where $\mathcal{R}, \mathcal{M}, \mathcal{A}$ denote the research, marketing and administration spend respectively, and n corresponds to the intercept. The final assumption of our model is that the observed profits are independent and homoscedastic around some mean. Due to the Central Limit Theorm, we assume that they are sampled from the normal distribtuion. We also put normal priors with sufficiently large variances on our parameters, while the longer tails of the half-cauchy prior are used for the variance before the MCMC process.

II. Composite relations of Investment sectors

The results of the MCMC process are shown in Table I. We notice that the weight of the research spend is by far of the largest magnitude, while β_m, corresponding to marketing, is the smallest. The size of β can be interpreted as importance of that parameter, when predicting profit with our multilinear model. We must me cautious, however, to not interpret the per-dollar spend in marketing to be the least important, as the horizontal axes of the top subplots in Figure 1 show that overall marketing spend was of a different order of magnitude. This naturally induces a smaller value of β_m.

III. State differences

The differences in β values between the three states are shown in Figure 2. In it, we can notice that the spend in the administrative sector was more instrumental in the state of New York, when compared to California and Florida - with a significant probability of being more important than research spend. New York's latter sector has the highest importance variance among the three states, while the distribution of marketing appears to remain consistent throughout the country.

TABLE I
The prior and posterior distributions of model variables, where N denotes the normal and HC the Half-Cauchy.

Value	Prior	Posterior
Intercept	$N(0, 20)$	$N(0.2179, 0.2987)$
β_r	$N(0, 20)$	$N(0.7187, 0.0014)$
β_m	$N(0, 20)$	$N(0.0818, 0.0005)$
β_a	$N(0, 20)$	$N(0.3277, 0.0005)$
σ	$HC(0, 20)$	$HC(13788.92, 23.35)$

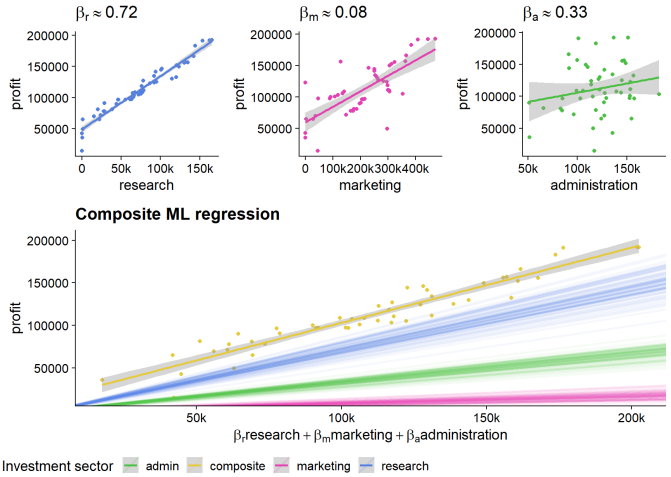

Fig. 1. The figure shows linear models fitted to all three investment sectors. The bottom figure shows the model (in yellow) fitted in composite space, calculated by our probabilistic programming approach. The blue, green and pink beams each correspond to lines, governed by the 100 beta samples from our final distributions.

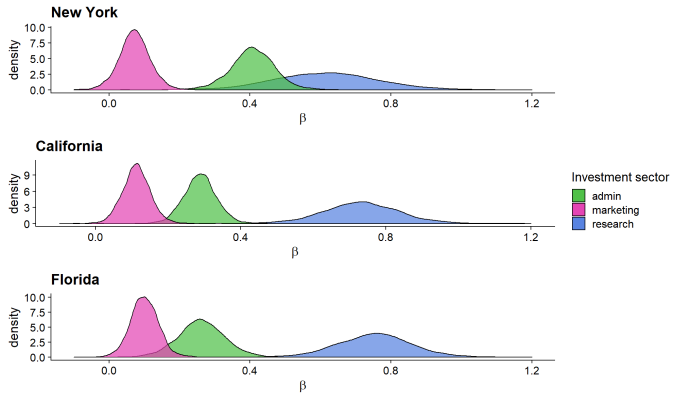

Fig. 2. The distributions of each linear weight in the multilinear model. The three subplots correspond to different states where the observed businesses are located.