Probabilistic programming

Urša Zrimšek

INTRODUCTION

In this report we show what would be our strategy if we were owners of a startup in the USA and would have to decide how to distribute our resources and where to open our offices. Our dataset contains information about how many resources 50 startup companies used for research, administration or marketing along with the location of the company and their profit. We used Bayesian inference to build a linear regression model to find out where we should have our offices and a multiple linear regression model to gain insight into how we should distribute our resources to maximize the profit. When building each model we performed basic diagnostics: inspected the traceplots, $\hat{R}$ and effective sample sizes.

COUNTRY SELECTION

To select **in which country we would open our office** (we have data for New York, California and Florida) we first built three linear regression models (one for each state), that predict the profit of the company based on summed investments in all sectors. We didn't use any intercept, since without any investment, we probably can't have any profit. The parameter infered from the model tells us the ratio between investment and the profit, and from it we can conclude in which state we can make the most out of our money. On figure 1 we can see the distributions of the ratio parameter for each state. From the
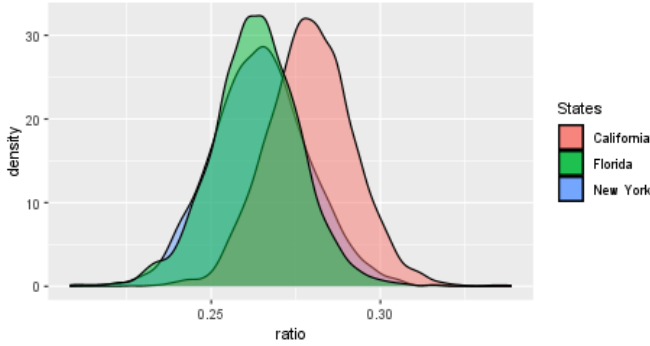

Figure 1. **Posterior distribution of ratio between investment and profit for each state.**

visualization we can see that the ratio is best in California, and that it is similar in New York and Florida. To be more precise, we calculated Monte Carlo standard error, and concluded that we can claim with **68% certainty** that opening our office in **California** is better than in New York or Florida.

DISTRIBUTION OF RESOURCES

To gain some insight into the optimal distribution of our resources, we built three multilinear regression models, again one for each state, all without intercept - for the same reason as above. We limited coefficients with a lower bound 0, because we believe that an investment in one of these three sectors can't have a negative impact on profit. On figure 2 we can see visualizations of the infered parameters.
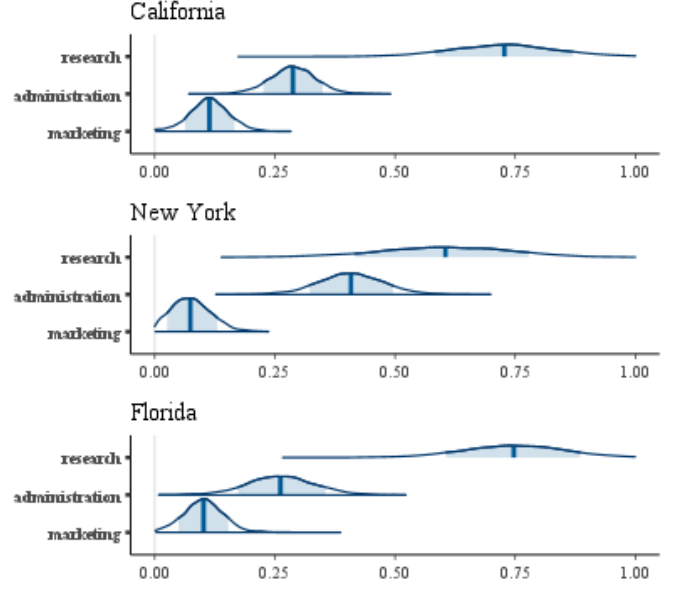

Figure 2. **Distribution of coefficients of sector investments for each country.** Above we see the distributions of the inferred parameters, their means and 80% probability mass intervals. The ordering of the sector coefficients is the same for all states.

We can further calculate (from ratio between coefficients) that in California, our state of choice, the money put in research is 11.7 times more valuable than the money put in marketing, and 2.6 times more valuable then put in administration. Standard errors for these two estimation are 1.3 and 0.02.

To get the optimal values of investment ratios, we calculated the ratios between samples of the coefficients. The numbers for all states are written in table I.

Table I
OPTIMAL INVESTMENT RATIOS.

State	**California**	New York	Florida
Research	**0.64** $\pm$ 0.002	0.55 $\pm$ 0.003	0.67 $\pm$ 0.002
Administration	**0.26** $\pm$ 0.001	0.38 $\pm$ 0.002	0.24 $\pm$ 0.001
Marketing	**0.1** $\pm$ 0.001	0.07 $\pm$ 0.001	0.09 $\pm$ 0.001

DISCUSSION

First we would like to point out that this analysis was made on small amount of data, which means that it has a lot of bias. We could improve that by gathering more data. Second problem arises at the optimal investment ratios. The ratio we suggested is far from any ratio of the companies in out dataset, where there are only few companies that invested more in research than in marketing, only two from California, with the biggest ratio between two being approximately 1.4, not more than 6, as we suggested. There could be a minimal ratio that needs to go into marketing, otherwise a company wouldn't be successful, and we could have made a mistake ignoring that. We should again gather more data and try to analyze this.