# Probabilistic programming

Žiga Trojer 63200440

## I. Specifying the model

We defined two models. The first model (A) is designed to identify the relationship between resources invested in research, administration and marketing and the company's profit. The second model (B) was designed to identify in which State to start the startup. Table I describes both models, where $\beta_1$ represents slope parameters for monetary features and $\beta_2$ for state locations. We should mention how we constrained $\beta_2$ in model B - we assumed that startups do not make profit larger than $1e+08$.

Table I: Models description

| model | A | B |
|---|---|---|
| likelihood | $y \sim \mathrm{N}(X\beta_1, \sigma_1)$ | $y \sim \mathrm{N}(X\beta_2, \sigma_2)$ |
| intercept | without | without |
| prior | $\beta_1 \sim \mathrm{Cauchy}(0, 5)$ | without |
| constraints | $0 \leq \beta_1$, $0 \leq \sigma_1$ | $0 \leq \beta_2 \leq 1e+08$, $0 \leq \sigma_2$ |

## II. Data description, preparation & diagnostics

We were given data about 50 startups across 3 states in the USA. It contains 4 independent variables (*research*, *administration*, *marketing* and *state*) and 1 dependent (*profit*). There are no empty values in the data.

First, we changed categorical variable *states* to one-hot-encoding variables. We then graphically analyzed the correlation of independent variables (only for monetary variables). The highest correlation is present between *research* and *marketing* (0.72), so we did not remove any variables, as correlation is below 0.9.

## III. Fitting the model

We fit both models where we used the default settings. It was important to perform some posterior checks to ensure that the model sufficiently converged. First we observed the $\hat{R}$, which was 1 for all parameters and the sample size was large enough. We then diagnosed the trace plots to confirm that everything had converged and that the chain trajectories looked correct.

## IV. Results

We now have 2 models from which we can figure out how to invest our start-up capital to maximize profits. We will also try to identify the state in which it is most worthwhile to start a startup. Let's start with how to invest our money. Table II shows summary data for $\beta_1$ coefficients for the first model. We can clearly see that coefficients make sense, as they are positive. For example, if we would invest 1 unit of money in *research*, we would gain 0.718 units of money on average. Similar for other coefficients $\beta_a$ and $\beta_m$. Instead, we can compare those coefficients to see, in which state we would like to start, so we would gain more money. Higher the coefficient, better the choice. But we need to be careful about standard deviation, as coefficients are very similar.

Table II: Summary of $\beta$ coefficients

| variable | mean | q5 | q95 |
|---|---|---|---|
| $\beta_r$ | $0.718 \pm 0.063$ | 0.608 | 0.828 |
| $\beta_a$ | $0.328 \pm 0.030$ | 0.274 | 0.382 |
| $\beta_m$ | $0.082 \pm 0.023$ | 0.045 | 0.119 |
| $\beta_{NY}$ | $101708 \pm 10579$ | 84851 | 119362 |
| $\beta_{CA}$ | $116684 \pm 10895$ | 98615 | 134122 |
| $\beta_{FL}$ | $112133 \pm 10428$ | 95011 | 129095 |

To determine the proportion, how to distribute our money, we need to calculate the proportion of coefficients. Table III shows the recalculated ratio of $\beta$ coefficients for monetary variables. From it, we can determine how to invest our money - we should invest around 64% in *research*, 29% in *administration* and 7% in *marketing*. The graph 1 is showing the optimal investment ratio with the 95% confidence interval.

Now we have to determine the state in which it is most worthwhile to start the startup. As we can see from the graph 2, it's a little better to start a startup in *California*, but we're not sure, as confidence intervals overlap. That's why we calculated $\mathrm{P}(\beta_{CA} > \beta_{FL}) = 0.64$ and $\mathrm{P}(\beta_{CA} > \beta_{NY}) = 0.84$. So we can claim that with 84% probability, we would choose to start the startup in *California* instead in *New York* and with 64% probability, we would choose *California* instead of *Florida*.

Table III: Ratio $\beta$ coefficients

| | mean |
|---|---|
| $\beta_r$ | $0.636 \pm 0.064$ |
| $\beta_a$ | $0.291 \pm 0.032$ |
| $\beta_m$ | $0.073 \pm 0.022$ |

Figure 1: Optimal investment ratio

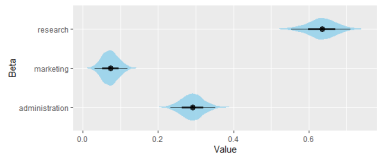
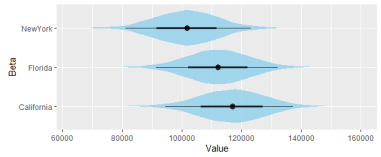



Figure 2: Expected profit based on state choice




## V. Conclusion

If you had to choose how to invest and in which state to start a startup, we would opt for California, as the results dictate, and Silicon Valley is also located there, which is also known for the success of startups. The funds would be distributed as follows: 64% *research*, 29% *administration* and 7% *marketing*.