02 - Probabilistic programming

Matej Miočić

Introduction

We developed 2 multiple linear regression models using Stan. We used the models on *50_startups.csv* dataset containing 50 instances of investments in *Research*, *Administration* and *Marketing* along with the location of the company and their profit. We used the models to gain insight into how we should distribute company's resources and where the company should open their offices to maximize the profit.

Methods

We used multiple linear regression (see Listing 1) with the intercept and uniform prior (**model_1**) to gain insight into where the company should open their offices. We used One-hot encoding technique to transform states into binary data. We removed one state to avoid multicollinearity. We did not use the intercept for resource distribution (**model_2**) since in our case without any investments (independent variables) there should probably be no profit (dependent variable). We used Cauchy as our prior with parameters 0 and 2.5. We checked all parameter outputs using stan and everything made sense.

Listing 1. Snippets of Stan code for both models.

```
parameters {
  real a;               // intercept
  vector[k] b;          // slope
  real<lower=0> sigma; // stdev
}
model_1 {
  // stan default prior used
  y ~ normal(a + X * b, sigma); // model
}
model_2 {
  b ~ cauchy(0, 2.5);        // prior
  y ~ normal(X * b, sigma); // model
}
```

Results

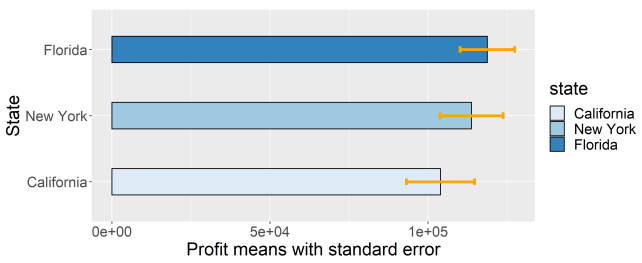

Figure 1. Profit means and their standard errors based on companies' state.

We started the analysis by comparing profits based on location of the companies. In Figure 1 we show profit means with standard error of company profits put together from the location of the companies.

We can see that the mean value of profits in **Florida** is a bit higher than the other two locations, but since we have so little data – the standard error is very high, we cannot make a certain claim that opening more offices there would be more profitable.

Then we analysed the optimal distribution of company's resources for every location. In Figure 2 we show the ratio between samples for β coefficients for *Research*, *Marketing* and *Administration* in **Florida**, **New York** and **California**.

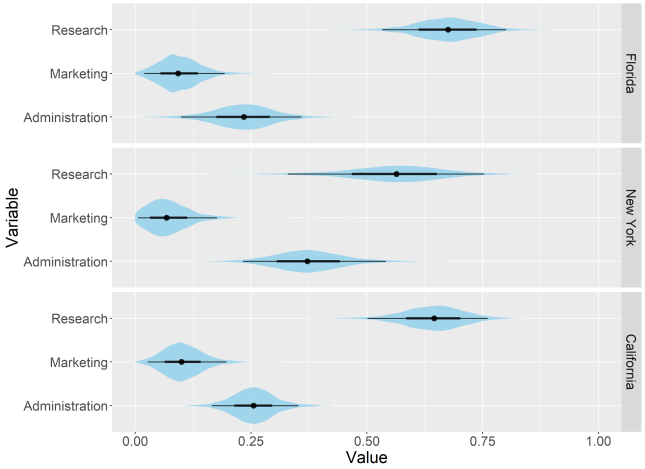

Figure 2. Ratio between samples for β coefficients for *Research*, *Marketing* and *Administration* in **Florida**, **New York** and **California**.

Florida and **California** share a similar distribution of resources where *Research* is the dominant resource for profit. Whereas a bit more resources should be put to *Administration* in **New York** even though *Research* is still the most profitable resource.

Conclusion

We constructed optimal resource distribution ratios for companies in **Florida**, **New York** and **California**. We discovered that *Research* is the most profitable resource for all locations. We also explored where the company should open their offices to maximize their profit but could not make a certain claim due to insufficient data.