Maj Gaberšček, 27212075

HOMEWORK 06: report

BASIC INFORMATION

I implemented this homework in Python. I used `pymc3` library for MCMC based Bayesian inference (as it was similar to `rstanarm::stan_glm` in R). The data was firstly normalized (so I could compare distance and angle's coefficients).

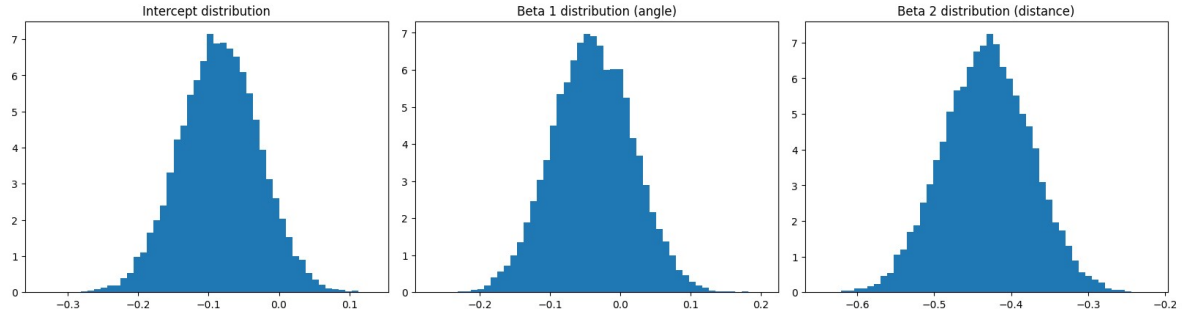
QUESTIONS

Before looking at any of the results, what is your personal opinion about the coefficient beta for distance? State it in probabilistic terms (as a distribution). Discuss after observing the results.

With my basketball knowledge, coefficient beta for distance should have a negative mean (as common knowledge suggests, that further the throw, lesser the chance of success). So, posterior distribution for beta 2 (beta for distance) should be normally distributed and $N(\mu, \sigma^2)$, where μ should be negative.

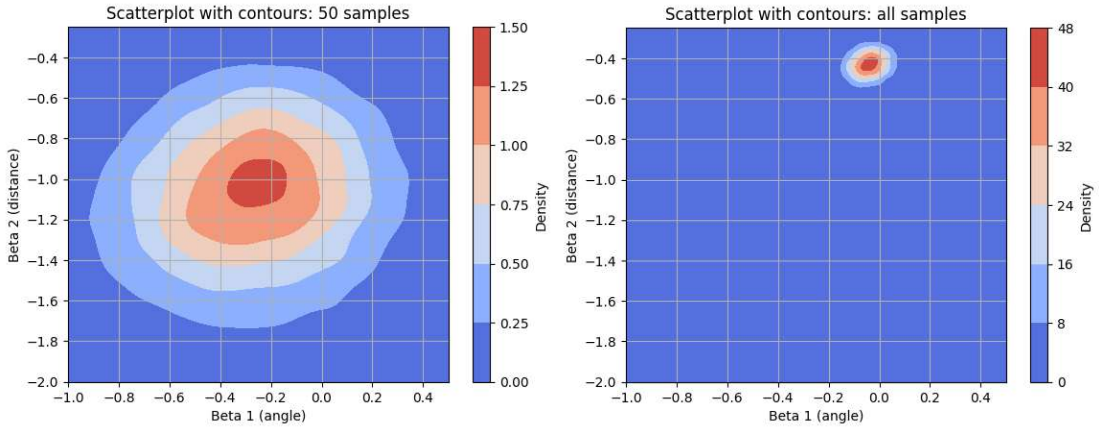
After observing the results, we get posterior distributions for coefficients and distribution plots:

coefficient	mean	standard error	95% CI
beta 0 (intercept)	-0.08	0.056	[-0.20, 0.03]
beta 1 (angle)	-0.04	0.057	[-0.15, 0.07]
beta 2 (distance)	-0.43	0.057	[-0.54, -0.32]



We can assume (from the distribution shape), that given mean μ and standard deviation σ, betas are distributed normally, $N(\mu, \sigma^2)$. So increasing distance indeed contributes towards throw not being successfully hit. We can say that for sure, as upper limit of the confidence interval still lays firmly below zero.

Plot a scatterplot with contours of the posterior samples of the angle and distance coefficients. Repeat the process for a smaller subset of only 50 shots (sampled at random). Discuss the similarities and differences.



As we can observe on the above contours, posterior sampling is much more concentrated, when using a large learning dataset as opposing to learning on only dataset of 50 random samples. This is because larger dataset provides more information about the relationship between the predictors and the response variable. As the sample size of data increases, the posterior distribution becomes more informed and reflects the underlying patterns in the data. With a larger dataset, the estimated coefficients also tend to have lower variance and higher precision, as can be seen from contour plots.

Formulate each of the following as a probabilistic question and estimate an answer based on the posterior using all available data:

- *Which is more important for shot success, angle or distance?*

If we compare beta coefficients for distance and angle (their magnitude), we can clearly see that distance contributes far more then angle. In fact, from observing the confidence interval for angle, we cannot even say for sure (in regard to confidence interval), whether increasing angle increases or decreases shot success probability.

- *Does shot success increase or decrease with increasing angle (the further on the sides we are)?*

As said above, it seems that increasing angle slightly increases chances of success, although we cannot say for sure (as 95% confidence interval also spans above 0.