03 - GLM

Matej Miočić

Introduction

We developed a gamma regression model to research how *genre* and *platform* influence the game's *sales* using stan. We used the model on *videogame_sales.csv* dataset containing 379 *sales* based on what *genre* and *platform* the games were sold for. First we looked into how *platform* affects *sales*, then we looked into *genres* and finally we looked into how combining these 2 features changes the outcome.

Methods

We used gamma regression with intercept (see Equation 1) with stan's default prior. Our target variable *sales* (y), resides on the positive side of real numbers, so we model y as:

$$\mathbf{y}|\mu,\sigma \sim gamma(\frac{\mu^2}{\sigma},\frac{\mu}{\sigma}), \tag{1}$$

where $\mu = e^{\alpha+\beta_1+\dots\beta_n}$ and σ is a parameter.

For the first part we one-hot encoded instances based on *genre* and removed **role-playing** to get rid of collinearity. Then for the second part we one-hot encoded instances based on *platform* and removed **PC**. For the final part we joined all possibilities of *genre* and *platform* and one-hot encoded them. We then removed the **Xbox** and **role-playing** combination, because it had the least instances. For every part we checked the diagnostics on our fits and they gave us no reasons for concern, even though we had warning at the warmup phase.

Results

We started the analysis by comparing *sales* based on *genre* of the game. In Figure 1 we show obtained parameters, *sales* mean based on parameters and data. We show that according to our model both *genres* are comparable, even though the means of data say that **shooter** *genre* is more profitable.

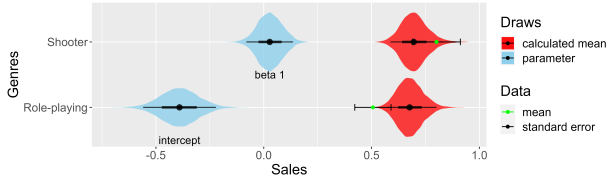

Figure 1. Parameter draws (blue eye), calculated mean with parameters (red eye) and data mean (green dot) with standard error (black error bars) for *sales* based on *genres*.

Then we compared *sales* based on *platform*. Similarly we show parameters and means obtained by the model and our data. In Figure 2 we show that the difference between model mean and data mean increases. We suspect that the gamma model might not be able to produce the best fit with the data we are given. However, both means agree that **Xbox** is the most profitable *platform* according to our data.

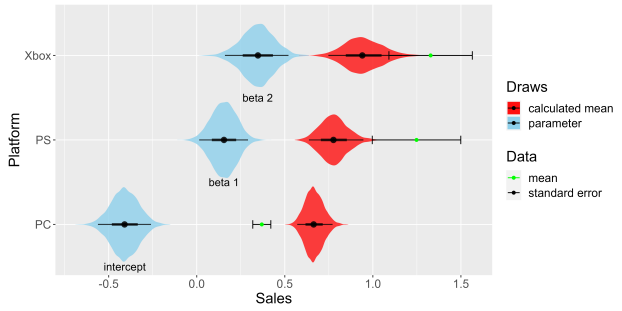

Figure 2. Parameter draws (blue eye), calculated mean with parameters (red eye) and data mean (green dot) with standard error (black error bars) for *sales* based on *platforms*.

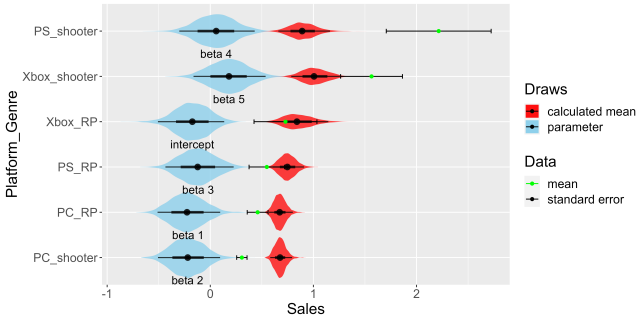

Figure 3. Parameter draws (blue eye), calculated mean with parameters (red eye) and data mean (green dot) with standard error (black error bars) for *sales* based on *genres* **and** *platforms*.

Finally we compare genres and platforms combined. In Figure 3 we show that according to our model **Xbox** and **shooter** combination is the most profitable, even though the data mean of **PS** and **shooter** says this combination creates the highest sales.

Conclusion

We researched how *genre* and *platform* influence the game's *sales*. We conclude that **shooter** *genre* is more profitable than **role-playing**. We also showed that **Xbox** is the most profitable *platform* for all *genres*. Combining *genre* and *platform* outputted different results. Our model says that **Xbox** and **shooter** combination produces the highest *sales*, meanwhile looking at the data means - we conclude that **PS** and **shooter** combination creates the highest *sales*. We also conclude that gamma regression model might not be the most suitable model for our data.