Generalized linear models

Urša Zrimšek

Introduction

We want to develop a successful video game. We used the video game sales dataset, that contain platforms (PC, PlayStation (PS) and Xbox), genres (shooters and role-playing games (RPG)) and sales of 379 different games. We used Bayesian inference to build a gamma regression model to research how genre and platform influence the game's sales. When building the model we performed basic diagnostics: inspected the traceplots, $\hat{R}$ and effective sample sizes.

Genre selection

To select **which genre generates more sales** we built a reparametrized **gamma regression** model, $y \sim Gamma(\frac{\mu^2}{\phi}, \frac{\mu}{\phi})$, that predicted sales values from genre - a column with zeros and ones, where 1 represents shooter. We also included intercept α, as it makes sense that the sales are not only dependent on the genre of the game. On α we put a Cauchy(0, 10) prior, and Cauchy(0, 2.5) on β. From the model we infer α and $\beta_{shooter}$, but since its value is transformed through the link function, we will get more information if we compare mean parameters μ:

$$\mu_{shooter} = e^{\alpha+\beta_{shooter}}, \mu_{RPG} = e^{\alpha},$$

whose posterior values can be seen on figure 1.

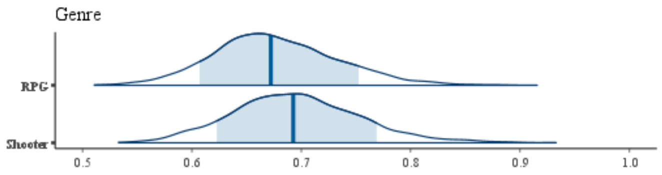

Figure 1. **Posterior distribution of mean sales for each genre.**

From the visualization we can see that the expected mean sales are higher for shooter games. To be more precise, we calculated Monte Carlo standard error, and concluded that we can claim with **67% certainty** that developing **shooter game** is better than developing role-playing game. The expected sales for shooter are 0.018 ± 0.001 higher than for RPG.

Platform selection

For selection of **best platform** we approached similarly as for genre. We used intercept with prior Cauchy(0, 10), and two columns that represented dummy variables for the three platforms with Cauchy(0, 2.5) priors. With reparametrized **gamma regression** we infered α, β_{PS} and β_{xbox}, and from them calculated mean parameters:

$$\mu_{PS} = e^{\alpha+\beta_{PS}}, \mu_{xbox} = e^{\alpha+\beta_{xbox}}, \mu_{PC} = e^{\alpha}.$$

Posterior values of these parameters are shown on figure 2.

Again we already see from visualization that the platform with highest expected sales is **Xbox**, if we calculate MCSE we see we can claim with **97% certainty** that it is better than both other platforms. Expected sales for Xbox are 0.277 ± 0.001 higher than for PC and 0.166 ± 0.002 higher than for PS.

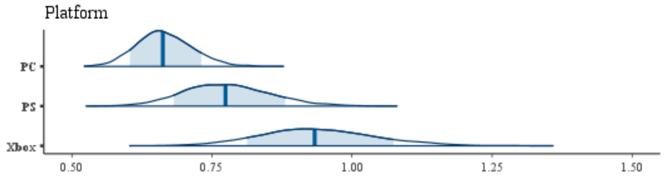

Figure 2. **Posterior distribution of mean sales for each platform.**

Discussion

Let's reason our prediction with data analysis. We found out (with some uncertainty) that the the better genre is shooter game, and best platform is Xbox. *But does that really mean that the best combination is shooter game on Xbox?* To answer that question, we plotted the distribution of sales in our data in figure 3. It looks like the sales for PS shooter are bigger than for Xbox shooter. Let's look at the means of datapoints: mean of all sales of Xbox shooter games is equal to 1.56 and mean off all PS shooters equals 2.22. Did we make a mistake when we separated genre and platform? To answer that, we made another model that combined both game properties. We made it in the same way as both above, and got similar results, that can be seen in table I. SE is between 0.001 and 0.002 for all μ's and s stands for shooter, X for Xbox and rp for RPG.

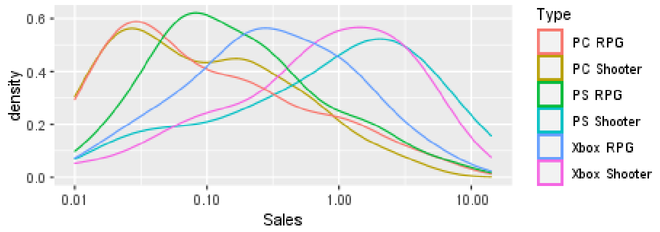

Figure 3. **Distribution of sales in our dataset, separated by both platform and genre.** The x axis is in log scale for better visibility, as the majority of sales values is very small.

Table I
Comparison of μ from combined model and the number of datapoints for game types.

type	s&X	rp&X	s&PS	rp&PS	s&PC	rp&PC
$\overline{\mu}_{type}$	0.97	0.91	0.81	0.77	0.68	0.65
# datapts	33	13	34	47	148	104

This combined model also tells us that we can claim with 94% certainty that **shooter on Xbox** yields bigger sales than shooter on PS. So we see that the model separation is not a problem. From the table we can see another explanation for this discrepancy – uneven number of datapoints for different categories. It is possible that our predictions for Xbox are better because more than 70% of games on Xbox are shooters, and only around 40% on PS. We could solve this problem by upsampling the underrepresented classes. The reason could also be the simplification with gamma regression, but to be sure, we would first need to eliminate the imbalanced data option.