GLM

Žiga Trojer 63200440

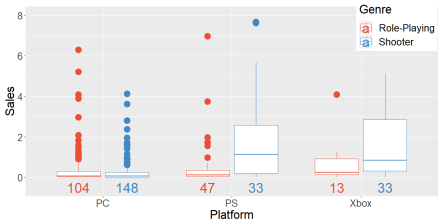
I. Introduction

In this homework, we investigated how the game's genre and platform influence its sales. Our goal was to identify the best platform and genre combination to maximize our sales.

II. Methods

We were provided with 397 records describing game sales across three platforms (*PC*, *PS* and *XBOX*) and two genres (*Role-Playing* and *Shooter*). Data contains 2 independent variables (*platform*, *genre*) and 1 dependent (*sales*). There is no empty values. Figure 1 shows the distribution of sales by platform and genre, as well as the number of data points for the combination of platform and genre.

Figure 1: Distribution of sales by platform and genre. Numbers below box plot shows the number of records for each category.



First, we changed categorical variables to one-hot-encoding variables. We removed one one-hot-encoded column for platform and genre due to potential multicollinearity. We then plotted the correlation of independent variables and found that it was low enough. We used 2 models, one for the *platform* and one for *genre* respectively. Table I describes the setting of both models, where we used only β_0 and $\beta_1 x_1$ for the second model's link (for *genre*). Here, β_0 stands for the intercept.

Table I: Model description.

likelihood	$y\|\mu,\phi \sim \text{gamma}(\frac{\mu^2}{\phi}, \frac{\mu}{\phi})$.
link	$e^{\beta_0+\beta_1 x_1+\beta_2 x_2}$.
transformed parameter	$\nu = \frac{\mu^2}{\phi}$.
prior	$\beta_1 \sim \text{Cauchy}(0, 10)$, $\beta_i \sim \text{Cauchy}(0, 2.5)$, $\nu \sim \exp(1)$.

We fit both models where we used 2000 iterations for the warm-up phase and 3000 for the sampling phase. We performed posterior checks to ensure that the model sufficiently converged. First, we observed the $\hat{R}$ and the sample size for all β_i. They indicated that sampling was successful, which was also confirmed when observing the trace plots.

III. Results

We can now calculate μ using the β_i coefficients we've obtained. We calculate it via the link function, provided in Table I. Figures 2 and 3 shows the posterior distribution of μ with mean and 95 % confidence interval, for *Platform* and *Genre* respectively. From both figures we already get some insights in which platform or which genre to choose for our video game. When we look at *Platform* distribution, we can see that they are very different from one another. Because we want to choose a platform that maximizes μ, we'll go with *XBOX*, which has the highest average value (0.830 compared to 0.752 and 0.684, SE less than 0.002) and a comparatively wide confidence interval. We see similarly from Figure 3, where we choose the *Shooter* genre (0.752 compared to 0.687, SE less than 0.002).

Figure 2: Posterior distribution of μ for the *Platform* with mean and 95% CI.

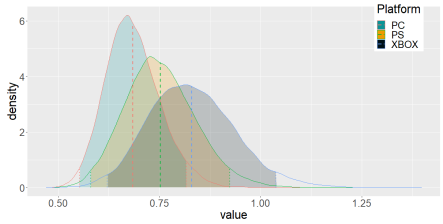
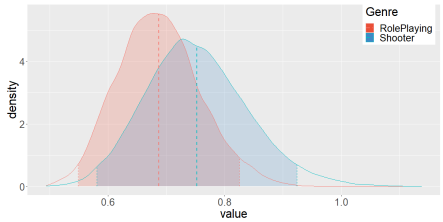


Figure 3: Posterior distribution of μ for the *Genre* with mean and 95% CI.



We want to confirm this. We calculated the probabilities that the proposed choice is really the best. Table II contains the estimated probabilities, along with the standard error.

Table II: Probability estimates for different choices.

probability	value	SE
$P(\mu_{XBOX} > \mu_{PS})$	0.825	0.004
$P(\mu_{XBOX} > \mu_{PC})$	0.982	0.002
$P(\mu_{PS} > \mu_{PC})$	0.907	0.003
$P(\mu_{Shooter} > \mu_{RolePlaying})$	0.719	0.006

IV. Discussion

As we have learned, it is most likely that we will make a higher sales from the game if we develop it as a shooter game for *XBOX*. If we compare statistics for PS shooters and *XBOX* shooters (see Figure 1), we see that above results deviate from the expected result. On average, *PS* shooters sold more than the *XBOX* shooters. If we examine Figure 1 closely, we can see that there are more outliers at *PS* platform, which have a large impact on the average with high sales.

Lastly, we wanted to check how likely the chosen combination is best. So we calculated the probability P(Shooter is best&XBOX is best) $= 0.574 \pm 0.007$. If I would be in a position to invest money in one of the games, I would not be yet persuaded in this decision. Maybe I would trust the statistics in Figure 1 more, especially if we had more data.