Forecasting Game Adoption Data Through Probabilistic Models

Executive Summary

This project aims to examine the adoption data from a popular video game released in 2014. Through the use of covariates, as well as the introduction of a spike at week 1 and a decaying factor, the finalized model – Burr XII with three covariates – aims the answer the managerial question: *By week 111 since its initial launch, how many players will adopt the game?* The paper then discusses the implications and future directions for using such models.

Background

In this project, we were tasked with illustrating appropriate models with sufficient rationales through a set of adoption data from a video game. The game, titled *The Binding of Isaac: Rebirth*, is a RPG (role-playing game) shooter game with a twisted theme at the root of its story. Players often describe the game as creepy, scary, and horrific – a great game to play during the Halloween season.

The data used were acquired by a Python-based crawler through Steam Community, an online gaming platform. It contained the weekly number of new adopters for a total of 111 weeks since its launch. For the purpose of this project, we will be using the first 61 weeks of data for model formulation, and the final 50 weeks for model validation.

Before Modeling

Upon first inspection of the data, before starting to build any model, I concluded that this is a timing data of continuous time with high heterogeneity. The key managerial question here is*: How many people have adopted the game by a given time after it was released?* Since people can choose to adopt the game at any time during the week (as opposed to only at specific instances in time), the data has continuous time. I also predicted that I would need to treat the number of adopters within the first week differently due to the high concentration of new users that week.

The Burr XII Distribution

The Burr XII distribution seems like a natural starting point. With the Weibull distribution at its individual level and the gamma distribution as the mixing distribution, the Burr XII distribution takes into account both the continuous time and the heterogeneity present in the data. I chose to start with this distribution, as opposed to the Pareto II distribution, because of my assumption that, due to the decreasing novelty of the game, the longer it had been since the initial launch of the video game, the fewer new adopters there would be within a given week. Upon fitting the model, my parameters were as follows:

	alpha	r	c	LL
Burr XII	353930	10489	0.346	-76187

While I did correctly predict the value of c to be less than 1, indicating negative time duration dependence, I was surprised to see high values of r and alpha, which indicate homogeneity. This led me to believe that these high values were due to the high concentration of players who purchased the game very early on (i.e. on week 1). I believed that, because there were so many “Fast Buyers” (comprising roughly 30% of all adopters), excel treated the group as homogeneous. Because of this, I decided to add a spike at week 1.

	alpha	r	c	p (first week)	LL
Burr XII Spike	381216	3416	0.58	0.025	-75880

Notice that, even though the value of r shrunk significantly, it is still very large. It is possible that the population is, indeed, homogeneous. However, since a model that allows for heterogeneity generally will not perform worse than ignoring heterogeneity(i.e. Weibull), I decided to keep this distribution, for now.

Introducing Covariates

It’s easy to observe that the data contains “bumps” where the number of adopters increases dramatically for a specific week. The reason for the poor fit of the previous model is the lack of covariates; factors that influence the number of adopters within a week. Correct identification of these covariates is vital to a successful model.

Covariate 1: Winter / Summer Sales

To those familiar with Steam: every year, Steam holds two major events: the winter and the summer sales. Lasting approximately two weeks each, these two “big sales” account for some of the biggest price dips for almost all video games. Gamers often target these two opportunities to purchase games at a lower price. A quick search of the sale dates provides the following chart:

	Start Date	End Date	# of Dates
Winter Sale 2014	Dec 18, 2014	Jan 2, 2015	16
Summer Sale 2015	Jun 11, 2015	Jun 22, 2015	12
Winter Sale 2015	Dec 22, 2015	Jan 4, 2016	14
Summer Sale 2016 *	Jun 11, 2016	Jun 24, 2016	14
Winter Sale 2016 *	Dec 22, 2016	Jan 4, 2016	14

*denotes predicted

In this model, the impact of the summer sale and winter sales are assumed to be equal (i.e. they have the same value for beta). Problems arise when we try to incorporate these covariates into the data. Since the data is recorded weekly, it is probable that a sale might end in the middle of a week. Such a sale might have a different (smaller) impact on that week’s count than if the sale was present for the entire week. One solution to this problem is to scale the covariate by the fraction of the week it lasted in a that week (e.g. 1 if the sale lasted for the whole week, 2/7 if the sale only lasted for two days of the week, etc.).

To project the future dates of these two major sales, I added one year from the start date of the summer sale 2015 and the winter sale 2015, respectively, to approximate the start date of the future sales. I then approximated the duration of each sale to be 14 days.

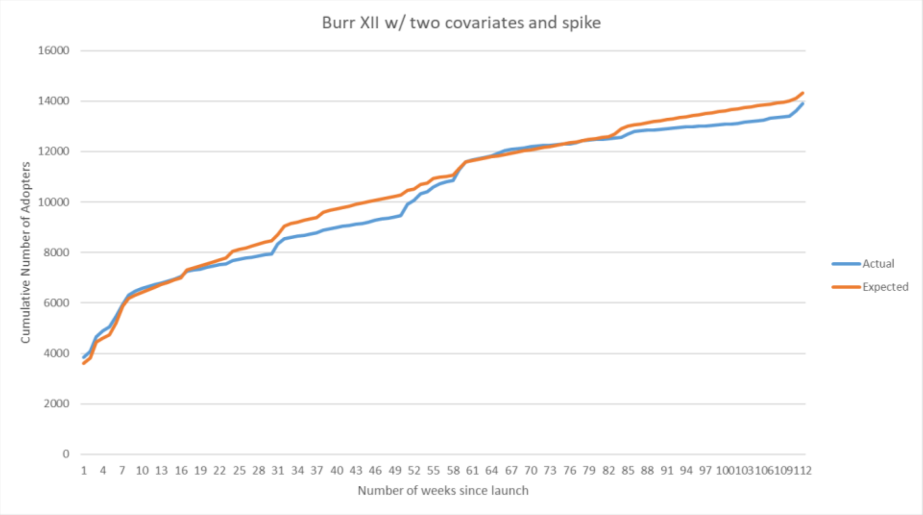
Covariate 2: Smaller sales that don't occur with a pattern (e.g. flash sales)

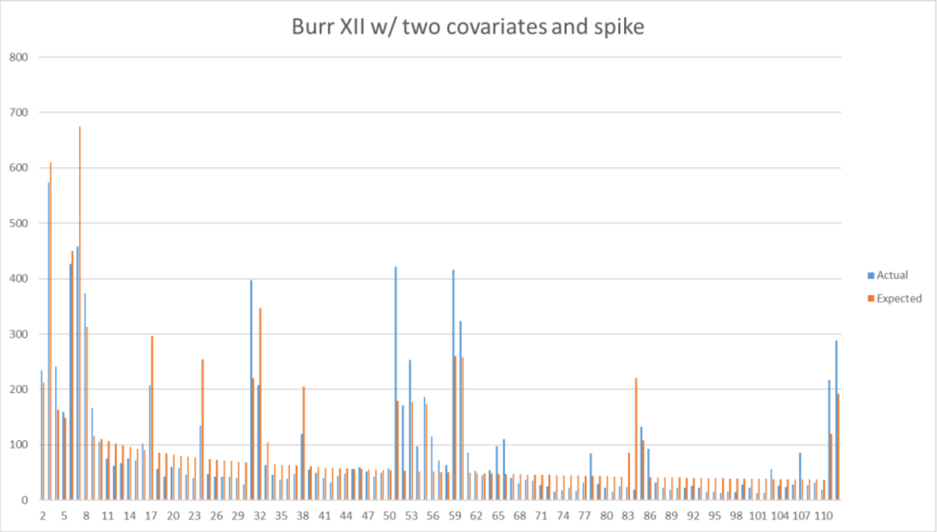
In addition to the annual winter/summer sales, Steam often promotes game-specific sales that often last for a short amount of time. Known as flash sales, these sales do not adhere to a repeating pattern of occurrence, but rather appear spontaneously. In order to track these sales, I found a dataset containing the exact dates of sales that the *Binding of Isaac: Rebirth* promoted during our observation period. While the exact date of each sale can be pinpointed using this dataset, the duration of each sale is not provided. Therefore, I made the assumption that most flash sales occur for the same length of time, so no scaling of dates was introduced for this covariate (i.e. 1 if there is a flash sale during the week, 0 if there isn't). Another assumption here is that flash sales do not overlap with the two major sales. That is, if a recorded sale date falls within either of the two major sales, I omitted this as a flash sale.

For the holdout period, I assumed that there were no flash sales. While it is obviously untrue, the unpredictability of the flash sales made it nearly impossible to predict when these sales would occur while forecasting. This is one limitation of using flash sales as a covariate.

Re-modeling after incorporating the two covariates into our dataset, we can see that we already have an "ok" fit of the data.

	alpha	r	c	p (first week)	B_big sale	B_flash sale	LL
Burr XII 2 Covariates w/ Spike	4078155	23121	0.572	0.028	1.652	1.209	-74073





*Note: On the incremental plot, I did not graph week 1, making the graph is more zoomed in, and making it easier to see the errors between each prediction. Week 1 would be a perfect fit, regardless, due to the spike.

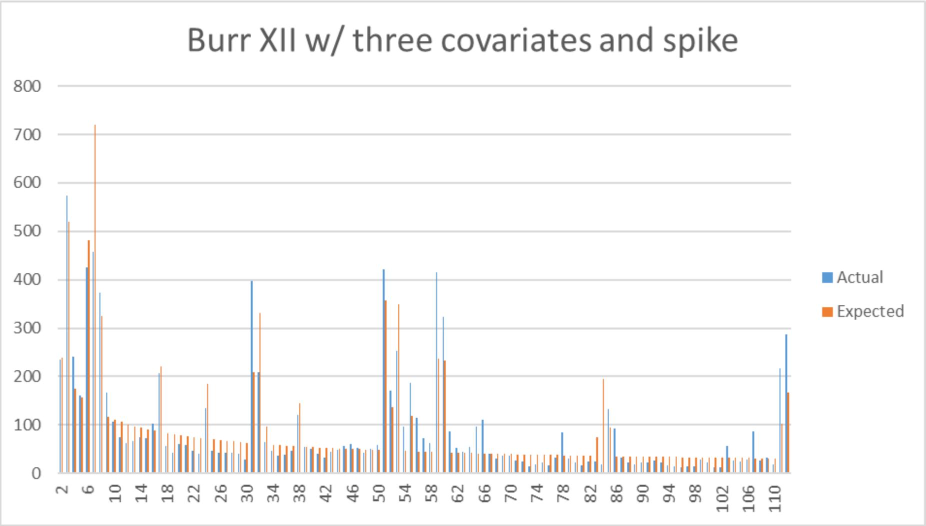
Both betas are positive, indicating that the existence of sales increases the number of new adopters. This makes logical sense. Again, r and alpha are still huge, implying homogeneity within the population.

Covariate 3: *The Binding of Isaac: Afterbirth*

Perhaps the biggest discrepancy between the actual and the expected was during week 52, which ran from 10/27/15 to 11/2/15. Coincidently, this was also the releasing week of the game's sequel – *The Binding of Isaac: Afterbirth*. I hypothesized that many new players would choose to purchase the two games in a bundle, and the publicity surrounding the new release would also result in an increase in the number of new players of the original game. This led me to add a new covariate (B_new) that models the increase in adopters within the first two weeks (where publicity is the heaviest) after the launch of the sequel game.

This took care of the spike at week 52.

	alpha	r	c	p (first week)	B_big sale	B_flash sale	B_new	LL
Burr XII 3 Covariates w/ Spike	3454372	27496	0.492	0.026	1.687	0.956	1.06	-73826



Decaying Sales

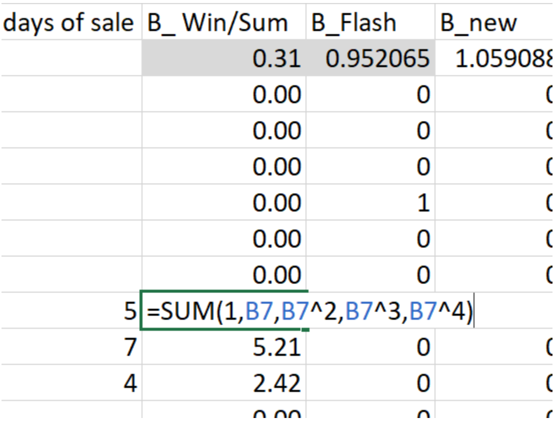
Focusing on the spikes where the highest error is present, I noticed that a majority of such "error spikes" occur when there's a winter/summer sale. Upon examination, an interesting observation prevails. Take a look at the 2015 summer sale as an example, which lasted from Jun 11 – Jun 22, a total of 12 days.

	Week number	Number of adopters	Days of the week that lie in the sale
6/9/2015	31	397	5
6/16/2015	32	208	7

Even though only 5 of the 12 days lie in week 31, the week has a lot more sales than the following week, which is a full week of sales!

It turns out that when there's a sale, especially a highly anticipated one such as the winter/summer sale, gamers who intend to purchase the game will do so as soon as the sale starts. Since the announcement of the annual sales is published as "big news" in Steam Community, it is safe to assume that gamers are well informed of the existence of the sale. This indicates that we can't treat this covariate as if it has a linear impact that only depends on the number of days (e.g. a sale that lasts two days has two times the impact as a sale that only lasted one day). Instead, we should have an exponential decay of the sale's impact to mimic the behavior of "first day purchases".

Using a new parameter called "decay factor" (d), I modeled the first day of the sale with an impact of 1, the second day of the sale with an impact of 1*d, third day with 1*d^2, and so on. The total impact with a week would be the sum of the impacts of each "sales day" in the week.



days of sale	B_ Win/Sum	B_Flash	B_new
	0.31	0.952065	1.05908
	0.00	0	0
	0.00	0	0
	0.00	0	0
	0.00	1	0
	0.00	0	0
	0.00	0	0
5	=SUM(1,B7,B7^2,B7^3,B7^4)		
7	5.21	0	0
4	2.42	0	0
	0.00	0	0

While the method of implementing the decay is slightly cumbersome, it yields a better model. Excel's approximates the decaying factor to be 0.96, confirming my theory that such decay exists.

	alpha	**r**	**c**	**p (first week)**	
Burr XII w/ 3 covariates, sales decay, and spike	3376872	26854	0.493	0.026	
	B_big sale	**B_flash**	**B_new**	**Decay factor**	
	0.305	0.952	1.059	0.963	

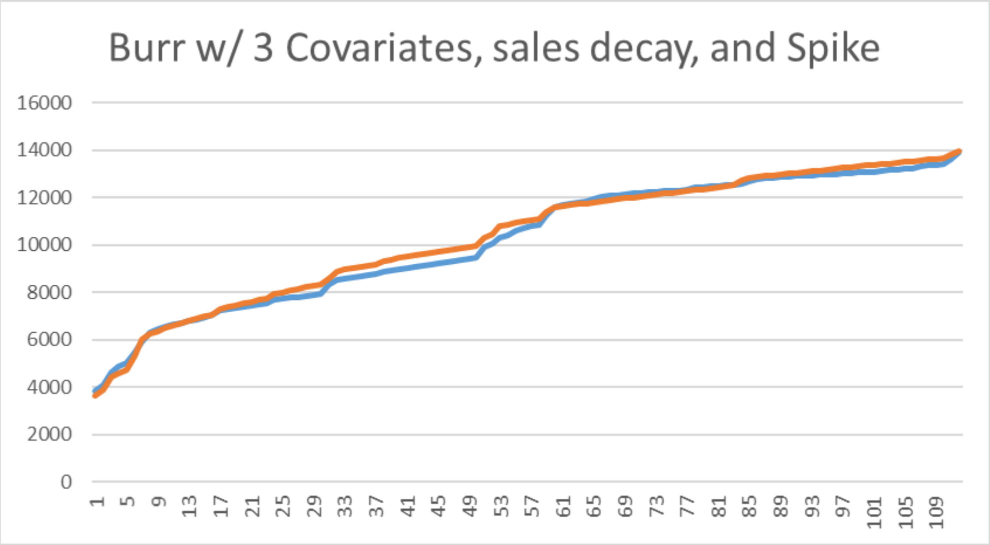
Final fit and summary

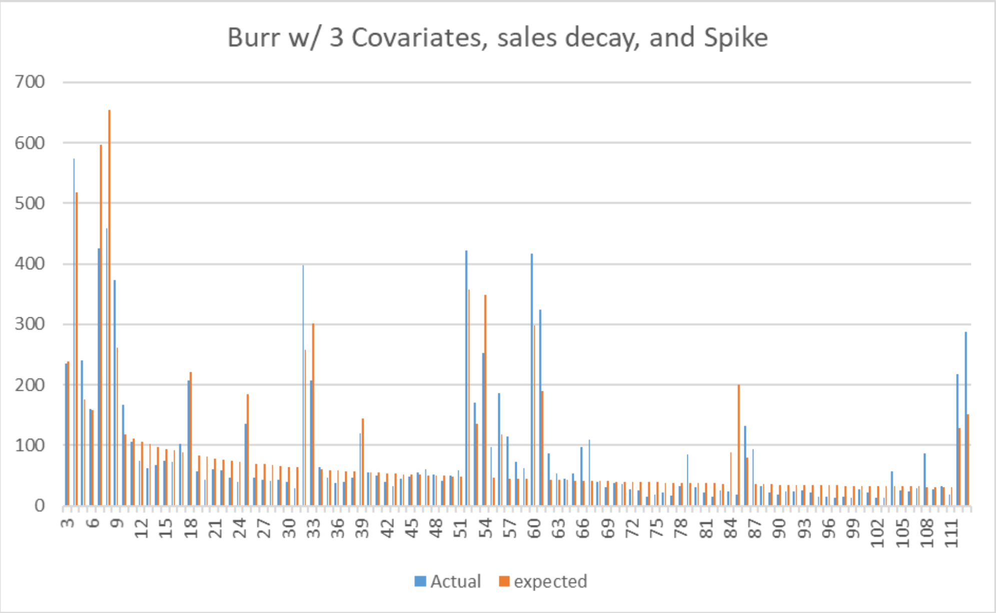
There are two possible reasons why the value of r is considerably high. First, it is possible that the population is, indeed, homogeneous. Second, the high r could be caused by covariates and c. That is, the differences between a gamer's propensity to purchase the game is less described by the heterogeneity of the population, and more by the different covariates. To validate whether the seemly high r is a result of true homogeneity or of influences from covariates, I decided to retreat from the Burr XII to a Weibull distribution to check its parameters.

The lambda for the Weibull distribution is 0.0078. Since the mean of a Weibull distribution is 1/lambda, it is calculated that on average, a gamer adopts the game after 128 weeks. This doesn't make logical sense, as our calibration period and modeling period, combined, total only 111 weeks (on average, they haven't bought the game yet!). This leads me to believe that the population is not truly homogeneous. The high value of r is purely a result of the use of covariates and c.

I also attempted to take away the spike at week 1, but decided against this when the fit became visually worse when no spike is introduced.

My final model is a Burr XII Distribution with 3 covariates, a decaying factor that simulates the decreasing impact of big sales covariate, and a spike at week 1.





	Number of parameters	LL	MAPE (median)	BIC
Burr XII	3	-76187	54.1%	152388
Burr XII Spike	4	-75880	42.3%	151777
Burr Spike 2 Cov	6	-74073	40.0%	148171
Burr Spike 3 Cov	7	-73826	33.3%	147680
Weibull Spike 3 Cov Decay	7	-73768	34.4%	147566
Burr Spike 3 Cov Decay	8	-73783	32.8%	147599

Future Models

While computing and fitting the sales decay factor, one assumption I made is that the decay rate is stationary. In this model, the impact of a sale that took place during the second week since initial launch has the same decay rate as the impact of a sale that started three years after the launch. I believe there should be time dependence on such decay. Similar to how games lose their "trendy status" as time goes on, it's possible that with time, fewer gamers would be "looking forward" to the sales events for the game. I believe that, as time goes on, the rate of decaying will decrease (i.e. the decaying factor will have a value closer to 1), but I have yet to come up with a way to model such behavior.

Managerial Problems

To answer the question: *How many players will adopt the game by 111 weeks after its launch*?, we just need to calculate the expected cumulative number. By our model, there should be approximately 14010 adopters by the end of week 111.

What's more interesting here is the sheer power of sales and promotions. As shown in the graph, a good sale has the chance of tripling, and even quadrupling the current adoption trend. From what I've read online, the sale's price of the game hardly ever dips under 50% of the original price. This means, as long as there are at least twice the number of adopters during the week of sales, the company obtains positive revenue. At the same time, too many sales might cause fatigue in potential buyers. If there's a sale every day, the feeling of the necessity of purchasing the game while on sale dwindles. Perhaps this is the true managerial problem, the true tradeoff: *How many sales should this company have?*

Sources

https://steamdb.info/app/250900/

https://store.steampowered.com/news/15308/

https://store.steampowered.com/news/17207/

https://store.steampowered.com/news/19813/

And all of my gamer friends who actually have a life outside of academics :)