

# DOES VIDEOGAME VIOLENCE REALLY WORK?

Georges Tippens  
9924293  
tippens@u.washington.edu

## EXECUTIVE SUMMARY

This paper analyzes the relationship between videogame sale revenues and factors such as ESRB rating, genre, release date, and others. I plan to examine if and how each property affects the total amount a videogame sells. For instance, one would imagine a game would sell more so than usual if it was a sequel because it implies the original game drew in enough profits to warrant newer release. What about other factors like the production company, genre, rating, or release date? Do they positively correlate with the total amount of games sold? With all of this in mind, I hope to be able to correctly forecast how well a game sells in the future.

The least squares regression shows that all variables are statistically somewhat important with the exception of possibly the holiday season. Genres and Ratings do determine how well a game will sell, with action games being very popular and racing games not so popular. Also it appears that games rated mature are a hot commodity, this is probably because videogames are marketed towards boys and young men who generally like blood and carnage.

The model also shows the three most important variables are if it is categorized as action, is a sequel of some sort, and if the publisher has released a top ten game. I admit that the latter variable is dependent because I consider the “Top Ten Games” as data points. This can be expected from intuition, because ‘action’ is a broad game category and the sequel generally is an improvement over the original game.

## DATA COLLECTION

The data collection was rather interesting. The NPD group provides sales and marketing information and releases a quarterly report on videogame console and game sales at the cost of \$1500, but luckily I was able to obtain this data from a videogame forum “<http://www.dkvine.com/interactive/forums/index.php?showtopic=445>”. I also used the site [www.ebgames.com](http://www.ebgames.com) to collect the ESRB ratings and genres of the games.

Important caveats about the data are as follows. For one, although some games are made for a variety of consoles, others are system specific. For example Halo has only been released for the XBOX platform. The playstation 2 has by far outperformed all other next-generation systems in units sold, so for the purposes of this model, I only consider PS2 games because it restricts the model to owners of the PS2. Another difficult problem to overcome is that games have a shelf life for as long as the system lasts. The data I have contains total sales figures from when the game was launched, not for a certain time period. To counter this problem I created a variable ‘age’ which counts the months a game has been available for purchase. This does not solve the problem completely because people are constantly becoming new owners of the console, and generally they will purchase newer rather than older games. For instance, someone who just bought a PS2 will probably buy Madden 2004 and not Madden 2001.

Along the same lines, it is difficult to distinguish one sequel from another. For instance, Madden football has been running since year 2000. To fix this I created two dummy variables, one which accounted for first year sequels, and the other which accounted for sequels beyond the first year. This is mainly because games which deserve

multiple sequels typically are such high caliber that the newer versions will sell extremely well.

## THE MODEL

Like most new things, games initially sell quickly and then as time passes begin to sell less frequently. Accordingly, I regressed with the log of the grossed units sold. The variables I took into account were: release date, the publishing company, the game ratings and genre, if it relates to another form of media, and if it was a sequel or a multiple sequel. I speculated that games would sell more during the summer and holiday season, mainly because games are generally targeted to young adults and children who have long summer vacations, and receive them as gifts for Christmas. I counted the holiday season as the month of December and the summer months as June, July, and August. I also created a dummy variable which determined if the publishing company had created a game which was one of the top ten sellers for the PS2.

The target audience of the game is also important. This is measured by both the ratings and genre. Action/Adventure games usually sell the best; whereas role playing games are targeted to a more peculiar group because the game takes more patience and are not as flashy (much like an Arnold Schwarzenegger film will fare better in the box office than a documentary about ants). To make the model simpler I constricted some genres into one, named 'OTHERS,' because they were the less popular genres and there were not enough data points to make the model realistic. The ratings given also affect how well a game sells. Parents are less likely to buy their child a game rated mature. Instead they will choose a tamer game which is rated for everyone or for teens.

Another important area is marketing. If a game is a sequel it signifies that the original did fairly well and with technological improves such as graphics, a sequel will fare well, especially if it is the third or fourth in the series. On the same level a game released that has ties with other forms of entertainment, be it television, books, or film would increase its marketability and desire. For example, a game like Harry Potter could have sub par game play and value, but because there is a phenomenon surrounding the books, product tie-ins like videogames perform well. Finally I factor how well the publishing company has done in the past. If they published a game that is one of the top ten sellers, then the publisher received a true value, and false otherwise.

## RESULTS

The model I began with was the following\*

$$\text{LNSOLD} = C(1) + C(2)*\text{AGE} + C(3)*\text{SEQUEL1} + C(4)*\text{SEQUELS} + C(5)*\text{MEDIA} + C(6)*\text{TOPTEN} + C(7)*\text{HOLIDAY} + C(8)*\text{ACTION} + C(9)*\text{SPORTS} + C(10)*\text{RACING} + C(11)*\text{EVERYONE} + C(12)*\text{TEEN}$$

The coefficients are as follows after running the regression:

$$\text{LNSOLD} = 12.84846353 + 0.01235728316*\text{AGE} + 0.3759376091*\text{SEQUEL1} + 0.3230889267*\text{SEQUELS} + 0.1960888902*\text{MEDIA} + 0.3794970072*\text{TOPTEN} + 0.09851808444*\text{HOLIDAY} + 0.3733052489*\text{ACTION} - 0.2331607868*\text{SPORTS} - 0.1741414043*\text{RACING} - 0.1056140306*\text{EVERYONE} - 0.1000451874*\text{TEEN}$$

The R-squared for this model is rather low at 0.377, but if you consider the outliers, like “GTA: Vice City” which sold close to six million copies in ten months, or “Enter the Matrix” which sold over 800,000 copies in one month my model holds pretty well.

The variable AGE is the only one that has a value other than true or false. It factors how long the game has been on the market, which would inevitably increase the number of units sold. The SPORTS, RACING, EVERYONE, and TEEN variables are

---

\* I decided to use the log of the quantity of games sold, because my sold variable did not follow a linear path when regressed with age and the other variables. Also it made viewing the scatter plot much easier.

negative because c(1) already accounts for a rating of 'mature' and a genre of 'other'. I had to do this so I would not receive a "near singular matrix" error because each game can only have one rating or genre type as a category.

I forecasted five games for which I had the sales figures and had excluded from my data. My model was incredibly close in forecasting the sales of "Nba Live 2002." It predicted a total sale of 742,544 copies whereas the actual figure was 737,187. The other numbers sadly were not as close. They ranged from being as far off as 350,000 copies (or a 22% over prediction) for "Bond: Agent Under Fire" to being an underestimate of 89,000 copies (or 14.3%) for "The Sims." I also forecasted the newest game in the NCAA football series and by my forecast it should reach 750,000 games sold within the first year. The previous year has already sold 940,000 copies in 11 months, so it appears there are a few kinks to work out.

## CONCLUSIONS

I was surprised the results came out as well as they did. I was half expecting the age coefficient wouldn't account the large discrepancies between the data points but the data shows it has a very low standard error. In Appendix A, I include the Scatter Plot with Regression of LogSold vs Age and you can see how widely dispersed it is. I only included two variables which controlled for sequels, but in reality some games like Grand Theft Auto was the fifth in the series. The data points gradually became smaller as sequels were made, as such it would not be realistic to create more than two variables which control for sequels. Another discrepancy may have been if the publisher published one of the ten top games in the past. This is because some publishers like Sega or Atari which are widely known and have been around for over 15 years have a mystique around

them that possibly could influence gamers buying their games, yet they have not published a top ten selling game.

In hindsight I should have split the genres into more dummy variables. Looking at the data it appears football games always sell better than basketball. My forecasts agree with this conclusion. When I predicted “NBA Live 2002” my forecast was off by a mere 5,000 copies, but it appears that my “NCAA Football 2003” forecast will be off by at least 200,000 copies. Both of the previous games have the same values for the dummy variables, but the differences in predictions were astonishing.

There are also inexplicable variables like marketing for which is impossible to find data. Also my model does not factor word of mouth if the game is quality or horrid. With these in mind distinctions I believe my model does a satisfactory job predicting total sales.

## APPENDIX A

Dependent Variable: LNSOLD

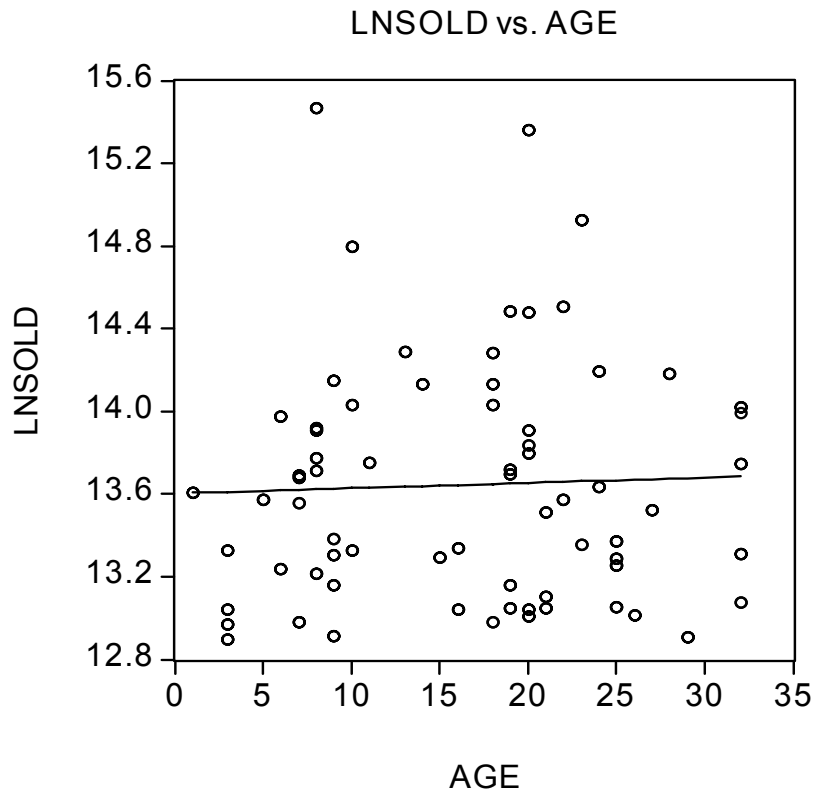
Method: Least Squares

Date: 11/12/03 Time: 14:34

Sample: 1 70

Included observations: 70

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.60272	0.151487	89.79446	0.0000
AGE	0.002534	0.008234	0.307700	0.7593
R-squared	0.001390	Mean dependent var	13.64413	
Adjusted R-squared	-0.013295	S.D. dependent var	0.578264	
S.E. of regression	0.582096	Akaike info criterion	1.783791	
Sum squared resid	23.04080	Schwarz criterion	1.848034	
Log likelihood	-60.43269	F-statistic	0.094679	
Durbin-Watson stat	0.021150	Prob(F-statistic)	0.759251	



Dependent Variable: LNSOLD  
 Method: Least Squares  
 Date: 11/12/03 Time: 15:01  
 Sample: 1 70  
 Included observations: 70

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	12.84846	0.279521	45.96607	0.0000
AGE	0.012357	0.007784	1.587484	0.1178
SEQUEL1	0.375938	0.158788	2.367549	0.0213
SEQUELS	0.323089	0.157700	2.048751	0.0450
MEDIA	0.196089	0.154645	1.267990	0.2099
TOPTEN	0.379497	0.134541	2.820676	0.0065
HOLIDAY	0.098518	0.147319	0.668740	0.5063
ACTION	0.373305	0.199094	1.875018	0.0658
SPORTS	-0.233161	0.202957	-1.148819	0.2553
RACING	-0.174141	0.205321	-0.848142	0.3998
EVERYONE	-0.105614	0.212796	-0.496315	0.6215
TEEN	-0.100045	0.210225	-0.475897	0.6359
R-squared	0.377741	Mean dependent var	13.64413	
Adjusted R-squared	0.259727	S.D. dependent var	0.578264	
S.E. of regression	0.497534	Akaike info criterion	1.596497	
Sum squared resid	14.35730	Schwarz criterion	1.981954	
Log likelihood	-43.87741	F-statistic	3.200802	
Durbin-Watson stat	0.728001	Prob(F-statistic)	0.001878	