Will Schmid

Introduction

Video games have become big business in the entertainment sector over the last few decades. What was once a niche market that catered to enthusiasts and arcade visitors has now expanded to a point where most households in America own a game console. Multiple games released in the last ten years have brought in over $500 million in revenue from just their sales in the first week of their release. This report attempts to find out what factors play an important role in what games consumers buy, with the most attention being paid to the scores each game receives from critics and journalists.

Data

The data for this report includes 326 observations and 8 variables, and was found from a multitude of online resources. Each of these observations represent a different game released in 2007. There are different observations of the same game for different platforms to explore any effects on sales based on what platform a game is released on, and I limited the observations to only games that were released on the three major home consoles of the time (Xbox 360, PS3, and Wii).

The dependent variable in this model will be retail sales of individual games. These data are represented by the *sales* variable, and are listed in millions of units sold. For example, a *sales* value of 2.4 means 2.4 million copies of that particular game were sold. All of the sales figures were found through the website vgchartz.com.

The next variable an aggregate review score for each game based off of any online reviews for it by professional game critics and journalists. It is represented by the *score* variable. Scores can range from 0 to 100, with a score of 0 meaning that game was extremely negatively reviewed, and a score of 100 meaning that a game got universally praised. These aggregate review scores were found through the website metacritic.com, which takes every review they can find for a game into account, and then calculates an aggregate score based on said reviews. It makes sense that if a game receives better review scores, it will sell better since people want to buy games that are an enjoyable experience, something that is reflected by a high review score.

The rest of the variables are dummy variables which reflect different aspects of each game. There is a variable for what platform the game was released on, which are represented by the *xbox* and *ps3* variables. During 2007, many more people owned an Xbox or Wii than a PS3. THe user base of a platform shouldn't affect the relationship between review score and sales though, so I would expect the effect of the *ps3* variable to have a negative effect on the constant term. In addition, I included a variable *publish*, which denotes whether or not a game was published by one the top 5 publishers in the video game industry (Activision, Nintendo, Take-Two, Ubisoft, and EA). Since these games probably benefited from the perks of being released by a larger publisher (notably bigger marketing budgets), it would be expected that the *publish* variable has a positive effect on sales.

I also included the variables *orig* and *multi* that represent if the game is an original intellectual property and whether or not it included multiplayer. I'm not sure there will be any effect from either of these, but it seems reasonable to assume that games that are based on

existing works would already have a fan base that would increase sales. For the *multi* variable, I would expect it to have a positive relationship with sales, since lots of people enjoy having multiplayer present in game.

Variable	Obs	Mean	Std. Dev.	Min	Max
sales	326	.4665337	.806766	.01	7.8
publish	326	.4509202	.4983503	0	1
xbox	326	.3588957	.4804139	0	1
ps3	326	.2423313	.4291524	0	1
score	326	66.1227	15.22519	19	97
orig	326	.2055215	.4047034	0	1
multi	326	.7791411	.4154633	0	1

A summary of these variable is listed below.

The Model

To begin with I ran a basic OLS regression of all of the explanatory variables on sales.

Source	SS	df	MS
Model	33.7950679	6	5.63251132
Residual	177.73812	319	.55717279
Total	211.533188	325	.650871348

Number of obs	326
F(6, 319)	10.11
Prob > F	0.0000
R-squared	0.1598
Adj R-squared	0.1440
Root MSE	.74644

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
score	.0213011	.0029025	7.34	0.000	.0155906	.0270115
xbox	-.129882	.0982944	-1.32	0.187	-.3232691	.0635052
ps3	-.3816817	.1120984	-3.40	0.001	-.6022273	-.1611361
publish	.080611	.085104	0.95	0.344	-.086825	.248047
orig	.031543	.105851	0.30	0.766	-.1767112	.2397972
multi	.0102859	.1031996	0.10	0.921	-.1927519	.2133236
_cons	-.8536881	.2057652	-4.15	0.000	-1.258516	-.4488598

The output is included below.
Right away we can see there are many variables which aren't deemed significantly significant. Both *multi* and *orig* seem to have no effect on sales, and the case that the *xbox* and *publish* variables are not 0 is weak. As expected, the *score* variable has a very strong positive coefficient, and the *ps3* variable is negative, meaning the constant term is lower for PS3 games than for other systems. The R-squared value suggests that about 16% of the variance in the data can be explained by this model.

Source	SS	df	MS		
Model	33.3837569	3	11.127919	Number of obs =	326
Residual	178.149431	322	.553259103	F(3, 322) =	20.11
				Prob > F =	0.0000
Total	211.533188	325	.650871348	R-squared =	0.1578
				Adj R-squared =	0.1500
				Root MSE =	.74381

sales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
score	.0196428	.0029305	6.70	0.000	.0138776	.0254081
ps3	-.3128183	.0988386	-3.16	0.002	-.5072692	-.1183674
pubscore	.0016995	.0012244	1.39	0.166	-.0007092	.0041083
_cons	-.8094551	.1861681	-4.35	0.000	-1.175714	-.4431958

For the second model, I dropped all of the variables deemed insignificant, and added in one m
pubscore, which is the *publish* dummy variable multiplied by the *score* variable. I thought maybe
the effect of being published by a large firm with market power was more reflected in the slope of the line than t
score and *ps3* coefficients are positive and negative respectively, although they have both decreased in magnit
pubscore variable has a higher t-statistic than the previous *publish* variable, but it still cannot be
considered significant at a 5% significance level. The R-squared value did go down, but by a
negligible amount, which means we didn't lose anything by omitting all of the variables from the previous regres

I next created the variable *lsales*, which is the log of the sales variable, and ran a

Source	SS	df	MS
Model	93.2656148	3	31.0885383
Residual	275.943369	322	.856966984
Total	369.208984	325	1.13602764

Number of obs	= 326
F(3, 322)	= 36.28
Prob > F	= 0.0000
R-squared	= 0.2526
Adj R-squared	= 0.2456
Root MSE	= .92573

lsales	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
score	.033067	.0036471	9.07	0.000	.0258918	.0402422
ps3	-.3032498	.1230111	-2.47	0.014	-.5452567	-.0612429
pubscore	.0031829	.0015238	2.09	0.038	.000185	.0061808
_cons	-3.622334	.2316983	-15.63	0.000	-4.078167	-3.1665

regression of *score*, *ps3*, and *pubscore* on it. The results are shown below.
Using this regression, the *pubscore* variable is significant at a 5% level, and has a positive coefficient as expected. The R-squared value jumped up to about .25, which means more of the variance in the data is explained by this model, and the *score* and *ps3* variables exhibit the same behaviour as before. I examined the residual plot was a little concerned about heteroskedasticity, so I ran a White test to test for it. It returned back a chi-squared value of 17.81, meaning there is strong evidence of heteroskedasticity. To counteract this, I decided to use robust errors in this model. Rerunning the regression with robust errors leads to the

Linear regression

Number of obs	= 326
F(3, 322)	= 28.60
Prob > F	= 0.0000
R-squared	= 0.2526
Root MSE	= .92573

lsales	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
score	.033067	.0042728	7.74	0.000	.0246608	.0414732
ps3	-.3032498	.1083038	-2.80	0.005	-.5163222	-.0901774
pubscore	.0031829	.001563	2.04	0.043	.000108	.0062578
_cons	-3.622334	.2636139	-13.74	0.000	-4.140957	-3.103711

regression output shown below.
The robust errors widened the confidencee intervals as expected, but all of the variables are still significant at a 5% level. This is the specification I chose to use for my analysis.

Analysis

The main variable of interest here is the *score variable*. It has a very high t-statistic of 7.74, so there is very strong evidence that it is positive and not-zero. It's point estimate of about .03 in the log-linear model means that a one unit change in the *score* variable will lead to about a 3% change in the *sales* variable. The 95% confidence interval for *score* is from .025 to .041, suggesting that there is very strong evidence that review scores have an effect on sales. This fits in very well with the theory that positive game reviews drive sales.

The *ps3* variable is also deemed statistically significant with a t-statistic of -2.8. It's point estimate of about -.30 suggests that if a game is released on the PS3, the constant term in the equation will be decreased by that much. This also falls in line with my expectations, since I wouldn't expect the relationship between *score* and *sales* to be any different just because less people own a certain system.

With a t-statistic of 2.04, the *pubscore* variable is the closest to being deemed insignificant. It's positive coefficient implies that if a game is released by one of the top publishers, that amount is added to the coefficient for the *score* variable, meaning that reviews will have even more of an effect on sales. The point estimate of about .003 means that if a game is released by a big publishers, a change in review score by 1 leads to only a .3% extra change in sales though, so its effect isn't that strong.

Conclusion

The goal of this report was to explore the relationship between video game's sales and the review scores it gets by various publications. Based on the regressions performed, I would say there is very strong evidence of a positive log-linear relationship between the two. Other factors were included to see if there are any other strong explanatory variables for video game sales, but most of them were not significant. The choice of platform on which a game is released had an effect on sales, but didn't change the relationship between scores and sales, while being released by one of the top publishers made review score affect sales slightly more. While I was able to draw these conclusions based on these data, there are many more things one could do with this basic idea. For one, I only used data from one year due to the time it took to manually enter it in. There are many more years of data available which could be used to draw stronger conclusions or find relationships I couldn't. In addition there are many other variables that could be included that I didn't for time's sake. As the video game industry continues to grow throughout the coming years, I expect even more research to be done to analyze what drives consumers to buy certain games. Review scores obviously play an important role, but I'm sure there is much more work to be done on this topic.