How can we predict the sale of video games across the world?

Aditya Singampalli

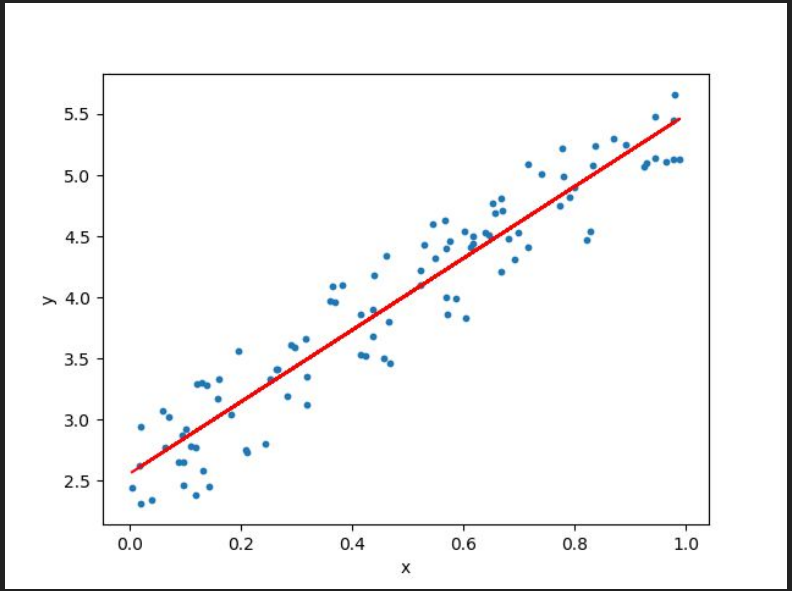
The Dataset

I found the data set that I used to answer this question on kaggle.com. It ranks ps4 games based on total number of units sold. In addition, the data set has 9 columns/variables, which lists, the game, year of release, genre, publisher, and units sold in various regions (North America, Europe, Japan, rest of world, global).

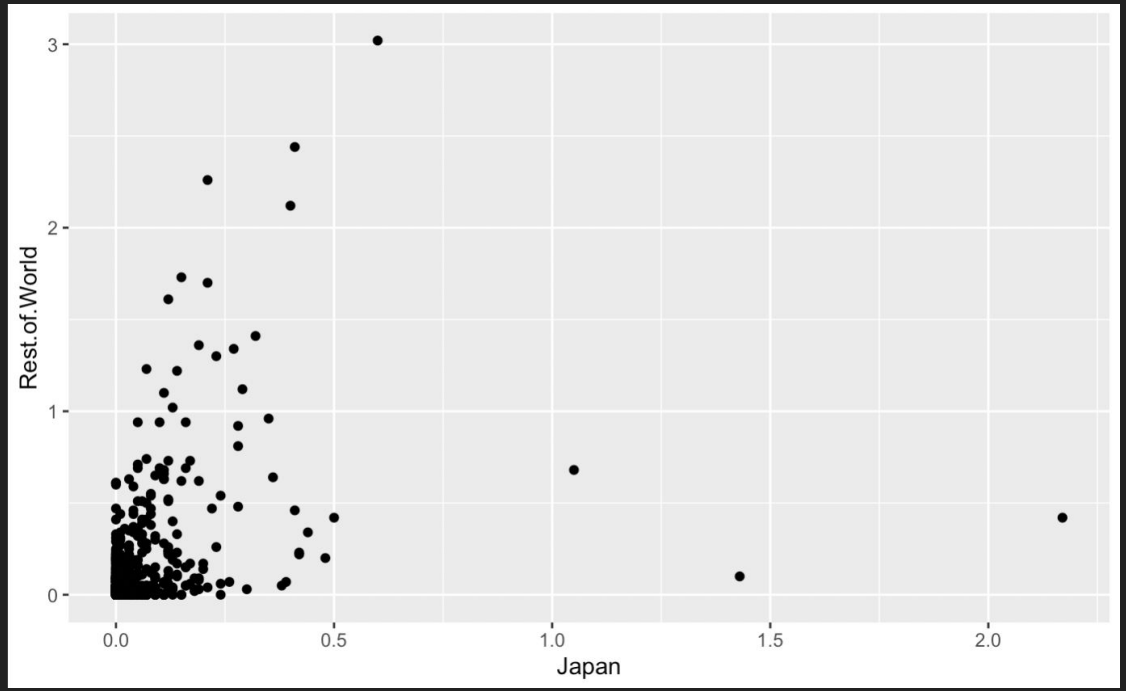
	Game	Year	Genre	Publisher	North.America	Europe	Japan	Rest.of.World	Global
1	Grand Theft Auto V	2014	Action	Rockstar Games	6.06	9.71	0.60	3.02	19.39
2	Call of Duty: Black Ops 3	2015	Shooter	Activision	6.18	6.05	0.41	2.44	15.09
3	Red Dead Redemption 2	2018	Action-Adventure	Rockstar Games	5.26	6.21	0.21	2.26	13.94
4	Call of Duty: WWII	2017	Shooter	Activision	4.67	6.21	0.40	2.12	13.40
5	FIFA 18	2017	Sports	EA Sports	1.27	8.64	0.15	1.73	11.80
6	FIFA 17	2016	Sports	Electronic Arts	1.26	7.95	0.12	1.61	10.94
7	Uncharted (PS4)	2016	Action	Sony Interactive Entertainment	4.49	3.93	0.21	1.70	10.33
8	Spider-Man (PS4)	2018	Action-Adventure	Sony Interactive Entertainment	3.64	3.39	0.32	1.41	8.76
9	Call of Duty: Infinite Warfare	2016	Shooter	Activision	3.11	3.83	0.19	1.36	8.48
10	Fallout 4	2015	Role-Playing	Bethesda Softworks	2.91	3.97	0.27	1.34	8.48
11	FIFA 16	2015	Sports	EA Sports	1.15	5.77	0.07	1.23	8.22
12	Star Wars Battlefront 2015	2015	Shooter	Electronic Arts	3.31	3.19	0.23	1.30	8.03
13	Call of Duty: Advanced Warfare	2014	Shooter	Activision	2.84	3.34	0.14	1.22	7.53
14	Battlefield 1	2016	Shooter	Electronic Arts	2.20	3.65	0.29	1.12	7.26
15	The Last of Us	2014	Action-Adventure	Sony Computer Entertainment	2.70	2.86	0.11	1.10	6.77
16	MineCraft	2014	Misc	Sony Computer Entertainment	1.89	3.13	0.35	0.96	6.33
17	FIFA 15	2014	Sports	EA Sports	0.83	4.49	0.05	0.94	6.32
18	God of War (PS4)	2018	Action	Sony Interactive Entertainment	2.83	2.17	0.13	1.02	6.15
19	Horizon: Zero Dawn	2017	Action	Sony Interactive Entertainment	2.20	2.43	0.28	0.92	5.82

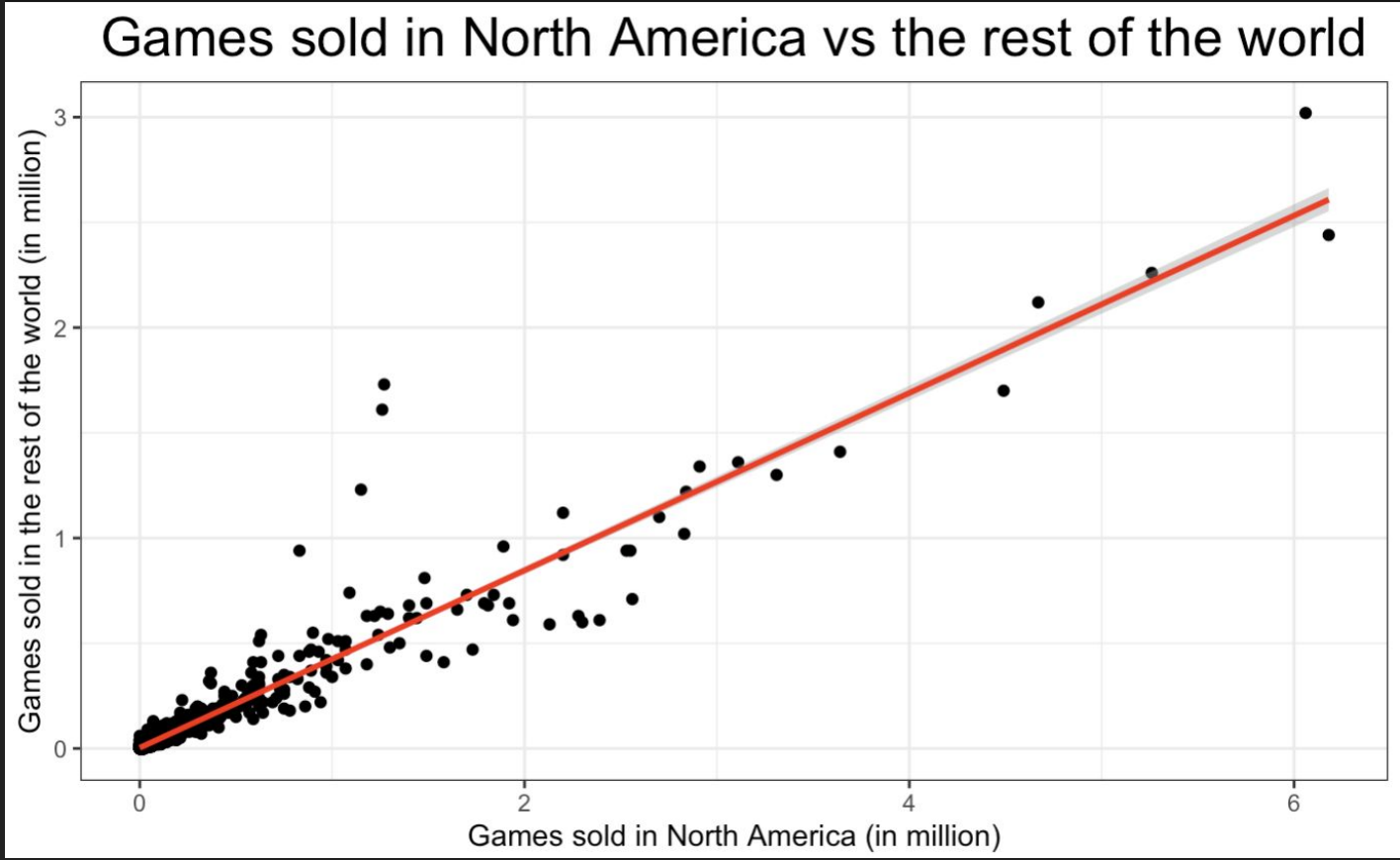
The Model:

Since the relationship we are trying to understand is between two numerical variables, I decided that a linear regression would be the best model to use in this situation.



Number of games sold in Japan vs the rest of the world




Games sold in North America vs the rest of the world
Games sold in the rest of the world (in million)
0
1
2
3
0
2
4
6
Games sold in North America (in million)

Accuracy:

- I calculated the accuracy of my model based on the mean squared error or mse for short

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$$

Accuracy:

$$MSE = 0.003744471$$

Improving my model

- To make my model a bit more accurate I decided to add in an extra predictor variable, so now I would be predicting game sales across the rest of the world based on North America AND Japan.

Improved Accuracy

$$MSE = 0.003658904$$

Conclusion:

There were quite a few missing values in my data, and since I removed those rows outright, I could have changed how my data looked entirely by reducing the sample size, possibly inflating the statistical significance, and overfitting. Overall, finding quality data sets was a bit tricky and it took me some time to even find a large enough one. Still, my model appeared to perform alright and scored pretty well in accuracy.

Link to model:

https://github.com/aditya-singam/atdpFinalProject/blob/main/AdityaSingampalli_Section2_FinalProject_VideoGameSales.Rmd