UV6767
Rev. Feb. 3, 2014

# LOGISTIC REGRESSION

Almost all of us are familiar with odds. What are the chances one thing will happen versus another? What are the chances you will succeed at work today? What are the chances your favorite game-show contestant will win today versus the chances he or she will lose?

What we might not be familiar with is how odds can be applied to marketing analytics. What are the chances a customer will buy your product versus the chances he or she won't? What are the chances you will retain a customer versus the chances you will lose him or her?

When you are using odds, you are examining two opposing outcomes. Any such unknown (i.e., one that can only be one thing or another) is known as a dummy variable. But if you know how to examine dummy variables properly, the results are anything but dumb.

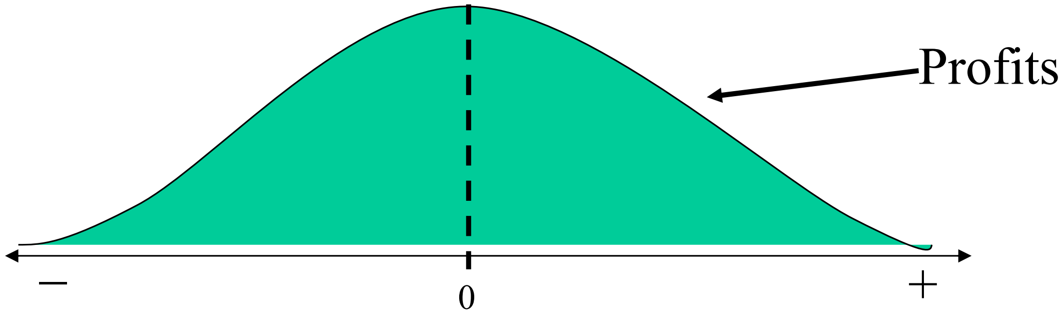
## When Logistic Regression Trumps Linear Regression

A logistic regression is similar to any linear regression but with one important variation that has critical consequences.

Think about an important metric in marketing: customer retention. If Keepmoney Bank wants to use a regression analysis to examine whether it will retain a customer, it will set retention as its dependent variable. Rather than being normally distributed in a bell curve in the manner of continuous variables (**Figure 1**), however, a 1 will be assigned to represent customer retention and a 0 will represent customer loss. Only those two outcomes are possible. Again, this is a dummy variable, wherein what you are trying to predict is one of two options.

This technical note was prepared by Shea Gibbs, Research Assistant, and Rajkumar Venkatesan, Bank of America Research Professor of Business Administration.

Figure 1. A normal distribution.



Source: All figures created by case writer unless otherwise specified.

Studies have shown that logistic regression is the best model for examining dummy variables such as customer retention.[1] But why can't Keepmoney use its trusty linear regression to determine the likelihood of customer retention given a set of independent variables? Again, linear regressions assume a bell-curve distribution of outcomes (what is known as a normal distribution) from negative infinity to infinity. Most things in life follow this sort of distribution. Think of human height or school grades—a few people typically earn Cs, a few more earn a B−, the majority will earn Bs, and a very few will earn an A+. But when examining a dummy variable such as customer retention, there is no curve across a range of outcomes. The outcome can only be 1 or 0.

If Keepmoney attempts to use a linear regression to examine customer retention, nonsensical predictions may result. The bank may find its chances of customer retention are greater than 1, meaning it has even better than a 100% chance of retaining a customer. Or the bank may find its chances are less than 0. One can round up for those predictions that are less than 0 or round down for those greater than 1, but the results of the regression will not be precise.

## Choice Behavior

The objective of logistic regression in this example is to represent consumers' choice behavior as accurately as possible. When individual consumers choose products, the value they place on the product does not typically increase linearly with increases in a preferred feature of the product. Instead, research indicates consumer valuation of a product typically follows an S-shaped curve with increases in the levels of a preferred attribute.

[1] Scott A. Neslin et al., "Defection Detection: Measuring and Understanding the Predictive Accuracy of Customer Churn Models." *Journal of Marketing Research* 43, no. 2 (2006): 204–211.

We can test whether the S-shaped curve represents consumers' choice behavior with a simple exercise. Imagine that on the x-axis we have the level of discount on a $300 plane ticket from Charlottesville, Virginia, to New York. Ask a group of your friends how many of them would purchase the flight. Then offer a discount of $20. How many additional people said they would buy the ticket? Probably not many. Increase the discount to $40. Maybe one person half-heartedly jumps in. At $60, you are likely to see a spike in purchasers. And from $60 to $100, the number of purchasers should increase at every level; however, at about $100, the number of additional purchasers will taper off, as you have reached the upper threshold.

In most real-life situations, this S-shaped curve represents how people make decisions. As a discount (i.e., promotion) increases, the odds that people will make the choice to buy will increase. In this example, at a $60 discount, 2 in 10 people are likely to purchase the flight to New York; 8 in 10 are unlikely to purchase the flight.

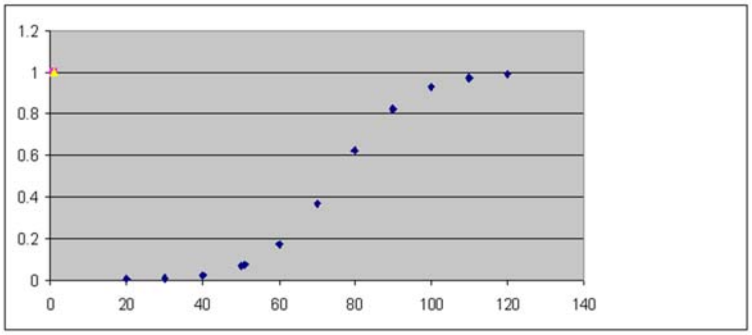
## The Logistic Transformation

We now see that a linear regression would be insufficient to accurately represent individual consumers' choices. In **Figure 2**, we show a distribution of probabilities from 0 to 1 representing the logistic function

1 ———

where $u_p$ = utility consumer obtains from product $p = a + b_1X$.

Figure 2. Distribution of probabilities for a logistic distribution.

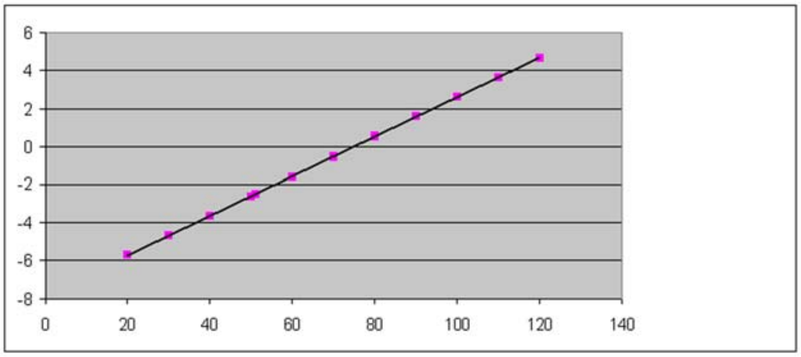


The utility function ($u_p$), otherwise known as a value function, is used to describe the value a person places on a certain good or service. Take coffee, for example. To find the utility, or value, you might derive from a cup of coffee, you must consider all of the variables that might go into the decision to buy that particular cup: the taste, the price, the logo, the location of the store from which you buy it, your personal habits, and the jolt it gives you in the morning. For convenience purposes—and based on behavioral studies indicating how people process variables in an additive way—the value function is assumed to be linear.

The logistic function used to describe the ways in which consumers make choices takes the form of the exponent of the value function over 1 plus the exponent of the value function. The resulting distribution looks like an S-shaped curve, as shown in **Figure 2**. The predictions from this function are bound between 0 and 1 (meaning if one outcome is 0.1, the opposite outcome is 0.9).

Furthermore, the probability of success (retention) versus failure (churn) is $P \div (1 - P)$, where $P$ is the probability of retention. For example, if there are 10 outcomes with 1 success and 9 failures, the odds are 1/9. This ($P \div (1 - P)$) is what is known as the "odds function." Substituting for $P$ using the logistic function above, the odds function is equal to $e^{(a+b_1X)}$. If we are to make a transformation of this exponential function to a linear function via the natural log,[2] we will find the log odds function, which is $ln[P \div (1 - P)] = a + b_1X$ (**Figure 3**). This is equivalent to the value function.

Figure 3. Log odds function.



Essentially, we have assumed a person has a linear value function or utility underlying his or her decision, then we have transformed that value into something useful about the chances he or she will make a decision. Therefore, the critical output of a logistic regression is the probability, or percent chance, a customer will stay with a company or leave the company, and that probability is defined in terms of the value the customer places on the company's product.

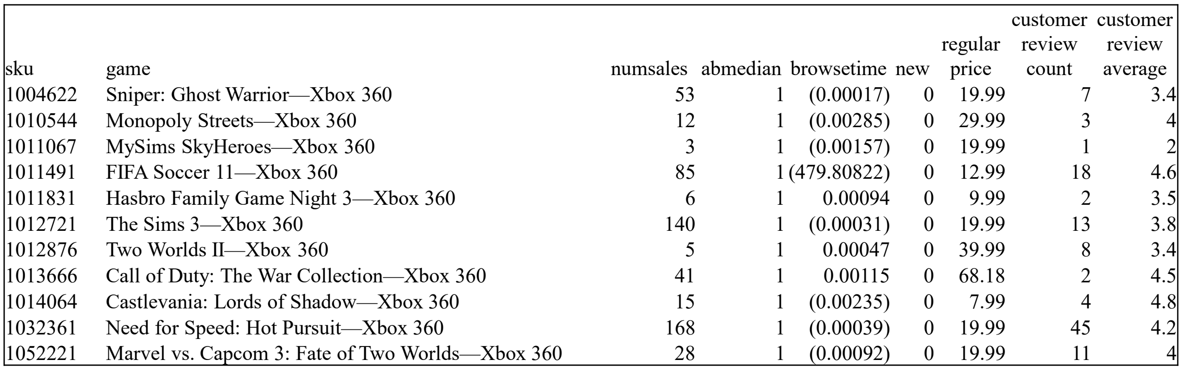
## Assessing Video Game Purchasers

How can a marketing manager use logistic regression techniques to find useful information about the ways people behave? Consider the data in **Figure 4**, which tally the number of sales of Xbox games through Best Buy's mobile app, as reported by Kaggle.[3]

[2] See **Appendix 1** for more information on transforming an exponential function to a linear function via the natural log.

[3] Kaggle is a user-generated business analytics community. For more information, visit http://www.kaggle.com.

Figure 4. Sales of Xbox games through Best Buy's mobile app.



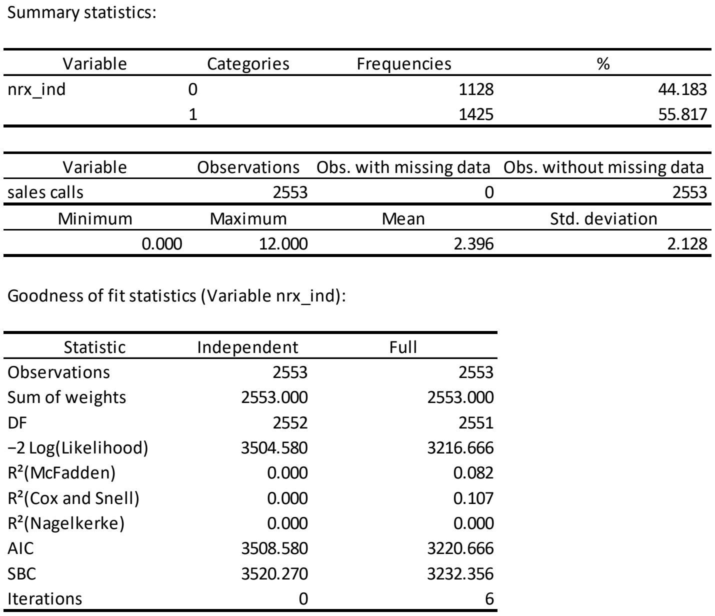
| sku | game | numsales | abmedian | browsetime | new | regular price | customer review count | customer review average |
|---|---|---|---|---|---|---|---|---|
| 1004622 | Sniper: Ghost Warrior—Xbox 360 | 53 | 1 | (0.00017) | 0 | 19.99 | 7 | 3.4 |
| 1010544 | Monopoly Streets—Xbox 360 | 12 | 1 | (0.00285) | 0 | 29.99 | 3 | 4 |
| 1011067 | MySims SkyHeroes—Xbox 360 | 3 | 1 | (0.00157) | 0 | 19.99 | 1 | 2 |
| 1011491 | FIFA Soccer 11—Xbox 360 | 85 | 1 | (479.80822) | 0 | 12.99 | 18 | 4.6 |
| 1011831 | Hasbro Family Game Night 3—Xbox 360 | 6 | 1 | 0.00094 | 0 | 9.99 | 2 | 3.5 |
| 1012721 | The Sims 3—Xbox 360 | 140 | 1 | (0.00031) | 0 | 19.99 | 13 | 3.8 |
| 1012876 | Two Worlds II—Xbox 360 | 5 | 1 | 0.00047 | 0 | 39.99 | 8 | 3.4 |
| 1013666 | Call of Duty: The War Collection—Xbox 360 | 41 | 1 | 0.00115 | 0 | 68.18 | 2 | 4.5 |
| 1014064 | Castlevania: Lords of Shadow—Xbox 360 | 15 | 1 | (0.00235) | 0 | 7.99 | 4 | 4.8 |
| 1032361 | Need for Speed: Hot Pursuit—Xbox 360 | 168 | 1 | (0.00039) | 0 | 19.99 | 45 | 4.2 |
| 1052221 | Marvel vs. Capcom 3: Fate of Two Worlds—Xbox 360 | 28 | 1 | (0.00092) | 0 | 19.99 | 11 | 4 |

Data source: Kaggle, "Data Mining Hackathon on BIG DATA (7GB) Best Buy mobile web site," http://www.kaggle.com/c/acm-sf-chapter-hackathon-big (accessed November 5, 2013).

Each of the games shown in this data set boasts above-median sales compared with the other games available. In other words, a dummy variable has been set where "above-median sales" is represented by a 1, and "below-median sales" is represented by a 0. Now, which independent variables shown in the chart (time browsed, whether the game is new, price, number of reviews, and review average) are good predictors of being a 1—that is, above-median sales?

The output of a logistic regression of this data (**Figures 5** and **6**) looks similar to the output of a linear regression, and the most important data points, in addition to the coefficients, are r squared and p-value; other predictors of accuracy and significance go by a variety of names.

Figure 5. Output of logistic regression.



Summary statistics:

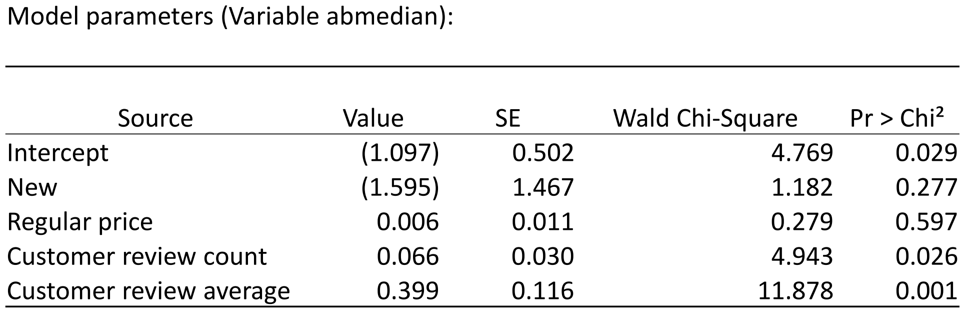
| Variable | Categories | Frequencies | % |
|---|---|---|---|
| nrx_ind | 0 | 1128 | 44.183 |
| | 1 | 1425 | 55.817 |

| Variable | Observations | Obs. with missing data | Obs. without missing data |
|---|---|---|---|
| sales calls | 2553 | 0 | 2553 |
| Minimum | Maximum | Mean | Std. deviation |
| 0.000 | 12.000 | 2.396 | 2.128 |

Goodness of fit statistics (Variable nrx_ind):

| Statistic | Independent | Full |
|---|---|---|
| Observations | 2553 | 2553 |
| Sum of weights | 2553.000 | 2553.000 |
| DF | 2552 | 2551 |
| −2 Log(Likelihood) | 3504.580 | 3216.666 |
| $R^2$(McFadden) | 0.000 | 0.082 |
| $R^2$(Cox and Snell) | 0.000 | 0.107 |
| $R^2$(Nagelkerke) | 0.000 | 0.000 |
| AIC | 3508.580 | 3220.666 |
| SBC | 3520.270 | 3232.356 |
| Iterations | 0 | 6 |

Figure 6. Model estimates.



Model parameters (Variable abmedian):

| Source | Value | SE | Wald Chi-Square | Pr > Chi² |
|---|---|---|---|---|
| Intercept | (1.097) | 0.502 | 4.769 | 0.029 |
| New | (1.595) | 1.467 | 1.182 | 0.277 |
| Regular price | 0.006 | 0.011 | 0.279 | 0.597 |
| Customer review count | 0.066 | 0.030 | 4.943 | 0.026 |
| Customer review average | 0.399 | 0.116 | 11.878 | 0.001 |

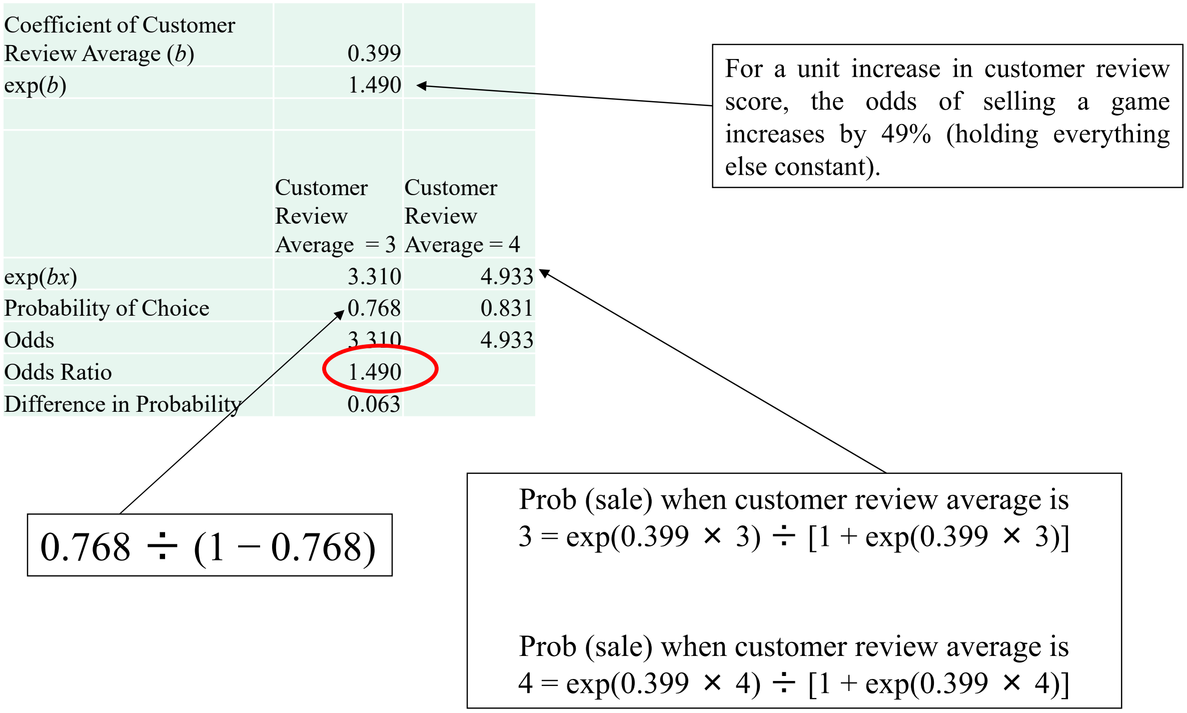
The key difference in the logistic regression output is that the coefficients are not interpreted as such. In order for the coefficients to add value to your analysis, you must calculate the odds ratio. For example, if a logistic regression yields a coefficient *b* of 2.303, the odds ratio says that for every one unit increase in the independent variable (e.g., number of promotions), the odds that the dependent variable will be equal to 1 (e.g., the product is purchased) will increase by a factor determined by taking the exponent of the coefficient: $e^b = e^{2.303} = 10$. This is not the same as a direct linear transformation.

So, examining the p-values shown in the far-right column of **Figure 6**, which variables can we say are predictive of whether a game will be a top seller? Customer review average, followed by the number of customer reviews, is the most significant variable. Price is relatively insignificant, in this case most likely due to the fact that the price range of the games is small.

Using the coefficients determined in the regression analysis, the marketing manager can then determine how much the odds of a game being a top seller increase if review average increases by one point (**Figure 7**). In other words, if a customer review average of 3 yields a certain probability of success, what happens if the average increases to 4? On average, the coefficient of customer review (coefficient *b*, the slope of the line) is 0.399, and the exponent of *b* is 1.49, which means that a single-point increase in reviews increases the odds by a factor of about 1.5.[4]

[4] For more information on how the odds ratio can be calculated, please see **Appendix 2**.

Figure 7. Equivalence of log odds ratio and logistic probabilities.



| | | |
|---|---|---|
| Coefficient of Customer Review Average (*b*) | 0.399 | |
| exp(*b*) | 1.490 | |
| | | |
| | Customer Review Average = 3 | Customer Review Average = 4 |
| exp(*bx*) | 3.310 | 4.933 |
| Probability of Choice | 0.768 | 0.831 |
| Odds | 3.310 | 4.933 |
| Odds Ratio | 1.490 | |
| Difference in Probability | 0.063 | |

For a unit increase in customer review score, the odds of selling a game increases by 49% (holding everything else constant).

$0.768 \div (1 - 0.768)$

Prob (sale) when customer review average is 3 = $\exp(0.399 \times 3) \div [1 + \exp(0.399 \times 3)]$

Prob (sale) when customer review average is 4 = $\exp(0.399 \times 4) \div [1 + \exp(0.399 \times 4)]$

## Conclusion

Marketing managers often want to predict customer behaviors that are not distributed across a range of outcomes. These are cases where only one of two behaviors is possible: buy or don't buy, customer retention versus customer loss, and so on. Here, if the manager attempts to use a traditional linear regression to examine the behaviors, nonsensical predictions can result.

But a logistic regression can be used to represent consumers' choice behavior. By transforming the value function into a logistic function, we can model how the value a consumer places on a product increases with a preferred feature of the product. The critical output of the logistic regression is therefore the increase (or decrease) in the percent chance a customer will perform a behavior based on a unit increase in a variable correlated with that behavior.

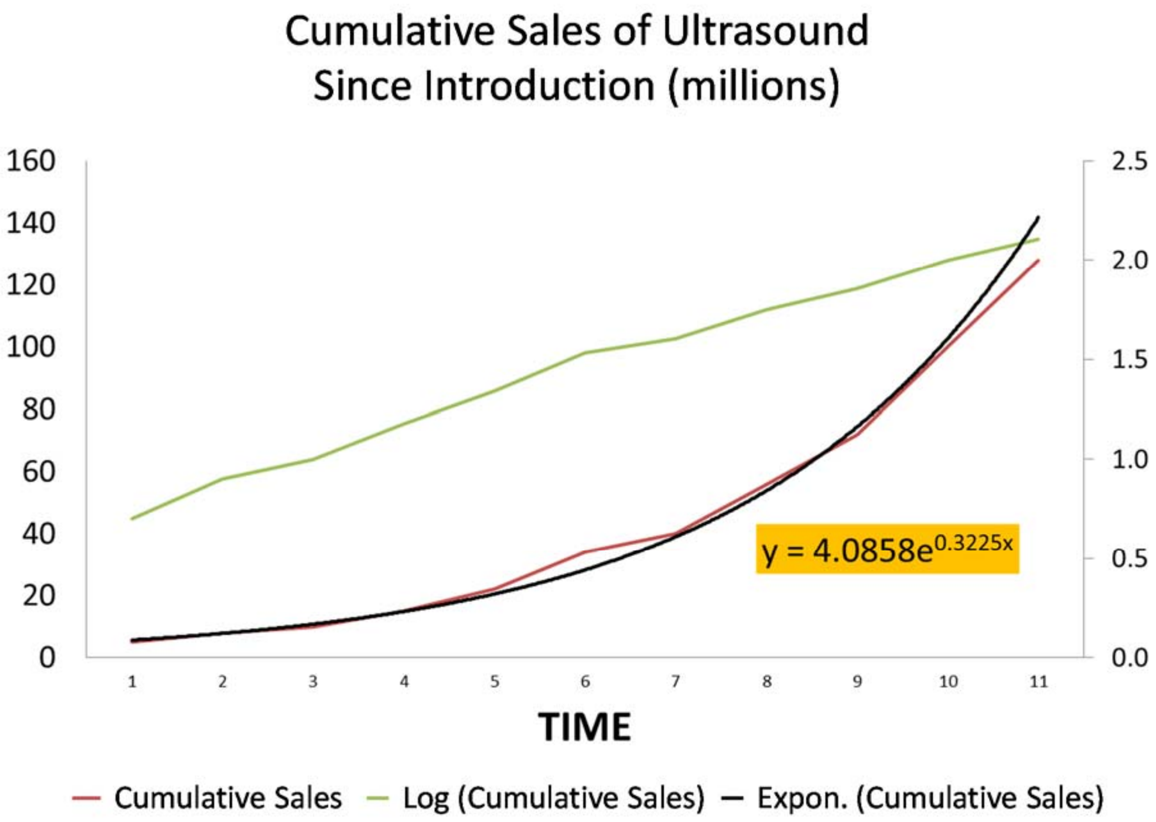
Appendix 1

## LOGISTIC REGRESSION

Understanding Exponential Functions

In order to understand logistic regressions, it is helpful to first examine exponential functions. **Figure 1** shows the classic example of an exponential distribution. When considering the cumulative sales of a product that has gained market acceptance over time (such as ultrasound machines), we see that sales are slow at first but begin to increase at a greater and greater rate once they have reached critical mass. In the graph, the red line is the actual data, or number of sales per year since introduction. What stands out is that the curve is not a straight line, whereas the ones used in linear regressions are. This is an exponential distribution.

Figure 1. An example of an exponential distribution.



Source: Created by case writer.

Appendix 1 (continued)

The black line represents a function, created using a computer program,[1] which best accounts for the data shown in the graph. The regression analysis of the available data has produced a line defined by the form $y = 4.0858e^{0.3225x}$, where 4.0858 is the intercept of the line and the slope (0.3225) changes exponentially. (The constant $e$ is an irrational number approximately equal to 2.71828, which is related to the rate of change in an exponential function and is the base of the natural logarithm. This function is found in a similar way as a straight-line function when performing a linear regression analysis.

One thing to note about this analysis is that the regression line fits almost perfectly. Because of the volume of data used, r squareds of up to 99% are possible, as compared with the r squareds of 20% to 30% one finds when running linear analyses. This is because the data are aggregate and viewed retrospectively, whereas linear regressions attempt to describe the behavior of individuals. If the same analysis of cumulative ultrasound sales was conducted in year two, however, it would be difficult to predict what would happen in years three, four, or five, because r squared breaks down at that point.

What does this have to do with logistic regressions? Consider the green line in **Figure 1**, which represents the natural log of cumulative sales at each time period $x$. The line is nearly straight, meaning a linear regression analysis could produce an accurate function describing the data. In other words, a logistic transformation of exponentially distributed data allows you to view the outputs of the regression in the same way you would a linear regression.[2]

[1] For more information on how to perform a logistic regression using computer software, please visit http://dmanalytics.org/.

[2] In algebraic terms, if $y = 4.0858e^{03225x}$, the natural log of $y$ will equal $4.0858 + 3.225x$, a linear function where the intercept is 4.0858 and the slope is 3.225.

Appendix 2

## LOGISTIC REGRESSION

Calculating Odds Ratio

Let us consider the log odds ratios presented in **Figure 7** and the logistic regression output in **Figure 6**. The log odds ratio is defined as the probability of observing an event (*p*) versus the probability of not observing and event ($1 - p$). In the context of the choice of games on the mobile app, we are considering the factor by which the log odds of purchasing a game increases when the review for the product increases from 3 to 4. A simple way to calculate this would be to take the exponent of the coefficient of reviews from the logistic regression output. In our case, the coefficient of reviews equals 0.399. So the log odds will increase by a factor of 1.49 or 149% (exp(0.399)) when the reviews for a product increases by one unit.

In Figure 7 we show that formula for calculating the log odds factor is equivalent to (a) computing the predicted probability of product choice when the reviews for the products are 3 and 4, and (b) then taking the ratio of these respective probabilities. The probability of product choice when average product review equal 3 is 0.768 and the corresponding log odds is 3.3. Similarly, the probability of choice when average product review equals 4 is 0.831 and the log odds is 4.933. The ratio of log odds (4.933 ÷ 3.3) equals 1.4. Hence the log odds increases by a factor of 1.4 or 140% when the average reviews for the product increases by one unit.