Practical Assessment, Research, and Evaluation

Volume 17 *Volume 17, 2012* Article 11

2012

Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results

Jason W. Osborne

Practical Assessment, Research & Evaluation

An open-access, peer-reviewed electronic journal.

Volume 17, Number 11, June 2012 ISSN 1531-7714

Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results

Jason W. Osborne
Old Dominion University

Logistic regression is slowly gaining acceptance in the social sciences, and fills an important niche in the researcher's toolkit: being able to predict important outcomes that are not continuous in nature. While OLS regression is a valuable tool, it cannot routinely be used to predict outcomes that are binary or categorical in nature. These outcomes represent important social science lines of research: retention in, or dropout from school, using illicit drugs, underage alcohol consumption, antisocial behavior, purchasing decisions, voting patterns, risky behavior, and so on. The goal of this paper is to briefly lead the reader through the surprisingly simple mathematics that underpins logistic regression: probabilities, odds, odds ratios, and logits. Anyone with spreadsheet software or a scientific calculator can follow along, and in turn, this knowledge can be used to make much more interesting, clear, and accurate presentations of results (especially to non-technical audiences). In particular, I will share an example of an interaction in logistic regression, how it was originally graphed, and how the graph was made substantially more user-friendly by converting the original metric (logits) to a more readily interpretable metric (probability) through three simple steps.

Use of logistic regression has been growing over recent years as more social scientists are trained in the procedure. In the last few years, popular statistics books have incorporated chapters on logistic regression (Cohen, Cohen, West, & Aiken, 2002; Field, 2009; Pedhazur, 1997; Tabachnick & Fidell, 2001), and some standalone books have been published with the social scientist in mind (Menard, 2002). Unfortunately, reviews of application of logistic regression show some continuing misunderstanding of this important and fun technique, even in the biomedical sciences (Holcomb Jr, Chaiworapongsa, Luke, & Burgdorf, 2001). In particular, many who wish to understand logistic regression are not clear on how odds ratios are calculated, what a logit is, how to convert between probabilities and odds and logits, and how this can dramatically improve the comprehensibility and communication clarity of results from logistic regression analyses. The goal of this paper is to briefly (and gently) walk readers through the mathematics of how these things are calculated, and how this knowledge can be used for the benefit of the reader.

The example I will use throughout this paper comes from the National Education Longitudinal Study of 1988 (NELS88) from the National Center for Educational Statistics (http://nces.ed.gov/surveys/nels88/), a survey of students in 8th grade in the US in 1988. These students were followed for many years on thousands of variables, similar to other studies from NCES. In particular, we will predict DROPOUT before completing 12th grade (1=yes, 0=no)[1] from a variable I calculated called POOR (1= the student falls below the average

[1] For those of you who are interested, I considered students who dropped out and returned as dropouts as well.

family income, or 0= the student falls above the average family income).[2]

Probabilities, conditional probabilities, and odds

If you are like most, the thought of calculating odds and probabilities may make you cringe or bring memories of slogging through endless problems from your introduction to statistics class(es). I will try to make this as painless as possible, because (a) I really don't like slogging through endless example calculations either, and (b) these are relatively simple concepts that are actually pretty fun once you understand them.

Let us begin our example of looking at student dropout from high school and family income. I have presented a crosstabulation of the variables in Table 1. We will start with simple counts of students in each group, and quickly use those numbers to calculate complex things like odds ratios and logits.

Table 1 Crosstabulation of family income and dropout

		DROPOUT		Total	Conditional prob.	Odds	Odds ratio
		No	Yes				
POOR	Yes (1)	7312	1244	8556	0.145	0.170	5.67
	No (0)	7821	233	8054	0.029	0.030	
	Total	15133	1477	16610			

Looking at the row labeled "Total," you can see that 1477 out of our sample of 16610 were classified as having dropped out. The probability of an event is calculated as the frequency of the event divided by the total observations (in this case, 1477 dropouts out of 16610 total students).

Probability of dropout ($P_{dropout}$) = number dropouts / total students

$P_{dropout}$ = 1477 / 16610

$P_{dropout}$ = 0.0889

Thus, in the overall sample, 8.89% of the sample dropped out, giving us a probability of dropout of 0.0889. When there are two categories (as with this dropout/retained variable), the probability of a student falling into the "retained" category is (1- $P_{dropout}$):

Probability of retained (1-$P_{dropout}$) = number retained / total students *or* 1- $P_{dropout}$

1-$P_{dropout}$ = 15133 / 16610 *or* 1- 0.0889

1-$P_{dropout}$ = 0.9111

Conditional probabilities. While it is important to know the overall dropout (or retention) rate, in Table 1 it is clear that there are more students from "poor" households dropping out of school, and fewer from "not-poor" households. Hopefully you are beginning to think about what percent of each group dropped out, or what is the probability that a student from a particular group dropped out. The probability of dropout within a group is called a *conditional probability*. Thus, for example, we can calculate the conditional probability of dropout for students coming from "poor" (below-average income) households. In this group, 1244 students dropped out (from a total of 8556), yielding a conditional probability of 0.145 . Likewise, we can calculate the conditional probability for those students coming from households with above-average income (233 students in this group dropped out from a total of 8054, yielding a probability of 0.029). In other words, by knowing one piece of information about a student's background, we have a more nuanced view of dropout probability. Students coming from below-average income households are much more likely to drop out than students coming from above-average income households.

In fact, those of you with a background in OLS regression might find it interesting to note that

[2] A brief note on interpretation: I am using this public data for demonstration purposes only. I intentionally did not weight the data or do any of the methodologically important steps necessary to appropriately use data from this type of complex multistage sample for drawing substantive conclusions. Therefore, you should not draw any substantive conclusions about dropout and family income based on these data. They are for illustrative purposes only. For more on the importance of weighting complex samples such as this, I will refer you to my paper on the topic: http://pareonline.net/pdf/v16n12.pdf, (Osborne, 2011)

when you have dichotomous variables in OLS regression, with both variables coded 0 and 1, the conditional probabilities of dropout are the predicted variable. Putting the exact same data into an OLS regression analysis produces the following results:

Table 2: OLS regression results of the same data

	Unstandardized Coefficients		Standardized Coeff	t	Sig.
	B	Std. Error	Beta		
(Constant)	.029	.003		9.318	<.001
poor	.116	.004	.204	26.923	<.001

And the following prediction equation:

Conditional probability of dropout = 0.029 + 0.116 (Poor)

As you can see from Table 2,, when the IV is 0 (not poor), the conditional probability is 0.029, which matches our calculated conditional probability in Table 1, above. Likewise, when POOR=1, the predicted probability is 0.029 + 0.116, or 0.145, which again matches the conditional probability we calculated.

Before the widespread availability of logistic regression, OLS regression of this type was one of the few options available to researchers wanting to study questions such as this. Unfortunately, it cannot be considered a best practice as the assumptions are difficult to match, and the predicted probabilities can become impossible when the IV is continuous (i.e., below 0 or above 1.0).

A brief thought experiment on the logistic curve. From these data and common sense, we can see something that is usually presented in discussions of logistic regression but not delved into deeply: the logistic curve. If poverty was strongly related to the probability that a student would drop out, the conditional probability of dropout would increase as poverty increased, but at some point, increased poverty doesn't substantially increase the probability of dropout. There may be a threshold above which the probabilities don't change substantially. Conversely, as you move downward toward very low poverty (increasing affluence), the probabilities might quickly asymptote toward 0. The probability of dropping out might be similar if a student's family makes $100,000.00 per year or $100,000,000.00, but it might make a large difference in dropout probabilities if the family makes $25,000.00 or $35,000.00. This theoretical relationship is presented below in Figure 1. As you can see, there is a relatively narrow window of poverty where changing makes a large difference, and outside that window, the probabilities don't change a great deal. In this fictitious example, when poverty (on whatever scale we are using) reaches 1.40, the probability of a student dropping out is about .80. conversely, at -1.40, the probability is about .20. Beyond these points, the slopes flatten out, giving less change in probability despite rather large changes in X.

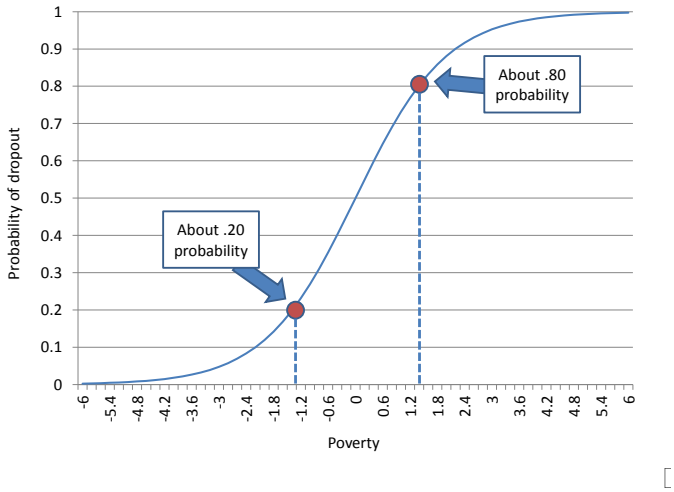

Figure 1. Hypothetical logistic curve relating poverty to probability of dropout

Think about this relationship in another way. Let's imagine that we were looking at the dosage of a hypothetical drug and the probability that we could cure a disease. The hypothetical drug is very effective and has no known side effects. If x is dosage and y is the probability of cure, you might well get a similar curve. At very low doses, there are very small probabilities of cure, but as the doctors increase the dosage, there will come a point where it

begins becoming effective, and as dosage increases (to a point) probability of cure will also increase. Then at some point, the benefit of increasing the dosage will level off as probability of cure reaches a maximum threshhold and increasing beyond that point will not materially increase the probability.

The benefits of odds. So one reason we don't use OLS regression in this sort of example is that we can get impossible predicted probabilities (below 0 or above 1.0). We can partly get around the issue of impossible values if we look at odds rather than probabilities. There are drawbacks to odds—such as being difficult to accurately interpret—but their benefits are that they only range from 0.00 to infinity. Conditional odds are calculated as the probability of that event divided by the probability of the event not happening:

$Odds_{(dropout)}$ = probability of dropout/ probability of not dropping out.

Thus, as you can see in Table 1, the odds of a student from a non-poor family dropping out are about 0.03, and the odds of a student from a poor family dropping out are 0.17. But odds are not perfect—predicted conditional odds can still be impossible—they go below 0.00. So the solution mathematicians and statisticians have come to is to take the natural logarithm of the odds, which has the benefit of having no restriction on minimum or maximum values. But before we move beyond odds, let's stop at the most commonly reported index of effect in logistic regression, the odds ratio.

The odds ratio. The conditional odds we have been discussing are the odds that an outcome (i.e., dropping out) will happen given a particular value of another variable (i.e., being below average in family income). As you can see in Table 1, those are interesting, but without something to compare it to, interpretation is difficult. So the odds ratio is used in logistic regression to represent the ratio of the conditional odds of the outcome at one level of *x* (for example, 1) relative to the conditional odds of the outcome at another level of *x* (for example, 0). In this way, the odds ratio (literally, a ratio of the odds of an outcome for two groups) helps us capture the effect of the independent variable. In our example in Table 1, we only have two levels of *x*: poor or not poor (1 or 0). If we calculate the ratio of those two odds, we get an odds ratio of 5.67 (0.17/0.03). the interpretation is straightforward (although as I discussed in (Osborne, 2006) there are common ways to misinterpret this number). In this example, the odds of students from "poor" households dropping out are 5.67 times that of students from "not poor" households. This is not a surprising statistic, given what we know of the importance of poverty in education.

In general odds ratios are calculated as the change in odds for every 1.0 increase in the IV. In the case of binary IVs, it is the comparison of those in the "1" group to those in the "0" group. In the case of a continuous IV, it would be the change in odds for each increase of 1.0 in the IV.

So to summarize, we have used simple division to move from numbers in boxes to the relatively important odds ratio statistic. Obviously things get more complex when there are multiple IVs in the equation, but conceptually everything is as simple as how we have discussed it thus far.

The logit. The natural logarithm of the odds is called the *logit*—the term that logistic regression derives its name from. Now we come to the crux of the issue—the initial question that prompted me to investigate this issue —what is the thing that logistic regression is really predicting? What is it exactly that we are graphing if we graph results from a logistic regression, and how do we interpret it coherently?

For those of you who are more than a few years removed from high school mathematics, let's do a *brief and painless* review of a logarithm before continuing. A logarithm is actually a class of mathematical operations where numbers as we are used to them can be represented by other bases. A logarithm is the power (exponent) a base number must be raised to in order to get the original number. Any given number can be expressed as y to the x power in an infinite number of ways. For example, if we were talking about base 10, 1 is 10^0, 100 is 10^2, 16 is $10^{1.2}$, and so on. Thus, $\log_{10}(100)=2$ ($100=10^2$) and $\log_{10}(16) = 1.2$ ($16= 10^{1.2}$). However, base 10 is not the only option for logarithms—you can literally use any number, although base 10 is one of the more common. Another common option is

the Natural Logarithm, where the constant *e* (2.7182818…) is the base.[3] In this case the natural log of 100 is 4.605 ($100 = e^{4.605}$).

Table 3. Examples of logarithms

Base:	10^6	10,000	100	1	0.01	0.0001	10^{-6}
2	19.93	13.28	6.64	0	-6.64	-13.28	-19.93
e	**13.81**	**9.21**	**4.60**	**0**	**-4.60**	**-9.21**	**-13.81**
3	12.58	8.38	4.19	0	-4.19	-8.38	-12.58
4	9.97	6.64	3.32	0	-3.32	-6.64	-9.97
5	8.58	5.72	2.86	0	-2.86	-5.72	-8.58
10	6.00	4.00	1.00	0	-1.00	-4.00	-6.00

As you can see in Table 3, the same number can be represented in a variety of ways across a variety of bases. Perhaps more germane to this discussion is the natural logarithm, of base *e*. If you notice, the natural logarithm of numbers above 1.0 grows from 0 toward infinity as the numbers being log transformed get larger. Interestingly, as numbers go from 1 toward 0, the log of those numbers becomes moves toward infinity in the negative direction (the log of 0 or a negative number is undefined).

You may also notice an interesting pattern in these numbers—the log of 100 and the log of 0.01 are identical except for the sign, as are the logs of 10,000 and .0001, and 1,000,000 and 0.000001. This is because in exponents, raising something to a negative power (n^{-1}) merely means to calculate 1/n. Thus, the interesting property of logs is that they "pivot" at 1.0—are essentially symmetrical around 1.0, and the log of 100 and 1/100 being identical except for the sign. This is an important revelation that will help with interpreting logistic regression output where odds ratios are below 1.0.

Summarizing so far

We started off with a hand calculation of simple probabilities and simple odds, and moved into the shortcomings of OLS regression in predicting dichotomous variables—aside from violations of all sorts of assumptions (usually), you can get predicted conditional probabilities (outside the 0 to 1 acceptable range) and conditional odds that are impossible (below 0.00). To handle these shortcomings, the natural logarithm of the odds can conceivably range from -∞ to ∞. Thus, if we use the logit (natural logarithm of the odds ratio) as our dependent variable we no longer face the issues that probabilities or odds have given us. The dependent variable then becomes logit(y), and the simple regression equation becomes:

$$\text{Logit}(y) = a + b_1x_1$$

which is the form that logistic regression takes. So with some division and a simple conversion of an odds ratio to a natural logarithm, we get a logit that solves the initial problem of having predicted probabilities or odds that are outside the possible range. There are a lot of technical details about why logistic regression uses maximum likelihood estimation rather than ordinary least squares estimation, but those issues are beyond the intended scope of this paper. There are two important things to note: (a) OLS regression is not appropriate under most circumstances when DVs are not continuous (technically, ordinal or interval), and (b) even when these assumptions of OLS regression are met, OLS regression and logistic regression using maximum likelihood estimation will produce identical coefficients (e.g., Menard, 2002). Thus, there seems to no significant drawback to using logistic regression where appropriate.

Still more fun with logits, odds, and probabilities

The *logit*, this metric of logistic regression, is the natural logarithm of the odds of something happening (whatever is 1 when the dependent variable is coded 0 and 1). The log of a number is difficult for most people who are not professional

[3] Sometimes referred to as Euler's number, but usually credited to Bernoulli, who attempted to solve the following formula which was applied to calculations of compound interest: $\lim_{n\to\infty}\left(1+\frac{1}{n}\right)^n$, *e* has applications in many fields beyond economics and statistics, including being particularly useful in calculus, probability theory, physical sciences, and beyond. It has been calculated to a trillion digits thus far, and like pi, is an enigmatic and interesting number.

mathematicians to comprehend in a deep way (or in an accurate way). So in logistic regression you are going to get these logits as the intercept and coefficients. But most statistical packages also provide odds ratios (sometimes abbreviated "OR," or in SPSS, labeled "Exp (B)") to make interpretation a bit simpler. It is important to recognize that these are all essentially the same bit of information, presented in slightly different form. If you have followed to this point, you can see each is a simple mathematical transformation of the other. Because of this, it is also relatively simple to reverse the process, and in reversing the process, we can bring significant clarity (and accuracy) to reporting our logistic regression findings. We can start with logits (again, the natural log of the odds of an outcome) and work our way back to conditional probabilities, which are generally easier for people to understand. This is particularly true for those of you who will be communicating to non-technical audiences (practitioners, policymakers, or the public) and is even useful when talking to other researchers who may not be as well-versed in logistic regression as you are.

From logit to odds ratio. Most statistical programs will present both logits and odds ratios. Below is a sample of the output from SPSS for this same data:

Table 4. SPSS logistic regression output for POOR and DROPOUT analysis

	B	S.E.	Wald	df	Sig.	Odds ratio
Poor	1.742	.073	566.339	1	<.001	5.711
Constant	-3.514	.066	2793.147	1	<.001	.030

Starting with the odds ratios, the constant is the predicted odds when X=0 (when students are *not* coming from poor households). The 0.03 should look familiar—in Table 1 we calculated the odds of dropping out when when POOR = 0 to be 0.03. This is the same number. And the logit for the constant (intercept) is -3.51, the natural log of 0.03. In other words, this is the natural log of the odds of dropping out if you are in the "0" category on the independent variable.

Now let's look at the variable of interest, POOR. The odds ratio is 5.71—which is within rounding error of what we calculated by hand. Converting to logits, the natural log of 5.71 is 1.742, which is what we see under the "B" column. If you have a calculator that can handle natural logs, exponents, and such (or access to Excel or similar spreadsheet programs) I encourage you to play with the output from your statistical software like this to help cement your understanding of the relationships between the numbers you are seeing on your output.

So to convert any logit to an odds ratio, we reverse the process. To get the natural log of a number, we raise *e* to a particular power.

Natural log of 5.71= $e^{1.74}$

And thus we say the natural log of 5.711 is 1.74. To reverse this, moving from logit to odds ratio, we *exponentiate* the logit—in other words to convert from logit to odds ratio we raise *e* to the logit power:[4]

$e^{1.74} = 5.71$

The importance of this seemingly simple process will hopefully become clear in a moment—but it clarifies why SPSS calls the odds ratio EXP(b)—if you exponentiate *b* you get the odds ratio.

Converting from odds ratio to conditional probability. In the same way we converted from conditional probability to odds ratio, we can reverse this process as well through two steps. Recall that to get from conditional probability to odds, we computed

conditional odds = P(dropout)/1-P(dropout) [=0.145/(1-0.145) or =0.029/(1-0.029)]

and then to compute an odds ratio, we divided one conditional odds by the other (0.17/0.03 = 5.67). To reverse engineer the process we can multiply the odds ratio by the conditional odds for the intercept (in the SPSS output this is the odds ratio multiplied by the EXP(B) constant, or 5.71 *

[4] Note that there is minute rounding error in all these calculations. If you are using a scientific calculator, excel or some similar process, you use the EXP(x) command, where x is the logit you want to convert back to an odds ratio.

0.03, which gets us back to 0.17, the conditional odds for the group of interest). To get from conditional odds to conditional probabilities divide the conditional odds by 1+ conditional odds:

Probability (dropout) = conditional odds / (1+ conditional odds) [0.17/(1+0.17)]

which leaves us with 0.146, which is within 0.001 rounding error of the original conditional probability we started off with back in Table 1.

More routinely, we will have predicted scores (predicted logits) for a group that we want to convert to a predicted conditional probability. Using the logistic regression equation from Table 4:

Logit' = -3.514 + 1.742(POOR)

We can calculate a predicted logit for poor students as -1.772. We can collapse all the steps above into one simple equation to convert predicted logits to conditional probabilities:

$$\text{Conditional probability of (Y=1)} = \frac{Exp(b)}{(1+Exp(b))}$$

Probability (dropout) = exp(-1.772)/(1+exp(-1.772)

Probability (dropout) = 0.145

which gets us back to the original hand-calculated conditional probability of students from poor households dropping out of high school. Likewise, we could perform the same calculation on the predicted logit of students from non-poor households and get back to the original conditional probability of that group as well.

Benefits of conditional probabilities. So why go through all these mathematical machinations? We already have what we want to know when we perform a logistic regression—what variables are significant predictors of the outcome, and the magnitude of the relationship (as well as direction), right? Yes, except that most of your audience won't intuitively understand odds ratios or logits. If you have an interaction effect or curvilinear effect in logistic regression and want to graph it, it is accurate and appropriate to graph it in logits, and explain what they are (natural logarithm of the odds). But what if you could graph the results as conditional odds or conditional probabilities (i.e., the probability that something will happen at a particular point of the independent variable for a particular group)? Wouldn't that be simpler to understand than the natural logarithm of the odds of the dependent variable being 1.0 at a particular point on the X axis?

Advantages of graphing logistic regression interactions as conditional probabilities

For this graphing example we are going to look at more data from NELS88—in this case, we will look at the same DV—DROPOUT—as a function of family socioeconomic status (SES, a continuous variable converted to z-scores so that the mean is 0.00 and the SD is 1.0) and student composite achievement test scores from 8th grade (ACH, also converted to z-scores).[5] A brief summary of the results from SPSS are presented in Table 5.

Table 5. SPSS logistic regression output predicting DROPOUT from ACH, SES

	B	S.E.	Wald	df	Sig.	Odds ratio
ACH	-1.174	.055	459.395	1	<.001	.309
SES	-0.857	.054	251.593	1	<.001	.429
ACH x SES	-.209	.051	16.597	1	<.001	.811
Constant	-3.174	.054	3458.32	1	<.001	.042

As you can see in Table 5, student achievement has a significant effect on DROPOUT, in that for every one standard deviation increase in achievement, the odds of dropping out decreases (logit = -1.17, OR= 0.31). SES also has a significant effect, in that for every one standard deviation

[5] A brief digression on continuous variables: I think it is most appropriate to convert all continuous variables to z-scores as (a) it centers them all at 0, which is valuable when looking at interactions, and (b) it converts them all to the same metric so that it is more straightforward to compare effects across variables.

increase in SES, the odds of dropping out decreases (logit = -0.86, OR=0.42). You can see by comparing logits[6] that ACH has a stronger effect on dropout than SES, but there is also a significant interaction between achievement and socio economic status. To explore the nature of this interaction, we can plot the interaction. This analysis gives us a prediction equation of:

$$Logit_{(Y=1)} = -3.174 - 1.174(ACH) - 0.857(SES) - 0.209(ACH*SES)$$

Choosing -2 to represent "low" and +2 to represent "high" for both IVs (again, because they are z-scores, that represents 2 SD below and 2 SD above the mean, which are reasonable to graph), we produce the following predicted logits, presented in Table 6 and graphed in Figure 2.

Table 6. Predicted logits and conversion to predicted probabilities

Group	Predicted Logits	Odds Ratio	Conditional Prob
Low ACH, Low SES	0.088	1.092	0.522
Low ACH, High SES	-1.74	0.176	0.149
High ACH, Low SES	-3.008	0.049	0.047
High ACH, High SES	-8.036	0.00032	0.00032

As you can see in Figure 2, logits remain relatively high for low SES students while logits drop for high SES students. High achieving students tend to have lower logits and the effect of SES appears to be stronger on them. From this graph, we would say that the natural log of the odds of dropping out tends to decrease as family SES increases, but that effect appears to be stronger for high achieving students. One of the things that is striking is about this graph is that low-SES students appear to drop out at relatively high rates regardless of achievement, and that for high-SES students, there appears to be a large gap between low- and high-achieving students.

What is striking about this graph is that it does not necessarily reflect what one sees in *actual* probabilities of dropout. The same data, graphed in 0.5 standard deviation increments and graphed in dropout probabilities separately by high- and low-achieving students (merely grouped into those below the mean and those above the mean for purposes of this exploration; note that there were too few high-achieving students at -2 SD or lower to graph) reveals a more intuitive and very different picture.

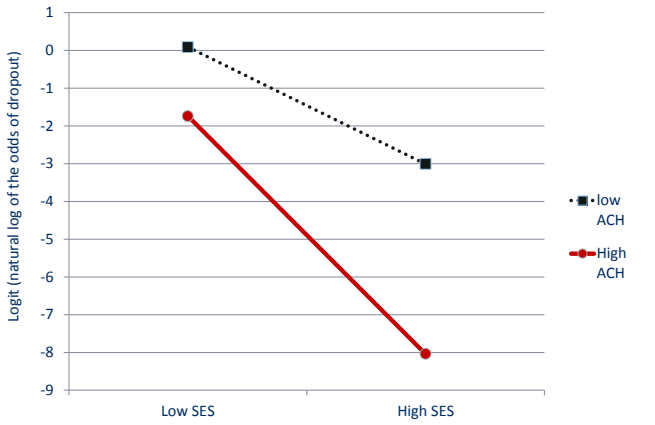

Figure 2: Interaction of achievement and family SES in logits

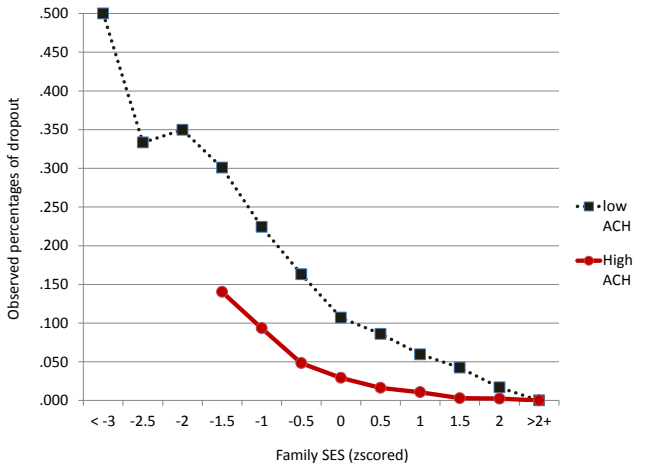

Figure 3: Observed probability of dropout

[6] Which is only possible because all continuous variables were standardized as z-scores

Converting Figure 2 to a more interpretable metric

I believe this highlights a second issue in using logits as a metric in graphing: logarithms can minimize very large differences (in $\log_{10}$ for example, the difference between 1000 and 10 is the difference between 3 and 1) and can also make small differences apparently large, especially as numbers asymptote toward 0 (in $\log_{10}$ again the difference between 0.01 and 0.000001 is the difference between -2 and -6). In other words, logits can make what for our purposes are very small differences in probabilities and make them look large, when graphed, and can minimize what are large magnitudes of difference. In the observed data, there is a real difference between high- and low-achieving students in dropout rates, and there is a real effect of family SES. Furthermore, there is an interaction between the two, but looking at the actual probabilities of dropout, it appears that achievement becomes more important as family SES decreases, and less important as family SES increases, which is a bit different than what we would conclude from Figure 2, looking at logits.

In Table 6 I have a brief summary of the calculations I used to convert these four logits to conditional probabilities, using the shortcut equation presented above.

$$\text{Conditional probability of (Y=1)} = \frac{Exp(b)}{(1+Exp(b))}$$

The same data converted to predicted probabilities (rather than logits) are presented in Figure 4. In my opinion, Figure 4 is a much better representation of the pattern of dropout in the observed data, and at the same time is easier for readers to interpret. For example, high achieving students have a lower probability of dropout regardless of SES, and low-achieving students have higher probability of dropout regardless of SES, but that difference is substantially more magnified for lower-SES students than for high-SES students. This interpretation is more closely aligned with the actual data. Further, the predicted probabilities are not far from the actual probabilities at -2 and 2 SD. If I had modeled a curvilinear relationship it is likely that the observed dropout probabilities would have been closer to the predicted dropout probabilities.

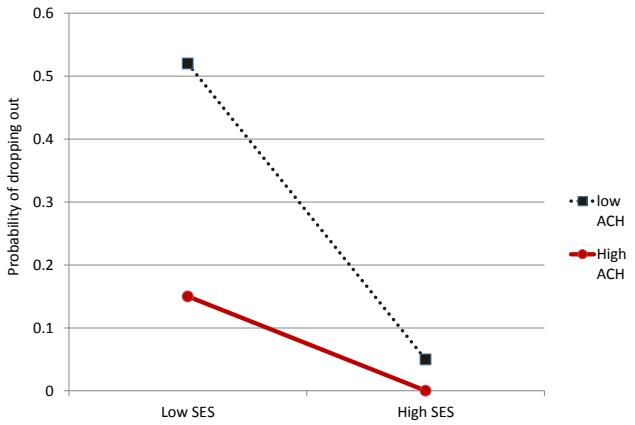

Figure 4: Interaction of achievement and SES predicting dropout graphed as probabilities

Summary

Once you understand some simple math of probability, odds, and logits, and how to convert between them, it becomes relatively straightforward to present the results from logistic regression analyses (particularly graphs) in metrics that consumers of your research can easily understand—conditional probabilities.

This is just one simple example, and it may not always make sense to make this conversion from logit to predicted probability. I think in the social sciences, it is more likely that this is a useful way to present the data, but researchers need to be thoughtful and careful about making decisions in presenting their data and results so that it is most easily understood and most likely to accurately represent the data to the reader.

References

Cohen, J., Cohen, P., West, S., & Aiken, L. S. (2002). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Mahwah, NJ: Lawrence Erlbaum.

Field, A. P. (2009). *Discovering statistics using SPSS*: SAGE publications Ltd.

Holcomb Jr, W. L., Chaiworapongsa, T., Luke, D. A., &

Burgdorf, K. D. (2001). An odd measure of risk: use and misuse of the odds ratio. *Obstetrics & Gynecology, 98*(4), 685.

Menard, S. W. (2002). *Applied logistic regression analysis* (Vol. 106): Sage Publications, Inc.

Osborne, J. W. (2006). Bringing balance and technical accuracy to reporting odds ratios and the results of logistic regression analyses. *Practical Assessment Research & Evaluation, 11*(7).

Osborne, J. W. (2011). Best Practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment Research & Evaluation, 16*(12), 1-7.

Pedhazur, E. J. (1997). *Multiple Regression in Behavioral Research: Explanation and Prediction.* Fort Worth, TX: Harcourt Brace College Publishers.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using Multivariate Statistics* (4th ed.). New York:: Harper Collins.

Citation:

Osborne, Jason W. (2012). Logits and tigers and bears, oh my! A brief look at the simple math of logistic regression and how it can improve dissemination of results. *Practical Assessment, Research & Evaluation*, 17(11). Available online: http://pareonline.net/getvn.asp?v=17&n=11

Author:

Jason W. Osborne, Ph.D.
Associate Professor, Educational Foundations and Leadership
120 Darden College of Education
Old Dominion University
Norfolk, VA 23529
919-244-3538
Jasonwosborne [at] gmail.com