


Please cite the Published Version

Voas, David  and Watt, Laura (2024) The odds are it's wrong: correcting a common mistake in statistics. Teaching Statistics. ISSN 0141-982X

DOI: <https://doi.org/10.1111/test.12391>

Publisher: Wiley

Version: Published Version

Downloaded from: <https://e-space.mmu.ac.uk/637020/>

Usage rights:  [Creative Commons: Attribution 4.0](https://creativecommons.org/licenses/by/4.0/)


Additional Information: This is an open access article which first appeared in Teaching Statistics

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

ORIGINAL ARTICLE

The odds are it's wrong: Correcting a common mistake in statistics

David Voas¹  | Laura Watt²

¹Social Research Institute, University College London, London, UK

²Department of Sociology, Manchester Metropolitan University, Manchester, UK

Correspondence

David Voas, Social Research Institute, University College London, London, UK, Email: d.voas@ucl.ac.uk

Abstract

Binary logistic regression is one of the most widely used statistical tools. The method uses odds, log odds, and odds ratios, which are difficult to understand and interpret. Understanding of logistic regression tends to fall down in one of three ways: (1) Many students and researchers come to believe that an odds ratio translates directly into relative probabilities. (2) Alternatively, they learn that coefficients tell us whether the variables make the outcome more or less likely, without knowing how to interpret changes in the odds. (3) They may be instructed in how to calculate predicted probabilities, but the additional steps are too complicated for them to follow. Our key aim is to highlight and correct the common mistake of confusing differences in odds with relative risks. Simply reporting the odds ratio is unhelpful, however, so we describe an easy method of estimating probabilities for both binary and continuous variables.

KEYWORDS

log odds, logistic regression, odds ratio, predicted probability, relative risk, teaching statistics

1 | INTRODUCTION

Regression is probably the most widely used method in inferential statistics. Linear regression is one of the first techniques that students learn, and binary logistic regression tends to follow swiftly. That is understandable, because we often want to predict a binary outcome: win/lose, pass/fail, mover/stayer, infected/not, and so on. Logistic regression generates odds ratios, and “there is a problem with odds: unlike risks, they are difficult to understand.”¹

If odds ratios were merely viewed as mysterious signs that a variable has a positive or negative influence on the outcome, with no further attempt at interpretation, the situation would be regrettable but not disastrous. Unfortunately, “most people misinterpret odds ratios as risk ratios,” that is, as relative probabilities, in the view of

Norton et al.^{2, p. 492} The assertion that an exponentiated coefficient of 2.0 means that increasing the predictor variable by 1 makes the outcome twice as likely is easy to find in online teaching material, never mind student papers.

The authors of the key publications in the Sage Quantitative Applications in the Social Sciences series (popularly known as the “little green books”) are aware of this problem. In the second edition of *Logistic Regression: A Primer*, Fred Pampel writes

In interpreting the exponentiated coefficients, remember that they refer to multiplicative changes in the odds rather than probabilities. It is incorrect to say that an additional year of education makes smoking 16.7% less probable or likely, which implies

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *Teaching Statistics* published by John Wiley & Sons Ltd on behalf of Teaching Statistics Trust.

probabilities rather than odds. More precisely, the odds of smoking are .833 times smaller or 16.7% smaller with an additional year of education.

[3]

Similarly, Scott Menard reveals some frustration in a passage added to the second edition of *Applied Logistic Regression Analysis*:

I have repeatedly seen the mistake of equating the odds ratio (a ratio of two odds) with a risk ratio (a ratio of two probabilities), sometimes with the justification that the two are ‘approximately’ equal under certain fairly restrictive conditions (a base rate less than .10). In general, the use of an odds ratio to ‘represent’ a risk ratio will overstate the strength of the relationship. An odds ratio of about .22 for males ... does not mean that the risk of marijuana use is only a little over one-fifth as high for males as for females or that the odds ratio of 4.5 for females ... indicates that the risk of marijuana use is nearly five times as high for females as for males. To compare the relative risk of marijuana use for males and females, it is necessary to use the model to calculate the probabilities for each, assuming values of the other predictors. ... Suggestion: Do the math. There is no excuse here for approximations that can so easily be misleading.

[4]

These warnings seem clear, but they are brief and buried in the middle of long chapters on interpreting logistic regression coefficients. The problem needs to be made much more visible to students, teachers, and researchers. The idea that odds ratios express relative risks is persistent and (as we shall see below) highly misleading except when the probabilities in question are very low (less than 0.1 or so).

The misinterpretation of odds ratios has been a problem for decades. In 1991, Roncek highlighted the issue in the social sciences, pointing out that an article published the previous year in *Social Forces* (then one of the top three journals in sociology) presented findings in which odds ratios were described as “times as likely,” in other words as relative risks.⁵ The situation did not improve in the years that followed, leading one scholar to conclude a major review of a more technical problem in logistic regression with the statement that “odds ratios are frequently misunderstood as relative risks, so it is often a

good choice to present at least one effect estimate in terms of effects on probabilities.”⁶, p. 80 In the bio-medical arena, Holcomb et al.⁷ reviewed all of the articles published in 1998–1999 in the two leading journals in obstetrics and gynecology, finding that “Of 151 studies using odds ratios, ... In 39 (26%) articles the odds ratio was interpreted as a risk ratio without explicit justification.”

Even textbooks can be misleading. The first edition of *Statistics at Square Two* sets out the statistical concepts—even mentioning that odds ratios are not relative risks—but then interprets odds ratios using “times as likely” language when giving two practical examples concerning breastfeeding and hypertension.⁸, pp. 43–45 The breastfeeding example was corrected in the second edition, and the odds ratio is explicitly contrasted with relative risk, but the mistake persists in the other example: Based on an odds ratio of 2.24, “we would predict that the older subject would be 2.24 times more likely to have hypertension.”⁹, p. 41 Finally, in the third edition, the examples have been changed and the authors stress the difference between odds ratios and relative risks.¹⁰

The bestselling *Discovering Statistics using IBM SPSS Statistics* includes problems at the end of each chapter, with answers on the companion website. One task involves using logistic regression to analyze the factors that influence condom use, and the resulting odds ratios are interpreted as showing that the variable makes the outcome that many times more or less likely.¹¹, pp. 933–934, 12, Task 20.6 The same problem appeared in earlier editions, which suggests that no one has ever brought the issue to the attention of the author and publishers.

Even the excellent Sage little green book on logistic regression, possibly the best introduction to the topic, creates some confusion with an example of men’s and women’s differing levels of support for the legalization of marijuana.

According to the 2016 GSS ... The ratio of odds of men to women equals 1.93/1.34 or 1.44. This odds ratio is a group comparison. It reflects the higher odds of supporting legalization for men than women. It means specifically that 1.44 men support legalization for each woman who does.

[3]

The author of the book acknowledges that “It would have been better to say ‘It means specifically that odds for men of supporting legalization are higher by a factor of 1.44 than for women’” (F. Pampel, personal communication, January 2, 2024).

Our point is not to criticize individual authors but to illustrate how easy it is to misinterpret odds ratios—a

tendency that has no doubt been encouraged by the fact that odds ratios are approximately the same as risk ratios when the underlying probabilities are very small, as they often are in applications involving rare medical conditions. Unfortunately, the mistake is so pervasive that web searches using terms such as “interpreting logistic regression coefficients” do not reliably lead to clear and correct explanations; an unwary scholar is likely to find material that is misleading, confused, or confusingly technical.

Some researchers have sufficient skill in statistics to navigate these difficulties. In the social sciences, however, many students have limited mathematical knowledge and may even resist training in quantitative methods for ideological reasons.^{13,14} In any case, it can be challenging for them to grasp logistic regression, a tool based on equations involving e , natural logarithms, and the overlapping concepts of probabilities, odds, log odds, and odds ratios. Misunderstandings can easily take root—particularly when the mistaken interpretation is less complicated than the true one.

A partial solution is to teach students that odds ratios cannot be interpreted as relative probabilities. They can learn that coefficients tell us whether the variables make the outcome more or less likely, without knowing how to interpret changes in the odds. Many textbooks take this approach, but it is hardly very satisfying.

A better solution is to use the output from logistic regression to calculate probabilities. The initial barrier is the fact that relative risks, unlike odds ratios, change depending on the values of the predictors and the dependent variable. Once that is understood, the usual method is to specify a typical case as the baseline, but the subsequent calculations are laborious to perform manually and require further training if done by statistical software.

To summarize, understanding of logistic regression tends to fall down in one of three ways:

1. Many students and researchers come to believe that an odds ratio translates directly into relative probabilities.
2. Alternatively, they learn that coefficients tell us whether the variables make the outcome more or less likely, without knowing how to interpret changes in the odds. They will correctly write that “the odds for women are twice as high as the odds for men,” but such a statement means little on its own.
3. They may be instructed in how to calculate predicted probabilities, but the additional steps are too complicated for them to follow.

Our key aim is to highlight and correct the common mistake of confusing differences in odds with differences

in probabilities. Simply reporting the odds ratio is unhelpful, however. Predicted probabilities are much more informative, and students using statistical software like R or Stata should be taught how to produce them. SPSS it is poorly suited to this task, but it continues to be the package most widely used by academic researchers,^{15,16} particularly for teaching purposes and in the social sciences. Moreover, scholars often encounter journal articles that report on odds ratios but not predicted probabilities. We describe a simple method for estimating them for both binary and continuous independent variables.

2 | UNDERSTANDING ODDS AND ODDS RATIOS

2.1 | Odds

However one introduces logistic regression—and we suggest that it is best not to start with the technical material—it is necessary to talk about odds at an early stage. The odds are the probability p that something happens divided by the probability $(1 - p)$ that it does not. People who bet on horse races use them all the time, but for everyone else, odds seem odd.

Most of us are accustomed to thinking of how likely or probable an event is as a percentage, as in “there is a 25% chance of rain today.” That is equivalent to saying that we could expect rain in these circumstances 25 times out of 100.

A different way of expressing the same thing is that there is one chance in four that it will rain, which means that there are three chances in four that it will stay dry. The odds are one to three that it will rain, or to flip things on their head, three to one that it will be dry.

If you want to bet on a horse race, most bookies will make you an offer expressed as the odds that your horse will lose, such as 8 to 1 against Blaze winning the race. Why do it this way, instead of saying that the chances are one in nine that it will win? One reason is that the odds tell you how much you stand to gain in relation to what you risk. If you gamble £1 and your horse comes in, you will win £8 that you did not have before (as well as getting back your £1 stake).

It is not hard to convert probabilities into odds and vice versa, but we have to be aware that the two values are different. Let us go back to the 25% chance of rain. Probabilities are traditionally expressed as proportions (values between 0 and 1), so the probability of rain is 0.25. The odds are 1/3, which we can treat as a fraction, so the decimal value is 0.33. That is already a

considerable difference between probability and odds, but the gap is much larger when it comes to events that are especially frequent. The probability that it will stay dry is 0.75. The odds are 3/1 or 3.0.

Why does this difference increase so much as the event becomes more likely? Probabilities are all between 0 (impossible) and 1 (certain), but there is no limit to how high the odds can be. As an event becomes more probable, the chances of it happening divided by the chances that it will not happen go up and up. For example, a probability of 0.15 produces odds of $0.15/(1-0.15) = 0.18$, a probability of 0.85 gives odds of $0.85/(1-0.85) = 5.7$, a probability of 0.98 means odds of $0.98/0.02 = 49$, a probability of 0.999 takes us to odds of $0.999/0.001 = 999$, and so on towards infinity. Note the absence of a straight line (or linear) relationship between probability and odds. One can see that probability and odds are similar for very small values (up to about 0.1), but they can be very far apart when the values are much larger.

2.2 | Odds ratios and avoiding the big mistake in logistic regression

Statistical software like SPSS typically produces output for logistic regression that includes a list of independent variables with the coefficient B in one column and $\text{Exp}(B)$ in another. The exponentiated coefficient gives us the odds ratio, which could be a value below 1 (meaning that the variable makes the outcome less likely) or greater than 1 (meaning that the variable makes the outcome more likely).

Textbooks and research articles often go no further than providing a basic interpretation of the odds ratio: If it is 2.5, for example, a unit change in the independent variable increases the odds of the outcome by two and half times; if it is 0.8, a unit change in the variable reduces the odds of the outcome by 20%. Such statements are at least correct, but no one has an intuitive sense of how to interpret those values, which makes the exercise seem slightly pointless.

To do better, it helps to start with an example of how odds and odds ratios work. Let us say that we want to know whether women are more likely to pass a particular course, all else being equal. The dependent variable is pass/fail; the independent variable of interest is gender, with 0 = male and 1 = female.

Assume that 60% of men pass a course, and so 40% fail. That means that for a man, the odds of passing are $0.6/0.4 = 1.5$. Now suppose that we run a logistic regression, and the exponentiated coefficient for gender (remember that female = 1) is 2.0. That value for $\text{Exp}(B)$

is the odds ratio—the odds of passing for women divided by the odds of passing for men:

$$\frac{\text{odds of passing for women}}{\text{odds of passing for men}}$$

So, the exponentiated coefficient of 2.0 tells us that the odds of passing are twice as high for women as for men.

Here is where we often encounter the most common error in statistical interpretation. The coefficient does not mean that women are twice as likely to pass as men. We are dealing with odds, not probabilities. Given that 60% of men pass, it would obviously make no sense to assert that women are twice as likely to pass.

Recall that for men, the odds of passing are 60/40, or 1.5. Since the odds ratio for women relative to men is 2.0, the odds of passing for women are $1.5 \times 2.0 = 3.0$.

If the odds of passing for women are 3.0, that means that they are three times as likely to pass as to fail:

$$\frac{\text{probability of passing (for women)}}{\text{probability of not passing (for women)}} = 3.0.$$

The only question remaining is how likely women are to pass. By definition, $\text{odds} = p/(p-1)$. A few lines of algebra transform this equation to:

$$p = \frac{\text{odds}}{1 + \text{odds}}.$$

And so we can work out the probability from the odds. Since we know that the odds are 3, it is easy to calculate that the probability is $3/4 = 0.75$. That makes sense, because if the probability of passing is 0.75, the probability of failing is 0.25, which means that the odds are $0.75/0.25 = 3.0$. To summarize, our logistic regression model predicts that 75% of women will pass.

In this situation, the odds of passing are twice as high for women as for men: the $\text{Exp}(B)$ in the logistic regression output is 2.0. But women are only 25% more likely than men to pass (because their projected pass rate is 75% rather than 60%, and $75/60 = 1.25$). Interpreting a difference in the odds as a difference in probabilities would be a serious mistake.

If the event represented by the dependent variable is fairly rare, the odds ratio and relative risks are similar. For example, if 10% of men and 15% of women have experienced depression, the ratio of women's odds to men's odds is $(15/85)/(10/90) = 1.59$. The odds are 59% higher for women, and the probability is 50% higher (0.15 vs. 0.10), which is fairly similar. As the likelihood of the outcome rises, however, the divergence grows. The pass/

fail scenario described above is an example. When the chances of the event occurring are very high, the odds ratios are typically large while the difference in probabilities becomes small. For example, suppose that 98% of women pass a course, compared to 92% of men. The odds ratio is $(98/2)/(92/8) = 4.26$, but women are obviously not four and a quarter times more likely than men to pass.

In the main example above, gender had been coded with male = 0, female = 1. A curious student might wonder whether the results would be different if the coding had been female = 0, male = 1. Such a question offers the opportunity to show that only the calculations along the way would change. This time our starting point would be the probability of passing for women, which is 0.75, and so, their odds of passing are $0.75/0.25 = 3.0$. The logistic regression would now give us 0.5 as the odds ratio for men relative to women:

$$\frac{\text{odds of passing for men}}{\text{odds of passing for women}}$$

That means that the odds of passing for a man are half those for a woman. Once again, it would be a mistake to jump to the conclusion that men are half as likely to pass: odds are not probabilities. Since the odds for women are 3.0, we know that the odds for men are $3.0 \times 0.5 = 1.5$. And because $\text{odds}/(1 + \text{odds})$ gives us the probability, we can work out that the probability of passing for men is $1.5/2.5 = 0.6$. Thus 60% of men pass, which was where we started.

It does not matter whether we look at women relative to men or men relative to women; the odds ratio will be flipped over (2 in one case, $\frac{1}{2}$ in the other), but we arrive at the same result. In either case, it is important to avoid confusing odds and probabilities: women are not twice as likely to pass the course as men, nor are men only half as likely to pass as women.

3 | INTERPRETING THE COEFFICIENTS

3.1 | Straight lines versus S-shaped curves

By this point, everyone should understand that an odds ratio of X does not mean that the predictor makes the outcome X times more likely. Saying that the odds are X times higher, though, is correct but not very useful. We need to translate the results back into probabilities. To take that step, it helps to have a basic grasp of how logistic regression works.

With linear regression, the concept is simple: you have a variable (like a measure of health and wellbeing) that you want to explain, and at least one variable (like number of cigarettes smoked each day, or hours spent on social media each week) that may have an influence on it. You can create a scatterplot with health along the vertical axis (high is good, low is bad) and smoking or social media use along the horizontal axis, so there is one dot for every person in your dataset. If you draw a line through the dots in a way that fits the best—that is, the line is as close as possible to as many points as possible—you have done linear regression.

The key point about the linear or “straight line” model is that the effect of an independent variable on the dependent variable is constant at all levels. The interpretation of a coefficient in linear regression is simple: it is the change in the outcome that results from increasing the value of the independent variable by 1. No other information is needed, and no calculations are required.

Imagine, for example, that the dependent variable in linear regression is a happiness scale and the independent variable of interest is a measure of health, with controls for age, gender, and so on. If the coefficient for health is 2, that means that every additional point on the health scale adds 2 points to the happiness scale. The 2-point boost holds for young and old, men and women, and at all levels of happiness, high and low. One can state the effect of health on happiness without knowing anything else.

In binary logistic regression, the dependent variable—the thing we want to explain—takes one of two values: yes or no, good or bad, pass or fail. One might think that the binary outcome would make things simpler, but in fact, it complicates matters. Instead of predicting the value of the dependent variable, we now aim to predict the probability that it is 1 rather than 0. And instead of fitting a straight line (the “linear model”), we are going to use an S-shaped curve (the “logistic model”).

In its basic form, the logistic curve starts very close to 0, rises gradually, and then more rapidly before leveling off, ending up very close to 1. The effect of an independent variable on the probability that the dependent variable equals 1 is *not* constant at all levels. The effect is greatest when the probability is 0.5 and smallest at the extremes.

Consider the example of sporting events with a binary outcome: win or lose, no draws (ties). The predictors of whether a team wins a particular match include whether they are at home or away. We can ask how much home advantage increases the probability of winning, but the answer depends on the base probability, taking everything else into account. If Liverpool is playing another team that is equally strong, being at home might be a

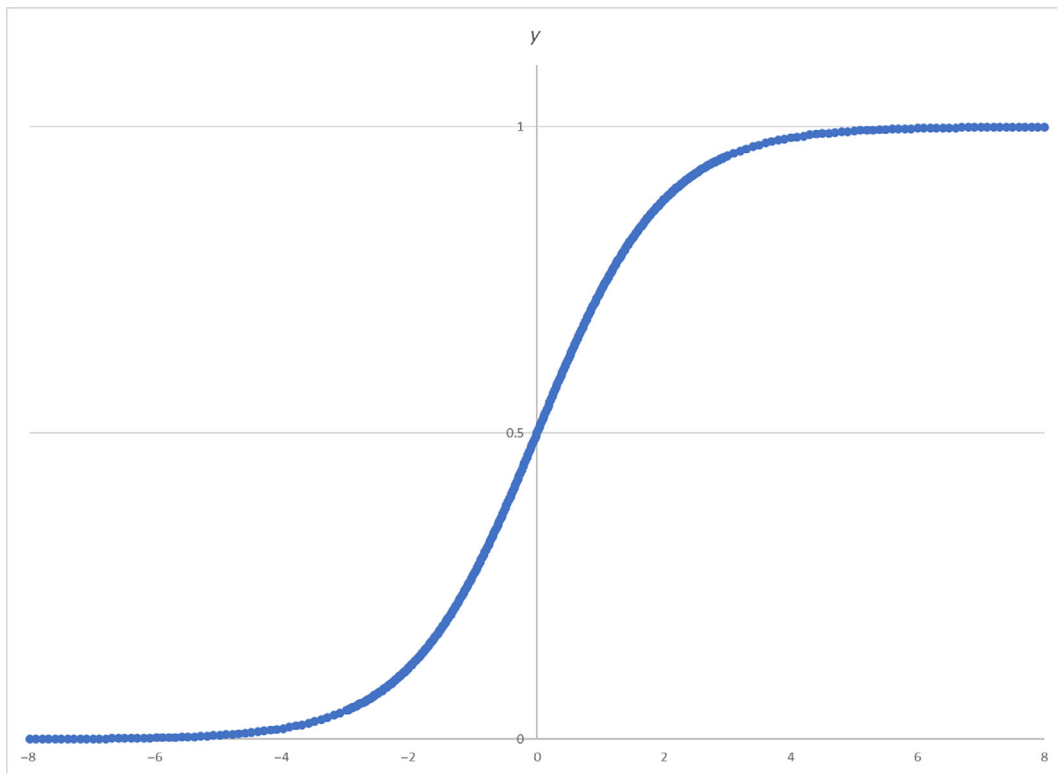


FIGURE 1 A simple logistic curve. [Color figure can be viewed at wileyonlinelibrary.com]

substantial advantage. If Liverpool were to play an amateur team, the venue would have little influence on the outcome.

The exact form of the logistic curve will be determined by the data to which the model is fitted. The basic pattern is always the same, however, as shown in Figure 1. There is an S-shaped curve representing—as a function of the independent variable(s)—the probability that the dependent variable equals 1. The probability is bounded by 0 and 1, which is an important reason the line on the graph is curved rather than straight. Shifting an independent variable has most influence on the probability when the binary outcome might go either way; it is least influential when the initial probability is close to 0 or 1. The logistic function is illustrated in Figure 1, and the underlying mathematics are outlined in the Appendix A.

3.2 | Going back to probabilities

We can convert the findings from logistic regression into probabilities, but the point made in the previous section is crucial: there is no single answer. The logistic curve is not a straight line, so the slope (the rate of change in probability as the x -variable changes) is not

constant. We can say how much difference an increment in the independent variable has on the probability of the outcome, but only with reference to a particular situation.

For example, if we are going to predict how a change in health affects the probability of being happy, we have to know how likely the person is to be happy anyway. But being happy or unhappy is the *dependent* variable: it depends on age, gender, income, and everything else, including how healthy the person is already. We therefore have to specify all of those values to calculate the probability of being happy (given that particular set of characteristics) and then how much that probability would go up or down if there is a change in health. The answer may be different if we chose another kind of person as the reference case.

The usual recommendation is to select an average or a “typical” case (such as a single white woman aged 20, employed, and so on) and then to insert all of the values of the independent variables into the logistic regression equation and do the calculations required to come up with an answer. Once we work out the probability of being happy for the baseline case, we can compare it to the probability we find when health goes up by a point. There are routines in R and Stata that will automate the process, but many students and researchers do not use those packages.

A simple alternative is to set all of the independent variables to zero, but the result is likely to be unrealistic unless the variables are standardized (so that zero represents the mean value). The predicted probability of passing a course for someone aged 1 rather than 0 is not helpful to know.

A third and arguably better option is to use the actual frequency from the sample. To return to the example of passing a course, a basic crosstab will give us the percentage of men who pass. By calculating $p/(1-p)$, that percentage can be turned into the odds of passing for men. Multiplying these odds by the odds ratio from the logistic regression gives the odds of passing for women, controlling for the other independent variables, and those odds can then be turned into a predicted probability or percentage.

Remember that the probability of passing depends on the individual's characteristics. If we follow the procedure just described, we would be predicting the outcome for women if they were like the general sample in every respect other than gender. The sample distribution by age, qualifications, ethnicity, family income, and so on might not be exactly what we want (though weighting can help), but whatever it is, the method gives us the approximate pass rate for female students compared with male students who were otherwise similar.

As shown earlier, we could just as easily compare men to women as women to men. The reciprocal of the odds ratio just considered represents the odds for men relative to those for women. Multiplying it by the observed odds for women and then converting odds to probability would give us the approximate pass rate for a set of male students who resembled the female students in all respects except for gender.

The take-away lesson should be reassuring for scholars who feel mystified by odds and odds ratios. The odds ratio does not mean much on its own. If we are content with knowing whether a particular predictor has a positive or negative effect (and whether that influence is statistically significant), then the logistic regression output will tell us. If we want more insight into the size of the effect, we need to convert odds back into probabilities.

3.3 | Predicted probabilities with binary variables

Descriptions of logistic regression often stop at odds ratios. When textbooks do explain how to calculate predicted probabilities from the model, the procedure typically involves lengthy calculations using the full regression equation with values assigned to every

TABLE 1 Pass rates by gender (%).

	Men	Women
Pass	62	70
Fail	38	30
Total	100	100

independent variable. The alternative method just described is easy to use, but it is worth reviewing the steps carefully.

If we are interested in the effect of a binary independent variable, the starting point is to run a crosstabulation with the dependent variable (Table 1). In our example, we have been looking at the effect of gender on a pass/fail outcome. The table gives us pass rates for both men and women, but they might differ in relevant ways, potentially including age, qualifications, ethnicity, family income, and so on. We performed a logistic regression in order to control for those socio-demographic differences. What we are aiming to do now is to take the percentage of men who pass and then estimate, for a hypothetical set of women whose other characteristics are the same, the percentage who will pass.

Let us label the probability a man will pass as p_0 ; the probability that a woman will pass is p_1 . Take the following steps:

1. Find the proportion of men who pass, p_0 , from a crosstab (converting percentages into proportions, e.g., 62% becomes 0.62).
2. Calculate the odds that a man will pass, which is $p_0/(1 - p_0)$.
3. Multiply that value by the odds ratio (the $\text{Exp}[B]$ from the logistic regression output); the result is the odds that a woman will pass.
4. Calculate $\text{odds}/(1 + \text{odds})$, which is the probability that a woman will pass, p_1 . It will differ from the value in the pass-by-gender crosstab if the other independent variables in the regression model have accounted for some of the gender gap.

With the pass rate by gender in hand, along with the output of the logistic regression, the entire calculation can be done in less than a minute. An alternative formula is described in the next section.

Recall that the object of logistic regression is to control for other factors. What we have done is to calculate a pass rate for women that can be compared to what was found for men, controlling for other variables included in our regression model. It is only an approximation—the most rigorous approach is to predict probabilities using a set of specified characteristics—but it is a quick and easy

way of estimating the magnitude of the variable's effect (here, the effect of gender on the probability of passing a course). Published articles that show odds ratios and confidence intervals would benefit from including some information about relative risks, even if they do not go as far as providing a full analysis of risk ratios, marginal effects, or predicted probabilities.^{2, 3, ch. 2}

3.4 | Predicted probabilities with continuous variables

If the independent variable of interest is not binary, the calculations are a little different. With a binary predictor, as described above, we start from the relative frequency of the outcome variable (like passing the course) for one value or the other (like male or female). With a continuous variable, we use the mean of the outcome variable as the baseline. For example, we might be interested in how the number of hours spent studying each week affects the chances of passing. If 68% of students pass the course, the odds of passing are $68/32 = 2.125$.

Suppose that on average students devote 15 h per week to their studies. The regression coefficient tells us what happens when that independent variable is incremented by one unit. In this example, it will give us the change in the odds for an additional 1 h per week of study. Suppose that the odds ratio is 1.2. Multiplying the odds of passing if you study for 15 h by the odds ratio gives us the odds of passing if you study for 16 h: $2.125 \times 1.2 = 2.55$. We can translate that back into a probability using the formula $\text{odds}/(1 + \text{odds})$, so $2.55/3.55 = 0.718$ is the probability of passing if you study for 16 h. An additional hour of study beyond the average, then, raises the chances of passing from 68% to almost 72%. An alternative approach to estimating the change in probability is described in the next section.

In some instances, a single unit increase will not be a useful indicator. If the variable is a scale from 0 to 100, for example, a change from 42 to 43 may not have much impact. In the case above, we might not be particularly interested in the marginal change represented by an additional hour per week. How much difference would it make if a student committed to studying 5 h more than the average: 20 h per week rather than 15? The odds of passing are multiplied by 1.2 for each additional hour of study, so the new odds ratio is $1.2 \times 1.2 \times 1.2 \times 1.2 \times 1.2 = 1.2^5 = 2.488$. That means that the odds of passing are now $2.125 \times 2.488 = 5.288$. Converting that into a probability, we have $\text{odds}/(1 + \text{odds}) = 5.288/6.288 = 0.841$. The extra 5 h a week have lifted the chances of passing from 68% to 84%.

It is interesting to compare what happens to the odds ratios and the predicted probabilities of passing with an extra 1, 5, or 10 h of study per week. The odds ratio quickly becomes very high, going from 1.2 to 1.2^5 to 1.2^{10} , that is, from 1.2 to 2.488 to 6.192. The predicted probabilities, by contrast, show that additional study has diminishing returns. The extra 1, 5, or 10 h take the probability of passing from 0.72 to 0.84 to 0.93. Reliance on odds ratios would give a misleading impression—though we encourage students to work hard!

4 | ALTERNATIVE WAYS OF ESTIMATING RELATIVE RISK

4.1 | Binary variables

A different route to obtaining the same result when the predictor variable is binary is the formula below, where OR is the odds ratio and p_0 is the probability of the outcome when the independent variable equals 0:

$$\frac{p_1}{p_0} = \frac{\text{OR}}{\text{OR} * p_0 + 1 - p_0}.$$

The expression p_1/p_0 is generally called the risk ratio or relative risk. This formula can be found in some articles and textbooks; it is usually attributed to Zhang and Yu,¹⁷ though it was anticipated by the “Mantel-Haenszel adjusted risk difference” in a technically sophisticated article by Holland.^{18, p. 1014}

If the aim is to estimate p_1 , as discussed above, we would calculate

$$p_1 = \frac{\text{OR} * p_0}{\text{OR} * p_0 + 1 - p_0}.$$

The problem with this formula is that it results from a very convoluted algebraic exercise—see the appendix in Shrier and Steele¹⁹—and so becomes yet another black box for producing a result. By contrast, our preferred approach helps to reinforce an understanding of the process at work: Take the probability of the outcome for the reference case ($x = 0$), calculate the odds, multiply by the odds ratio from logistic regression to obtain the odds for the comparison group ($x = 1$), and then convert those odds into a probability. Thus,

$$\text{Odds}_{x=0} = \frac{p_0}{(1 - p_0)},$$

$$\text{Odds}_{x=0} * \text{OR} = \text{Odds}_{x=1},$$

$$p_1 = \frac{\text{Odds}_{x=1}}{1 + \text{Odds}_{x=1}},$$

Which route one chooses to take is a matter of individual preference, of course.

The value obtained for p_1 should be regarded as an approximation. Some statisticians argue that because there is some bias and it is hard to calculate correct confidence intervals, the Zhang and Yu formula (which is equivalent to our procedure) should not be used. The contrary view, which we share, is that the best should not be the enemy of the good. The alternatives are sophisticated statistical methods that will be inaccessible to many students and researchers.

4.2 | Continuous variables

If the independent variable of interest is continuous, a different option is available. The change in the probability of passing can be calculated using a simple formula, but the method will be opaque unless the user is familiar with differential calculus. Again, the means are the starting point: students devote 15 hours a week on average to studying; the probability of passing is 0.68, and hence, the probability of not passing is 0.32. Instead of working out the odds and then multiplying by the exponentiated coefficient (the odds ratio), however, we can calculate the following: the coefficient $B \times$ the probability of passing \times the probability of not passing (see Agresti,^{20, p. 164} but note that the formula has been known for at least a century: see Pearl and Reed^{21, p. 365}). The result is the rate of change in the probability of the outcome at that specific point on the curve, or approximately, the amount by which a one-unit change in the independent variable alters the probability.

In the case discussed above, where the logistic regression output told us that the odds ratio was 1.2, it would also have given the coefficient B (prior to being exponentiated) as 0.1823. The formula would therefore be $0.1823 \times 0.68 \times 0.32 = 0.0397$. In other words, an extra hour of study would increase the probability of passing by almost 0.04, from 0.68 to about 0.72. This procedure produces what we want without using odds.

We are attracted to the slightly longer process (calculating odds, multiplying by an odds ratio, and then converting the result back from odds into a probability) because it is consistent with the recommended method for binary independent variables and the logic is clear. The alternative (going directly from the coefficient) is simple, but it compounds the inscrutability of logistic regression unless one understands the concept of finding

the derivative for a point on a logistic curve. It also becomes more difficult to estimate the probability of the outcome for a larger increase in the continuous variable because the amount by which the probability shifts is constantly changing.

5 | CONCLUSION

Our main aim has been to highlight a common error in presenting the results of binary logistic regression: Odds ratios must not be interpreted as relative probabilities. We need to ensure that students and researchers do not fall into this trap.

Relatedly, though, it is hardly satisfactory to say that a particular independent variable has a significant effect without giving some sense of the size of that effect. As the odds ratio itself does not express the magnitude of the change in likelihood, we need to translate the results into probabilities in some other way. The conventional approach is to calculate the probability of the outcome by putting average or characteristic values of all of the independent variables into the regression equation (so that age = 20, hours studied = 15, and so on); that probability can be compared with the value obtained when just one of the predictors is changed (for example, by increasing age or the hours studied). The calculations are laborious to perform by hand, however, and require further training if the job is done using R or Stata. We have described (in sections 3.3 and 3.4) a quick and easy way of estimating a particular variable's impact on the probability of the outcome when the other independent variables take their average values (or frequencies, in the case of dummy variables). We recommend this approach for students who are using SPSS rather than statistical software that includes straightforward routines for predicting probabilities. It can also be useful when looking at published output that gives only odds ratios and descriptive statistics.

Finally, the underlying problem is that logistic regression is based on a nonlinear model that is challenging for people with limited mathematical backgrounds to grasp. The concept of odds requires only basic numeracy, however. Going further requires knowing about natural logarithms, but with that knowledge in hand, it is not difficult to understand logistic regression (as shown in the Appendix A). The methods described above are accessible to all: anyone can make sense of odds ratios if they take a moment to do a quick conversion into probabilities.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

David Voas  <https://orcid.org/0000-0003-4094-1369>

REFERENCES

1. H. T. O. Davies, I. K. Crombie, and M. Tavakoli, *When can odds ratios mislead?* Br. Med. J. **316** (1998), 989–991.
2. E. C. Norton, M. M. Miller, and L. C. Kleinman, *Computing adjusted risk ratios and risk differences in Stata*, Stata J. **13** (2013), no. 3, 492–509.
3. F. C. Pampel, *Logistic regression: a primer*, 2nd ed., Sage, London, 2021.
4. S. Menard, *Applied logistic regression analysis*, 2nd ed., Sage, London, 2001.
5. D. W. Roncek, *Using logit coefficients to obtain the effects of independent variables on changes in probabilities*, Soc. Forces **70** (1991), no. 2, 509–518.
6. C. Mood, *Logistic regression: why we cannot do what we think we can do, and what we can do about it*, Eur. Sociol. Rev. **26** (2010), no. 1, 67–82.
7. W. L. Holcomb, T. Chaiworapongsa, D. A. Luke, and K. D. Burgdorf, *An odd measure of risk: use and misuse of the odds ratio*, Obstet. Gynecol. **98** (2001), no. 4, 685–688.
8. M. J. Campbell, *Statistics at square two: understanding modern statistical applications in medicine*, BMJ Books, London, 2001.
9. M. J. Campbell, *Statistics at square two: understanding modern statistical applications in medicine*, 2nd ed., Blackwell, Hoboken, NJ, 2006.
10. M. J. Campbell and R. M. Jacques, *Statistics at square two: understanding modern statistical applications in medicine*, 3rd ed., Wiley Blackwell, Hoboken, NJ, 2023.
11. A. Field, *Discovering statistics using IBM SPSS statistics*, 5th ed., Sage, London, 2017.
12. A. Field, *Discovering statistics using IBM SPSS statistics*, 5th ed., Student Resources, Smart Alex's Solutions, Sage, London, 2017. <https://edge.sagepub.com/field5e/student-resources/smart-alex-solutions>; https://milton-the-cat.rocks/dsus_alex.html#Task_206.
13. R. Crompton, *Forty years of sociology: some comments*, Sociology **42** (2008), no. 6, 1218–1227.
14. G. Payne, *Surveys, statisticians and sociology: a history of (a lack of) quantitative methods*, Enhanc. Learn. Soc. Sci. **6** (2014), no. 2, 74–89.
15. E. Masuadi, M. Mohamud, M. Almutairi, A. Alsunaidi, A. K. Alswayed, and O. F. Aldhafeeri, *Trends in the usage of statistical software and their associated study designs in health sciences research: a bibliometric analysis*, Cureus **13** (2021), no. 1, e12639. <https://doi.org/10.7759/cureus.12639>.
16. R. A. Muenchen. *The popularity of data science software*. <https://r4stats.com/articles/popularity/>, 2024.
17. J. Zhang and K. F. Yu, *What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes*, JAMA J. Am. Med. Assoc. **280** (1998), 1690–1691.
18. P. W. Holland, *A note on the covariance of the Mantel-Haenszel log-odds-ratio estimator and the sample marginal rates*, Biometrics **45** (1989), no. 3, 1009–1016.
19. I. Shrier and R. Steele, *Understanding the relationship between risks and odds ratios*, Clin. J. Sport Med. **16** (2006), no. 2, 107–110.
20. A. Agresti, *Categorical data analysis*, 3rd ed., Wiley, Hoboken, NJ, 2013.
21. R. Pearl and L. J. Reed, *A further note on the mathematical theory of population growth*, Proc. Natl. Acad. Sci. USA **8** (1922), no. 12, 365–368.
22. E. Maor, *e: the story of a number*, Princeton University Press, Princeton, NJ, 1994.

How to cite this article: D. Voas and L. Watt, *The odds are it's wrong: Correcting a common mistake in statistics*, Teach. Stat. (2024), 1–12, DOI [10.1111/test.12391](https://doi.org/10.1111/test.12391).

APPENDIX A: GOING FURTHER

$$y = \frac{1}{1 + e^{-x}}.$$

A.1 | UNDERSTANDING LOGARITHMS

It is possible to treat logistic regression as a machine for generating odds ratios, but many people will wish to know how the method works. To do so, one needs to have a basic understanding of logarithms and the exponential function e^x .

The logarithm (or “log” for short) is just the power to which you have to raise the base to obtain a given number. You have to raise 10 to the power 4 to produce 10,000, so $\log_{10}(10,000) = 4$.

Any number can be raised to a higher power (so, for example, $2^3 = 8$), and likewise, any number can be used as the base for logs. In statistics and most other branches of mathematics, we typically use the value e as the base, rather than 10.

But what, students will want to know, is e ? The simplest answer is that like π , it is a fundamental constant that pops up everywhere. Both e and π appear in the formula for the normal distribution, for instance. The value of e is approximately 2.71828. For anyone wishing to know more, a good starting point is the article on e (mathematical constant) in the online Britannica, the section on compound interest in the Wikipedia article headed “ e (mathematical constant),” or the first three chapters of Maor.²²

For practical purposes in understanding logistic regression, it is enough to know that e is a numerical value, just like 2 or 10 or π , and it can be raised to a power or used as the base of a logarithm. Logs to base 10 are called “common logarithms”; logs to base e are “natural logarithms.” We can use the symbol $\log_e(x)$ for logs to the base e , but it is worth knowing that $\ln(x)$ is widely used as the notation for natural logarithms.

Two mathematical facts are particularly important for what follows. The first is that logarithms change division into subtraction: $\log(a/b) = \log(a) - \log(b)$. The second is that exponentiation and logarithms are inverse functions: $e^{\log_e(x)} = x$ and $\log_e(e^x) = x$. Just as multiplication and division can cancel each other out, the same is true of the exponential and log functions.

A.2 | HOW LOGISTIC REGRESSION WORKS

The role of odds ratios in logistic regression can be understood if one knows about logs and can follow elementary algebra. The exposition could proceed along the following lines.

Figure 1 was produced by the simplest logistic model:

That is in fact exactly the equation used in logistic regression. It probably helps to use “ p ” rather than “ y ” because we should think of it as the probability that the dependent variable equals 1. In addition, “ x ” here represents the whole expression we know from linear regression: $a + b_1x_1 + b_2x_2 + \dots$. Thus, the starting point is:

$$p = \frac{1}{1 + e^{-(a+b_1x_1+b_2x_2+\dots)}}.$$

But here is the problem with logistic regression. We have something simple that we are interested in—the probability of the outcome, given the predictors—on the left-hand side of the equation, but the right-hand side is cumbersome. We can fix that problem, but it comes at the cost of turning the left-hand side into something more complicated.

Some basic algebra allows us to turn the equation into:

$$\frac{p}{1-p} = e^{(a+b_1x_1+b_2x_2+\dots)}.$$

So at this point, instead of a probability, we are now working with odds: $p/(1-p)$.

Things are going to get worse. Taking the natural logarithm of both sides of the equation gives us log odds on the left-hand side:

$$\log_e\left(\frac{p}{1-p}\right) = a + b_1x_1 + b_2x_2 + \dots$$

Now we can see the trade-off. We started with an S-shaped (logistic) curve to represent the probability of the outcome. We have transformed it to work just like linear regression in the sense that the right-hand side is a linear expression with independent variables x_i and their coefficients b_i . There is a crucial difference on the left-hand side of the equation, though. Instead of predicting the actual value of the dependent variable, we are predicting the log odds that it is 1.

The good news is that we can still discover whether our independent variables have a significant influence on the outcome and in what direction. The bad news is that interpreting the coefficients is not as easy as it is in linear regression. It is still the case that each coefficient gives us the effect on the outcome of a one-unit change in the

predictor variable, but here the outcome is the log odds rather than the value of the dependent variable—and no one has an intuitive sense of what log odds mean.

To make the situation more concrete, let us assume that x_1 is a dummy variable, such as 0 = male and 1 = female. The probability that someone passes a course is p , and more specifically, the probability that a man will pass is p_0 , while the probability that a woman will pass is p_1 . The coefficient of x_1 (representing gender) is b_1 . A one-unit change in x_1 (that is, the change from male to female) produces a change of b_1 in the log odds of passing the course, and so, the log odds for men plus b_1 gives us the log odds for women:

$$\log_e\left(\frac{p_0}{1-p_0}\right) + b_1 = \log_e\left(\frac{p_1}{1-p_1}\right),$$

which means that b_1 is the log odds for men minus the log odds for women:

$$\log_e\left(\frac{p_1}{1-p_1}\right) - \log_e\left(\frac{p_0}{1-p_0}\right) = b_1.$$

And because $\log(a) - \log(b) = \log(a/b)$ for any values a and b , that is equivalent to:

$$\log_e\left(\frac{\frac{p_1}{1-p_1}}{\frac{p_0}{1-p_0}}\right) = b_1.$$

Exponentiating both sides gives us:

$$\frac{\left(\frac{p_1}{1-p_1}\right)}{\left(\frac{p_0}{1-p_0}\right)} = e^{b_1}.$$

Thus, the exponentiated coefficient is the odds ratio. Again, there is good news and bad news. The good news is that the crucial information is easy to spot: a value below 1 means that women are less likely than men to pass, while a value greater than 1 means that women are more likely than men to pass. The bad news is that the actual value of the odds ratio is of limited help because it means very little on its own. We have to convert the odds into probabilities.