Logistic Regression

Kristin L. Sainani, PhD

Logistic regression is widely used in the medical literature for analyzing binary outcome data. Logistic regression has many similarities to linear regression, but it is more complex and harder to evaluate graphically. As a result, many researchers apply logistic regression without a deep understanding of the model and without sufficient plotting. This article provides a visual and mathematical introduction to logistic regression, along with tips for evaluating articles that contain logistic regression.

WHAT IS LOGISTIC REGRESSION?

Logistic regression is similar to linear regression; it attempts to fit a line (an intercept and slope) to the data. However, because the outcome has only two levels, it is not optimal to fit a line directly to such data. Figure 1 illustrates this point using hypothetical data (n = 100). The outcome is whether a student passed a particular statistics class; the predictor is the number of hours the student spent on homework each week. I have superimposed a linear regression line on the plot to demonstrate that a line is not a good fit.

Rather than fitting a line directly to the binary outcome (eg, pass/no pass), logistic regression instead uses a transformation of the outcome called a logit, or log odds. Figure 2 illustrates that a line fits the data better after this conversion.

The logit is directly related to the probability of the outcome. Unlike probability, however, it can take on any value from negative to positive infinity. The relationship between the logit and the probability is:

$$\text{logit} = \ln\left(\frac{p}{1-p}\right)$$

For example, if the probability of passing the course is 37.5%, then the logit of passing is:

$$\text{logit} = \ln\left(\frac{.375}{.625}\right) = -0.51$$

To calculate and plot logits (as in Figure 2), one must divide the data into groups. For example, I divided the students into deciles by homework times. In the lowest decile of homework time ($\leq$7 hours per week), 3 out of 8 students, or 37.5%, passed; this corresponds to a logit of −0.51. However, the model-fitting algorithm for logistic regression does not involve any arbitrary division into groups. The algorithm uses calculus to find the equation of the best-fit line. When I fit a logistic regression model to these data, the resulting equation is:

$$\text{logit}\ (\text{passing the class}) = -2.2 + 0.24 * \text{homework hours}$$

The Y intercept is −2.2, and the slope is 0.24. The Y intercept indicates that the logit of passing is −2.2 for a student who does no homework. The slope indicates that for every 1 additional hour of homework, the logit of passing goes up 0.24.

K.L.S. Division of Epidemiology, Department of Health Research and Policy, Stanford University, Stanford, CA. Address correspondence to: K.L.S.; e-mail: kcobb@stanford.edu
Disclosure: nothing to disclose

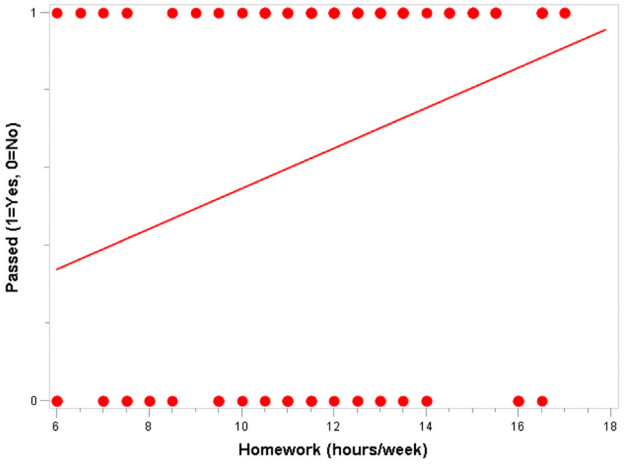

Figure 1. A scatter plot between a binary outcome (whether a student passed the class) and a continuous predictor (homework hours per week). A line does not fit these data well.

Most of us don't think in logits. Fortunately, the slopes or "beta coefficients" from logistic regression easily can be translated into odds ratios.

WHAT DO THE RESULTS MEAN?

Logistic regression yields information about the relationship between individual risk/protective factors and the outcome. It also can be used to calculate predicted probabilities and to generate receiver operating characteristic (ROC) curves, which reflect the discriminatory ability of the entire model.

INDIVIDUAL BETA COEFFICIENTS

The slopes (beta coefficients) from a logistic regression can be interpreted as odds ratios, a measure of relative risk. The mathematical details of this process are explained in the In-Depth Box.

Using the same example dataset, I fit a logistic regression model with 2 predictors: homework time and gender, coded as 1 for female and 0 for male. The resulting equation is:

$$\text{logit (passing the class)} = -2.4 + 0.21 * \text{homework hours} + 1.43 * (1 \text{ if female, } 0 \text{ if male})$$

If I exponentiate the slope for gender, I get the adjusted odds ratio for women versus men:

$$\text{OR}_{\text{women vs men}} = \exp^{1.43} = 4.18$$

Thus, independent of homework times, women have a 4-fold higher *odds* of passing the class than do men. The odds ratio for a continuous predictor is calculated in the same way but is interpreted as the increase in odds for every 1-unit increase in the predictor. The odds ratio for homework is:

$$\text{OR}_{\text{homework}} = \exp^{0.21} = 1.23$$

This means that for every additional hour of homework per week, a student's *odds* of passing go up by 23%. The odds ratio for homework would be different if the authors chose different units for this variable (eg, homework per day). Thus, authors need to specify the units, as in Table 1.

If the logistic regression model contains interaction terms, the calculation of odds ratios is more complex. For example, if we found a significant interaction between gender and homework time, this would mean that there were 2 odds ratios for homework time—one for males and one for females.

PREDICTED PROBABILITIES

The logistic regression model can be used to estimate the probability that a person will have the outcome. For

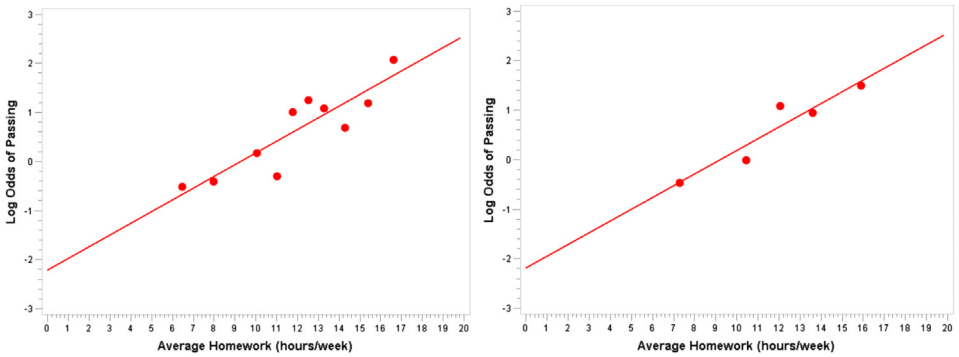

Figure 2. Scatter plots of the logit (log odds) of passing the course versus homework times, with a best-fit line superimposed. I divided my observations into deciles (left panel) or quintiles (right panel) by homework times. The equation of the best-fit line is −2.2 + 0.24 * homework hours. Note how both lines intersect the Y axis at −2.2.

Table 1. *Odds ratios and 95% confidence intervals from logistic regression*

Variable	Odds Ratio	95% Confidence Interval
Homework time (per hour/week)	1.23	1.04-1.45
Female gender (vs male)	4.18	1.60-10.96

example, the logistic regression model for passing the class is:

$$\ln\left(\frac{p}{1-p}\right) = -2.2 + 0.21 * \left(\text{homework hour}\right) + 1.43 * \left(1 \text{ if female, } 0 \text{ if male}\right)$$

In this equation, p represents the predicted probability of passing. We can algebraically rearrange the equation to isolate p:

$$p = \frac{e^{-2.2 + 0.21 * \text{homework} + 1.43 * \text{gender}}}{1 + e^{-2.2 + 0.21 * \text{homework} + 1.43 * \text{gender}}}$$

Thus, the predicted probability of passing for a woman who does 10 hours of homework per week is:

$$p = \frac{e^{-2.2 + 0.21 * (10) + 1.43 * (1)}}{1 + e^{-2.2 + 0.21 * (10) + 1.43 * (1)}} = \frac{3.1}{4.1} = 76\%$$

The predicted probability of passing for a man who does 10 hours of homework per week is:

$$p = \frac{e^{-2.2 + 0.21 * (10) + 1.43 * (0)}}{1 + e^{-2.2 + 0.21 * (10) + 1.43 * (0)}} = \frac{.74}{1.74} = 43\%$$

Table 2 shows the predicted probabilities for a limited number of observations from our hypothetical dataset. The difference between the observed outcome and the predicted probability is called the residual. For example, subject 1 has a value of 0 for passing and a predicted probability of 0.24, so his residual is 0-0.24 = −0.24. Residuals plots are useful for evaluating the model.

Table 2. *The 10 observations with the lowest predicted probabilities in our hypothetical dataset*

ID	Homework Hours Per Week	Gender (Female = 1, Male = 0)	Predicted Probability of Passing	Passed? (Yes = 1, 0 = No)
1	6.0	0	0.24	0
20	6.0	0	0.24	1
31	6.0	0	0.24	0
44	6.0	0	0.24	0
12	6.5	0	0.26	1
6	7.0	0	0.28	1
7	7.0	0	0.28	0
88	7.0	0	0.28	0
69	7.5	0	0.30	0
17	8.0	0	0.33	0

Note that if the data derive from a case-control study, the predicted probabilities are not valid because they reflect the ratio of cases to control subjects in the study, rather than the true probability of developing the disease.

ROC CURVES

Logistic regression also yields ROC curves, which graphically show how well a particular logistic model discriminates between 2 groups (eg, cases and non-cases). ROC curves are useful when the goal of the study is to classify people as accurately as possible, as in a diagnostic study.

ROC curves are plotted on a 1 × 1 square. If the ROC curve lies along the upper left-hand corner of the square, such that the area under the curve is 100%, then the model has perfect discriminatory power. If the ROC curve lies on the diagonal line, such that the area below is 50%, then the model discriminates no better than chance. Figure 3 shows the ROC curve for the model containing homework and gender; this model has moderate discriminatory power, with an area under the curve of 76%.

How exactly is the ROC curve drawn? The points correspond to the sensitivity and specificity of different predicted probabilities. For example, in our hypothetical

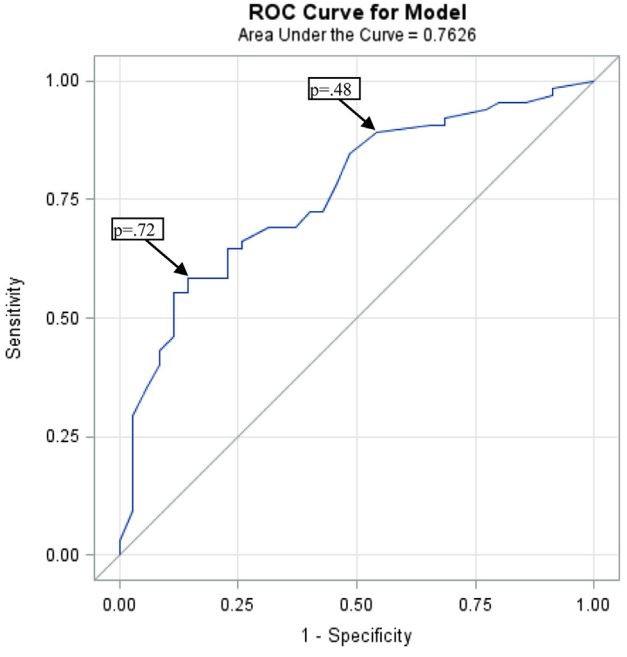

Figure 3. Receiver operating curve (ROC) for the logistic regression model that contains homework times and gender. The arrows indicate the points corresponding to predicted probability cutoffs of 48% and 72%.

dataset, 38 of 65 people who passed the course had predicted probabilities of 72% or higher, and 30 of 35 people who failed the course had predicted probabilities below 72%. Thus, a cutoff of 72% yields a sensitivity of 38/65 = 58.5% and a specificity of 30/35 = 85.7%. Because ROC curves plot sensitivity against 1-specificity, this corresponds to the point x = 0.143, y = 0.585, which is highlighted in Figure 3.

As a second example, consider the predicted probability of 48%. Of the 65 students who passed, 58 had predicted probabilities of 48% or higher, which corresponds to a sensitivity of 89.2%. Of the 35 who failed, 16 had predicted probabilities below 48%, for a specificity of 45.7% and 1-specficity of 54.3%. The point x = 0.543, y = 0.892 is highlighted in Figure 3. The remaining points on the ROC curve are calculated in this same manner.

The area under the ROC curve is equivalent to the C statistic. The C statistic gives the likelihood that a randomly selected case (a person who had the outcome) from the dataset will have a higher predicted probability than a randomly selected non-case.

WHAT ARE THE COMMON PITFALLS?

Logistic regression is often applied carelessly. Readers should check whether authors adequately plotted their data, interpreted odds ratios correctly, and recognized the limitations of ROC curves.

PLOTTING IS CRITICAL

Just as it is important to plot the data when performing linear regression, the same is true for logistic regression. Logit plots can reveal when the relationship between a continuous predictor and the outcome is nonlinear. To illustrate this capability, I created a second hypothetical dataset in which homework has a quadratic relationship with passing (Figure 4). The optimal homework time is a medium amount—students who do too much or too little are less likely to pass. Thus, one should include a homework-squared term in the model or should treat homework as categorical (eg, low, medium, and high). Without plotting the data, researchers would likely miss this effect.

Unfortunately, many authors skip plotting for logistic regression because logit plots are more difficult to generate than simple scatter plots. However, residual plots—which are automatically generated by most standard statistical packages—can also reveal nonlinear relationships. The trick is to superimpose a curved line (or "smoothing line") on the residual plot, as in Figure 5. Residual plots are also useful for identifying outliers.

There is no need to test the linearity assumption for binary or categorical predictors. Binary predictors have only 2 levels. Because 2 points always define a line, the relationship with the outcome is always linear, as Figure 6 shows. The same applies for categorical variables, which are entered into the model as a series of binary variables (called "dummy coding").

ODDS RATIOS CAN BE MISLEADING

In interpreting results for continuous predictors, readers should keep in mind that the magnitude of the odds ratio depends on the units chosen. If the authors choose smaller units, the odds ratio will appear smaller; if the authors choose larger units, the odds ratio will appear bigger. Readers should also check that authors have calculated odds ratios correctly when the model contains interaction terms. Finally, readers should be aware that odds ratios can distort effects [1]. In our hypothetical example, women have a 4-fold higher *odds* of passing the course compared with men, but it would be incorrect to say that they have a 4-fold higher *chance* of passing.

Researchers should consider alternatives to logistic regression when the odds ratios are too misleading. For example, binary data can be analyzed using Poisson

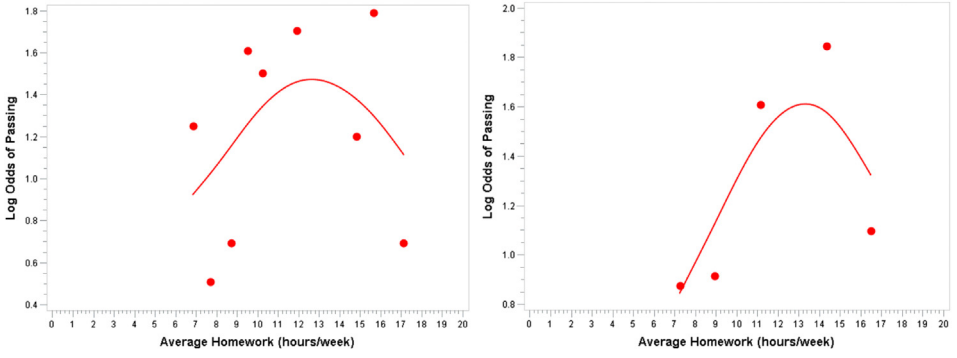

Figure 4. Logit plots for hypothetical dataset 2. The logit plots reveal that a quadratic rather than a linear relationship exists between homework times and the outcome. I created logit plots using deciles (left panel) and quintiles (right panel) of homework times. Researchers should try several different groupings when making logit plots.

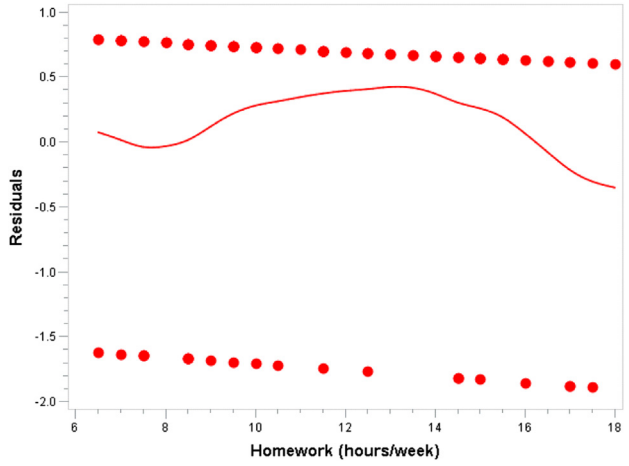

Figure 5. Residual plots can reveal nonlinear relationships between a continuous predictor and the logit of the outcome. I ran a logistic regression on hypothetical dataset 2 with homework time as the predictor and the logit of passing as the outcome. I plotted the resulting residuals (one per observation) against homework time and fit a smoothing line to these data. The curved pattern indicates that homework has a quadratic relationship with the logit of passing.

regression with robust standard errors to generate adjusted risk ratios, rather than odds ratios [2]. When I applied this method to the example data, I got risk ratios of 1.07 (95% confidence interval: 1.01-1.13) for homework and 1.54 (95% confidence interval: 1.14-2.10) for gender. Thus, although a woman's *odds* of passing are 4-fold higher than a man's odds of passing, her *chance* of passing is only increased 54%.

Odds ratio are most likely to exaggerate effects when the outcome is common and the effect sizes are modest to strong. Thus, readers should check that authors have interpreted odds ratios cautiously when the risk or prevalence of the outcome exceeds 10% in the reference group, particularly when the odds ratios exceed 2.0 (or, equivalently, are lower than 0.50).

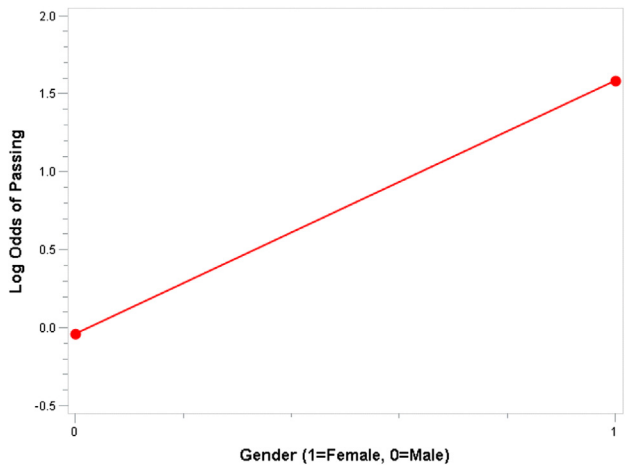

Figure 6. The relationship between a binary predictor (such as gender) and the logit of a binary outcome (such as passing the course) is always linear because 2 points define a line.

ROC CURVES HAVE LIMITATIONS

ROC curves are widely used in medical literature for the evaluation of logistic regression models. They do have some important limitations, however. First, the C statistic is always overly optimistic when it is calculated using the same observations that were used to fit the model (as above). If I were to apply the logistic regression model for homework times and gender to a new dataset, the C statistic will go down at least somewhat, which is why it is important to internally or externally validate the model.

Second, a high C statistic means that the model discriminates well, but it does not mean that the predicted probabilities are accurate. For example, a model would discriminate well if it assigns predicted probabilities of 15% to all students who fail the course and 16% to all students who pass, even if the absolute probability of passing is much higher or lower in either group.

Third, many researchers use the change in the C statistic to judge whether adding a risk factor to a model improves model performance. However, the change in the C statistic is not a sensitive measure. Adding a risk factor to a logistic regression model may improve model performance without significantly changing the C statistic [3].

Finally, the C statistic gives equal weight to sensitivity and specificity. However, in a clinical setting, it may be more important to avoid false-positive results than to avoid false-negative results, or vice versa.

CONCLUSION

Though logistic regression is widely used in the medical literature, many authors and readers are unaware of its assumptions and limitations. When considering an article with logistic regression, readers should look for evidence that the authors have graphically assessed their model, particularly if it includes continuous predictors. Readers and authors should be cautious in the interpretation of odds ratios from logistic regression and should consider alternative regression models that yield adjusted risk ratios rather than odds ratios. When the goal of modeling is predictive, the area under the ROC curve (or C statistic) is a useful measure, but readers and authors should keep in mind its limitations.

REFERENCES

1. Sainani KL. Understanding odds ratios. PM R 2011;3:263-267.
2. Zou GA. Modified Poisson regression approach to prospective studies with binary data. Am J Epidemiol 2004;59:702-706.
3. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation 2007;115:928-935.

In-Depth: From Beta Coefficients to Odds Ratios

The logistic regression equation gives us the predicted logit of the outcome for a particular type of person. For example, the predicted logit for a woman who does 10 hours of homework is:

$$\ln\left(\frac{p}{1-p}\right) = -2.4 + .21 * \left(10\right) + 1.43 * \left(1\right)$$

To convert this value to a predicted odds of passing, we exponentiate to remove the natural log:

$$\text{Odds} = \exp^{\ln\left(\frac{p}{1-p}\right)} = \exp^{-2.4 + .21 * (10) + 1.43 * (1)}$$

Similarly, the odds of passing for a man who does 10 hours per week of homework is:

$$\text{Odds} = \exp^{\ln\left(\frac{p}{1-p}\right)} = \exp^{-2.4 + .21 * (10) + 1.43 * (0)}$$

To calculate the odds ratio for women versus men, we divide the odds for women by the odds for men:

$$OR = \frac{\text{odds for women}}{\text{odds for men}} = \frac{\exp^{-2.4 + 0.21 * (10) + 1.43 * (1)}}{\exp^{-2.4 + 0.21 * (10) + 1.43 * (0)}}$$

Note that the intercepts cancel. The homework terms also cancel as long as we compare men and women with equal homework times. Thus, the odds ratio for gender is just the beta coefficient for gender exponentiated:

$$\frac{\exp^{-2.4 + 0.21 * (10) + 1.43 * (1)}}{\exp^{-2.4 + 0.21 * (10) + 1.43 * (0)}} = \frac{\exp^{1.43 * (1)}}{\exp^{1.43 * (0)}} = \exp^{1.43} = 4.18$$

We say that this is an "adjusted" odds ratio for gender, meaning that this represents the effect of gender, holding homework time fixed. Note that if there are interaction terms in the model, the calculation of odds ratios becomes more complex.