Overview of regression with binary outcomes

As we have seen through our study of linear regression, a regression model usually focuses on the conditional mean function $E[y|x_1, \ldots, x_p]$, which can be interpreted as the average value of the dependent variable (y) for fixed values of the predictor variables $x_1, \ldots, x_p$. In the setting of a linear model, the mean function is expressed as the linear predictor $b_0 + b_1x_1 + \ldots + b_px_p$. This linear approach is useful for many different types of dependent variables, especially quantitative variables like blood pressure which we have used as an illustrative example previously in this course.

It also frequently arises that we would like to conduct a regression analysis using a binary dependent variable, i.e. a dependent variable that can only take on two values. For illustration, we can consider a research study of a population, say the US adult population, and look at whether a given person in this population has ever been a smoker in their lifetime. This is a categorical variable with two levels, so it does not have an expected value (i.e. if we code the variable's values as "smoker" and "non-smoker" you cannot take the mean of these two labels). However we can choose one category, say the "smoker" category, and aim to model the probability that this outcome occurs. Equivalently, we can code the outcome numerically as 1 (for "smoker") and 0 (for "non-smoker") and then model the expected value of this binary quantitative variable. These two approaches turn out to be equivalent.

A binary dependent variable such as smoking status generally cannot be modeled with linear regression. The probability (or expected value if coded numerically) of a binary variable can only take on values between 0 and 1. There would be no practical way to constrain the linear predictor $b_0+b_1x_1+\ldots+b_px_p$ to remain within this interval for all possible values of the independent variables $x_1, \ldots, x_p$. Thus, if we tried to model a binary dependent variable with linear regression, we would often obtain fitted probabilities that are greater than 1 or less than 0. This is one of several ways in which linear regression is not ideal for modeling binary dependent variables.

When working with binary dependent variables, it is much more common to introduce a link function that is used to map the probability (or mean) to a value on the real number line, which can then be modeled using the linear predictor. The link function that is most commonly used for regression with a binary outcome is the logistic function, or log odds function, $\log(p/(1-p))$.

This gives rise to a very widely-used approach for regression analysis with binary outcome variables called logistic regression.

Odds, log odds, and the link function

We first review some definitions related to binary data values. If we have a binary random variable y that takes on two values (0 and 1), then we can write p as the probability that y is equal to 1, and it follows that $1-p$ is the probability that y is equal to 0. These probabilities must lie in the interval $[0, 1]$. The odds is a quantity that is derived from a probability, it is given by the expression $p/(1-p)$. The odds can take on any non-negative number. For example, if the odds is 3, this means that we are three time more likely to observe a 1 than to observe a 0. An odds of 1/3 means that we are 3 times more likely to observe a 0 than to observe a 1. An odds of 1 is "neutral" in the sense that in this case we are equally likely to observe either outcome. The neutral odds value of 1 corresponds to the neutral probability value of 1/2.

A further transformation that is often applied is to take the (natural) logarithm of the odds, yielding the log odds, $\log(p/(1-p))$. The neutral point for the log odds is 0 – that is, a log odds of 0 corresponds to an odds of 1 and to a probability of 1/2. The log odds is symmetric around 0, in the sense that a log odds of v and a log odds of $-v$ convey the same strength of association in terms of the bias relative to the neutral value. This symmetry makes the log odds a good choice for modeling via the linear predictor. The log odds transformation is identical to the logistic function, and it is the latter term that gives logistic regression its name.

Logistic regression

In a logistic regression, we model the log odds as being equal to the linear predictor. That is, we set $\log(p/(1-p)) = b_0 + b_1 x_1 + \ldots + b_p x_p$. For example, if the outcome we are modeling is smoking history, and the covariates are age and gender (coded as female=1 if the reference level of this categorical variable is "male"), then the model is $\log(p/(1-p)) = b_0 + b_1 \cdot \text{age} + b2 \cdot \text{female}$. If b_1 is positive, then older people are more likely to be smokers than younger people (comparing within a subpopulation of people having a fixed gender), whereas if b_1 is negative, older people are less likely to be smokers than younger people. If $b_1 = 0$, then age is unrelated to the probability that a person smokes (within a specific gender). Similarly, if b_2 is positive, then females are more likely to smoke than males, and if b_2 is negative, then

females are less likely to smoke than males.

Parameter interpretation

The quantitative interpretations of the logistic regression parameters are best expressed either additively in terms of the log odds, or multiplicatively in terms of the odds. If the age parameter is 0.04, this means that comparing two people of the same gender whose ages differ by one year, the older person has log odds for smoking that is 0.04 units greater than that of the younger person. This relationship is linear in age, so if we compare two people of the same gender whose ages differ by ten years, then the older person has 0.4 greater log odds for smoking than the younger person.

Since many people may be unfamiliar with the log odds scale, we can also express these relationships in terms of odds, by exponentiating the log odds differences. We must remember however that after exponentiating, the effects for the odds are multiplicative, while the effects for the log odds were additive. For example, since $\exp(0.04) \approx 1.04$, we see that when comparing two people of the same gender whose ages differ by one year, the older person has 1.04 times greater odds of smoking than the younger person. When comparing people with the same gender whose ages differ by ten years, the older person has $\exp(0.4) \approx 1.49$ times greater odds of smoking than the younger person.

Inference and uncertainty assessment

Uncertainty assessment for logistic regression can be conducted using many of the same tools that are used when working with linear models. The parameters $b_0, \ldots, b_p$ will be reported by any software as point estimates with standard errors. We can therefore construct a confidence interval for each parameter, and we can carry out a hypothesis test for the null hypothesis that a given parameter is equal to zero.

Our ability to accurately recover the parameters in a logistic regression analysis is limited by many of the same factors that limit parameter estimation in linear models. Recall that in the case of linear models, four critical factors influencing our ability to estimate a regression parameter are: (1) sample size, (2) unexplained variation (conditional variance), (3) variance of the predictors, and (4) correlations among predictors. In a logistic regression, sample size (1), variance of the predictors (3), and correlations among predictors (4) contribute to uncertainty in a similar manner to how they contribute in linear regression. But in a logistic regression, the conditional variance (2) is completely determined by the mean, and plays a very different role compared

to the setting of linear regression.

Also, in the case of linear regression, the true values of the regression parameters are not relevant for determining the uncertainty in the estimated value of the regression parameters. That is, we can estimate a strong effect and a weak effect with equal precision. However this is not the case in logistic regression. In general, in a logistic regression, the standard errors for estimating parameters that are larger in magnitude are greater than the standard errors for estimating parameters that are smaller in magnitude.

Visualization and diagnostics
The most effective visualization technique for a fitted logistic regression model may be the CERES plot. This is a modified version of an added variable plot that aims to show the role of one variable in a logistic regression model without pre-supposing that its role is linear. After inspecting a CERES plot, we may choose to introduce quadratic terms or interactions to improve the fit. More advanced methods such as basis splines (not covered here), are also used for this purpose.