LOGISTIC REGRESSION

ETHAN LEVIEN

Contents

1. Motivation 1
2. Logistic model 2
2.1. The logistic function 2
2.2. Fitting logistic regression 2
3. Interpreting coefficients 2
4. Model evaluation for logistic regression 3
5. Logistic regression vs. binning 3

1. Motivation

So far we have considered the linear model:

$$Y = b + \sum a_i X_i + \epsilon \tag{1}$$

where ϵ follows is mean zero Normal random variable. We understand that this can be formulated as the conditional distribution of Y given X:

$$Y|X \sim \text{Normal}\left(b + \sum_i a_i, \sigma_\epsilon\right) \tag{2}$$

We know that we have some formulates for unbiased estimators of b, a_i and σ_ϵ in terms of our data. Specifically, the formulas for the a_i come from the covariances and variances of the predictors and response variables in complicated ways (but are straightforward to compute). After recognizing that we can select both Y and X_i to be function of various variables in our data (by transformations and addition of features), this modeling framework gives the ability to expand our model indefinitely. The remaining limitation is that of the structure of the noise: Since the noise is Gaussian, we are limited to studying only certain times of randomness.

What if we want to predict a binary variable? As an example, let's consider the problem of predicting whether someone supports same sex marriage based on some information about them, such as age and gender. As a "warm up" to get us thinking about this problem, will start with the binary predictor sex (which is restricted to Male or Female is this dataset). Our response variable Y is one if someone supports same sex marriage and zero otherwise. Thus Y is a Bernoulli random variable and thus does not follow a normal distribution regardless of which predictors we condition on. In this case, we can frame the problem of modeling the association between X and Y as two separate inferences of Bernoulli random variables:

$$Y|(X = 0) \sim \text{Bernoulli}(q_0) \tag{3}$$

$$Y|(X = 1) \sim \text{Bernoulli}(q_1) \tag{4}$$

Hence, we can simply break the data up into two groups and estimate q_0 and q_1 as we've done before with Bernoulli random variables.

Alternatively, we can frame this as a regression problem

$$Y|X \sim \text{Bernoulli}(q(X)) \tag{5}$$

where

$$q(X) = q_0(1 - X) + q_1 X = X(q_1 - q_0) + q_0 \tag{6}$$

Structurally, this is similar to the linear regression. Our model for the response variable Y conditioned on X is a distribution in which the parameters depend linearly on X. In the linear regression context, it is the mean of the Normal distribution that depends linearly on X while here it is the chance for $Y = 1$. In fact, if in our data k people support and $N - k$ do not, we compute $\tilde{Y} = k/N$. We could

Date: April 2022.

then have a linear regression model (with Normal noise) for $\tilde{Y}$ in terms of X. We will later see that there are some advantages to working directly with Y!

More generally, in order to capture binary noise, we might start with the model

$$Y|X \sim \text{Bernoulli}(q). \tag{7}$$

Here, q is the sole parameter and therefore in order for the distribution of Y to depend on X, q must depend on X. Naively, we might simply take $q(X)$ to be

$$\sum a_i X_i \tag{8}$$

but this has a problem if X are continuous predictors, since we must have $0 < q < 1$. In order to ensure this is the case, we set

$$W = \sum_i a_i X_i \tag{9}$$

and set $q = h(W)$, but we want to select h so that as $W \to \infty$, $q \to 1$, and as $W \to -\infty$, $q \to 0$. The next task is to see how this is done.

2. Logistic model

2.1. The logistic function.

We would ideally like to come up with a function $h = h(w)$ that maps w to 0 and 1. The standard choice is

$$q(w) = \text{logit}^{-1}(w) = \frac{1}{1+e^{-w}} \tag{10}$$

This is called the inverse logistic function between if we solve for z, we get the **logistic function**

$$w = \ln\left(\frac{q}{1-q}\right) \tag{11}$$

To better understand how the slope and intercept b and a effect the plots, let's think about limiting cases. But first, think about the functions e^{-w} and $\text{logit}^{-1}(w)$.

- $e^{-w} = 1$ when $w = 0$. Thus $\text{logit}^{-1}(0) = \frac{1}{1+1} = \frac{1}{2}$.
- If w is a very large positive number, $e^{-w} \approx 0$, so $\text{logit}^{-1}(w) \approx \frac{1}{1+0} = 1$.
- If w is a very large negative number e^{-w} is huge, so $\text{logit}^{-1}(w) \approx \frac{1}{1+\infty} = 0$.

Let's imagine $b = 0$ (there is no intercept).

- If a is very large relative to all values of x, then $w = ax$ will quickly become large for small x and therefore y will very likely be 1 for positive x (and very likely be zero for negative x)
- If a is small relative to all values of x, then $w = ax$ will not change much and the chance that $y = 1$ will be around $1/2$ for most x.

To summarize, our the logistic regression model is

$$Y|X \sim \text{Bernoulli}\left(\frac{1}{1+e^{-b-\sum_i a_i X_i}}\right) \tag{12}$$

Example 1. *Generating data from logistic regression model*

2.2. Fitting logistic regression.

We can fit this in statsmodels as shown by the following example, and much of what we've learned turns out to carry over.

Example 2. *Logistic regression in statsmodels*

Example 3. *Logistic regression with real data*

3. Interpreting coefficients

In a logistic regression the meaning of the coefficients is a bit tricky. **This is because their "effect" depends on the value of the predictors.** Let's think about a model with multiple predictors.

Let's think about the intercept first. When the predictor (or all predictors if there are multiple) is zero,

$$P(Y = 1|X = 0) = \frac{1}{1+e^{-b}} \implies b = -\ln\left(\frac{1}{q} - 1\right) \tag{13}$$

We can rearrange terms to get

$$b = \ln\left(\frac{q}{1-q}\right) = \ln\frac{P(Y=1|X=0)}{P(Y=0|X=0)} \tag{14}$$

the expression $(1-q)/q$ is called the odds ratio, so b tells us the log odds ratio to get $y = 1$ when all predictors are zero.

More generally, a_i **tells us how much the log odds ratio changes when we chance** X_i **by** 1 **with all other predictors fixed.** To see this, first note that if the odds are $q = 1/(1+e^{-w})$, the odds ratio is

$$\ln \frac{1/(1+e^{-w})}{1-1/(1+e^{-w})} = \ln \frac{1+e^{-w}}{e^{-w}(1+e^{-w})} = \ln e^{w} = w \tag{15}$$

If we change X_i by 1, then w changes by a_i.

I find this really hard to think about odds ratios. Instead, I think it is easiest to interpret the coefficients when the logistic function is well approximated by a linear function. This happens when

$$w = b + \sum_i a_i X_i = 0. \tag{16}$$

At $w = 0$, $\text{logit}^{-1}(0) = 1/2$ and close to $z = 0$ (between -1 and 1)

$$\text{logit}^{-1}(w) = \frac{1}{1+e^{-w}} \approx \frac{1}{2} + \frac{w}{4} \tag{17}$$

This leads to the **divide by four rule**: When the sum of predictors time coefficients is close to 0, $a_i/4$ represents the difference in the chance that $Y = 1$ between data points for which X_i differs by 1, with all other predictors fixed.

The divide by four rule gives an **upper bound** on how much changing X_i changes the probability for Y to be one. In other words, the actual difference is always less than this.

Example 4. *Homocide victim data*

Exercise 1: *Testing divide by four rule*

4. Model evaluation for logistic regression

It's always useful to have some metric for accessing how much of the variation in the data the model explains the data we used to fit it, even though we know this is not the full story. For linear regression we use R^2. For logistic regression we have something called Pseudo R^2. Like R^2, it tells us, roughly speaking, what fraction of the variation in Y values is explained by the predictors, but in this case we don't have the usual notion of residuals. Instead, we compare how likely it is to see the particular sequence of Y values under the model, vs. how likely it would be to see them if the chance to get $Y = 1$ did not depend on X:

$$\text{pseudo } R^2 = 1 - \frac{\ln(\text{chance to see } Y_1, Y_2, \ldots, Y_n \text{ given our model})}{\ln(\text{chance to see } Y_1, Y_2, \ldots, Y_n \text{ given } x \text{ has no effect})} \tag{18}$$

Why does this make sense? Observe the following facts:

- The chance to see a given sequence of y values of y is between 0 and 1.
- The log of a value between 0 and 1 will be negative, and will be a larger negative number when the chance is smaller.

Thus, if we are much more likely to see the data given our model, the denominator will be closer to 0. If we are less likely to see our data, it will be a large negative number (but always smaller than the denominator).

It tells us how good our model is at predicting the Y **values**. Notice that if we do bin the data and perform a linear regression, the R^2 we get is generally MUCH larger than the Pseudo R^2. Think about why.

Example 5. *Understanding Pseudo* R^2 *with simulations*