# Notes on Logistic Regression

Paul F. Roysdon, Ph.D.

## I. Intuition of Logistic Regression

Logistic regression solves the classification problem by reformulating it as a regression problem, explicitly estimating probabilities. To formulate our toy example, we require a review of Bayes' Theorem, and we will assume linear model. Because we don't know the likelihood for the data, given our hypothesis, nor the prior probability for our hypothesis, we assume a linear relationship between the log-odds of our hypothesis and our data. The result is a problem transformed into a function that we can compute.

### A. *Example Set-up*

Our example seeks to model, probabilistically, the perfect cup of coffee. Brewing a great cup of coffee depends on several factors:

- Age of the beans.
- Coarseness of the grind.
- Weight of the grounds.
- Duration of the pour.
- Temperature of the water.
- Quantity of water.

While there might be other factors, we use the above as our *set-up*, $\mathcal{S}$, otherwise known as our *data*.

### B. *Linear Model*

The linear model $y$ is equal to a linear combination of $x$ multiplied by a parameter $\beta_1$ plus some constant $\beta_0$:

$$y = \beta_1 x + \beta_0.$$

Therefore, $y$ increases or decreases at a rate $\beta_1$, with an intercept $\beta_0$.

### C. *Logistic Function*

Logistic regression uses data to predict the probability of an hypothesis. Given a random input $x$, and some outcome $y$, the logistic regression seeks to *learn* the probability of $y$ given $x$:

$$\mathrm{p}(y|x).$$

Thus a perfectly *trained* model will output $1$ for success, and $0$ for failure. The *inverse logit*, i.e. *logistic function*, for a linear equation is

$$\mathrm{p}(y|x) = \frac{1}{1 + e^{-(\beta_1 x + \beta_0)}},$$

and shown in Fig. 1. Intuitively, this function "pushes" large positive input values toward $1$, and large negative input values toward $0$. The next few sections *intuitively* derive the function $\mathrm{p}(y|x)$.

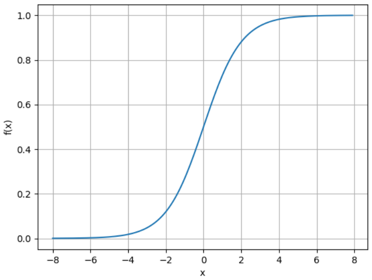

Fig. 1. Logistic Function

### D. *Bayes Theorem*

We seek the probability of our null-hypothesis $\mathcal{H}_0$, a "great cup of coffee", given the correct brewing set-up, $\mathcal{S}$, defined as $\mathrm{p}(\mathcal{H}_0|\mathcal{S})$. We use Bayes Theorem to find the probability of a hypothesis given our set-up:

$$\mathrm{p}(\mathcal{H}_0|\mathcal{S}) = \frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)\mathrm{p}(\mathcal{H}_0)}{\mathrm{p}(\mathcal{S})},$$

where

- $\mathrm{p}(\mathcal{H}_0|\mathcal{S})$ is the *posterior* probability that we make a great cup of coffee given our setup.
- $\mathrm{p}(\mathcal{S}|\mathcal{H}_0)$ is the *likelihood* in having the correct set-up.
- $\mathrm{p}(\mathcal{H}_0)$ is the *prior* probability in making a great cup of coffee.
- $\mathrm{p}(\mathcal{S})$ normalizes the problem such that the output is scaled between 0 and 1.

The intuition of $\mathrm{p}(\mathcal{H}_0)$ is, based on prior experience, that $n\%$ of our cups of coffee are great. The intuition of likelihood is "Given I had a good cup of coffee, how likely is it I had this set-up". The intuition of $\mathrm{p}(\mathcal{S})$, the probability of "this is the correct set-up for brewing great coffee", is discussed next.

### E. *Posterior Odds*

To understand, and evaluate $\mathrm{p}(\mathcal{S})$, we need to consider the alternate-hypothesis, $\mathcal{H}_i$, "the coffee is not great". Note, in the context of statistics, it is not correct to say "the coffee is bad," because our null-hypothesis is "great" therefore the alternate-hypothesis is "not-great." While there are $i$-many alternate-hypotheses, we will only consider one. If we consider the ratio of null-hypothesis to alternate-hypothesis

$$\begin{aligned}\frac{\mathrm{p}(\mathcal{H}_0|\mathcal{S})}{\mathrm{p}(\mathcal{H}_i|\mathcal{S})} &= \frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)\mathrm{p}(\mathcal{H}_0)\frac{1}{\mathrm{p}(\mathcal{S})}}{\mathrm{p}(\mathcal{S}|\mathcal{H}_i)\mathrm{p}(\mathcal{H}_i)\frac{1}{\mathrm{p}(\mathcal{S})}} \\ &= \frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)\mathrm{p}(\mathcal{H}_0)}{\mathrm{p}(\mathcal{S}|\mathcal{H}_i)\mathrm{p}(\mathcal{H}_i)},\end{aligned}$$

notice the $\mathrm{p}(\mathcal{S})$ term drops out. The result is an equation for computing the *posterior odds* for $\mathcal{H}_0$.

Odds express our uncertainty in terms of how many times more likely $\mathrm{p}(\mathcal{S}|\mathcal{H}_0)$ is than $\mathrm{p}(\mathcal{S}|\mathcal{H}_i)$. Because probabilities must add to one (first law of probability), then

$$\mathrm{p}(\mathcal{S}|\mathcal{H}_0) + \mathrm{p}(\mathcal{S}|\mathcal{H}_i) = 1.$$

To simplify notation, let the prior odds of our hypotheses be denoted as

$$\mathrm{o}(\mathcal{H}_0) = \frac{\mathrm{p}(\mathcal{H}_0)}{\mathrm{p}(\mathcal{H}_i)},$$

then the likelihood ratio in terms of odds is

$$\mathrm{o}(\mathcal{H}_0|\mathcal{S}) = \frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)}{\mathrm{p}(\mathcal{S}|\mathcal{H}_i)}\mathrm{o}(\mathcal{H}_0)$$

### F. Transformation by Logarithm

To simplify our model, enabling easier intuition of the maths, we'll use $\log_{10}$. However, in practice we use $\ln(x)$ and $e^x$ notation. The $\log_{10}$ of the prior equation is

$$\begin{aligned}\log_{10}(\mathrm{o}(\mathcal{H}_0|\mathcal{S})) &= \log_{10}\left(\frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)}{\mathrm{p}(\mathcal{S}|\mathcal{H}_i)}\mathrm{o}(\mathcal{H}_0)\right) \\ &= \log_{10}\left(\frac{\mathrm{p}(\mathcal{S}|\mathcal{H}_0)}{\mathrm{p}(\mathcal{S}|\mathcal{H}_i)}\right) + \log_{10}\left(\mathrm{o}(\mathcal{H}_0)\right).\end{aligned}$$

Note that $\mathrm{o}(\mathcal{H}_0)$ is independent of $\mathcal{S}$. This fact allows us to use this term in the linear model for $\beta_0$ such that

$$\beta_0 = \log_{10}\left(\mathrm{o}(\mathcal{H}_0)\right),$$

where $\beta_0$ is the log of the prior odds.

Assuming the log-likelihood ratio is a linear function of the data, $\mathcal{S}$, then the log-odds can be represented in terms of the linear model,

$$\mathrm{lo}(\mathcal{H}_0|\mathcal{S}) = \beta_1\mathcal{S} + \beta_0,$$

where $\mathrm{lo}(\cdot)$ represents the log-odds function. Applying this function we can *learn* the likelihood ratio and prior odds, in log form, as a linear function of the data.

Because the model is in terms of odds, we can observe

- $\mathrm{o}(\mathcal{H}_0|\mathcal{S}) = 10$ means that coffee is 10 times more likely to be great.
- $\mathrm{o}(\mathcal{H}_0|\mathcal{S}) = \frac{1}{10}$ means that coffee is 10 times likely **not** to be great.

This inverse relationship can be transformed using the $\log_{10}$, and observe the exponential relationships

- $\log_{10}(\mathcal{H}_0|\mathcal{S}) = 1$ means that great coffee is 10 times more likely.
- $\log_{10}(\mathcal{H}_0|\mathcal{S}) = 2$ means that great coffee is 100 times more likely.
- $\log_{10}(\mathcal{H}_0|\mathcal{S}) = -1$ means that great coffee is 10 times **less** likely.
- $\log_{10}(\mathcal{H}_0|\mathcal{S}) = -2$ means that great coffee is 100 times **less** likely.

### G. The Inverse Logit

Note that *logit* is short for *log-odds*, therefore the *inverse logit* is simply the inverse log-odds. The current model is in terms of log-odds, which cannot be computed. The odds can be converted to probabilities by applying the rule

$$\mathrm{p}(\alpha) = \frac{\mathrm{o}(\alpha)}{1 + \mathrm{o}(\alpha)}.$$

Replacing $\log_{10}$ with $\ln$ and converting to an exponential function, then

$$\mathrm{o}(\mathcal{H}_0|\mathcal{S}) = 10^{(\beta_1 x + \beta_0)} = e^{(\beta_1 x + \beta_0)}.$$

Then

$$\mathrm{p}(\mathcal{H}_0|\mathcal{S}) = \frac{e^{(\beta_1 x + \beta_0)}}{1 + e^{(\beta_1 x + \beta_0)}}.$$

From the exponent relation

$$\frac{e^{\alpha}}{1 + e^{\alpha}} = \frac{1}{1 + e^{-\alpha}},$$

the inverse logit results

$$\mathrm{p}(\mathcal{H}_0|\mathcal{S}) = \frac{1}{1 + e^{-(\beta_1 x + \beta_0)}}.$$

Notice the logit function takes probabilities and transforms them into log-odds, whereas the inverse logit takes log-odds and turns them into probabilities!

## II. Formal Notation

Add formal notation and theory here...

## III. Example: Multinomial Regression for Ordinal Responses

Consider the automotive dataset for cars built between 1970 and 1990, with information

- Acceleration.
- Engine displacement.
- Horsepower.
- Vehicle weight.
- Manufacturer.
- Miles per gallon (MPG).

The logistic regression problem seeks the factors that are most significant to MPG. The predictor variables are the acceleration, engine displacement, horsepower, and weight of the cars. The response variable is MPG.

First create an ordinal response variable categorizing MPG into four levels from 9 to 48 MPG by labeling the response values in the range 9-19 as 1, 20-29 as 2, 30-39 as 3, and 40-48 as 4.

Then fit an ordinal response model for the variable *miles*.

```
load carb
X = [Acceleration Displacement Horsepower Weight];
miles = ordinal(MPG,{'1','2','3','4'},[],[9,19,29,39,48]);
[B,dev,stats] = mnrfit(X,miles,'model','ordinal');
```

resulting in

```
B = [—16.6895, —11.7208, —8.0606, 0.1048, 0.0103, 0.0645,
    0.0017]
```

The proportional odds model in this example is

$$\ln\left(\frac{p(MPG \leq 19)}{p(MPG > 19)}\right)$$
$$= -16.6895 + 0.1048x_a + 0.0103x_d + 0.0645x_h + 0.0017x_w$$
$$\ln\left(\frac{p(MPG \leq 29)}{p(MPG > 29)}\right)$$
$$= -11.7208 + 0.1048x_a + 0.0103x_d + 0.0645x_h + 0.0017x_w$$
$$\ln\left(\frac{p(MPG \leq 39)}{p(MPG > 39)}\right)$$
$$= -8.0606 + 0.1048x_a + 0.0103x_d + 0.0645x_h + 0.0017x_w$$

The coefficients express the relative risk or log odds of the MPG of a car being less than or equal to one value versus greater than that value. For example, the coefficient estimate of $0.1048$ indicates that a unit change in acceleration would impact the odds of the MPG of a car being less than or equal to 19 versus more than 19, or being less than or equal to 29 versus greater than 29, or being less than or equal to 39 versus greater than 39, by a factor of $e^{0.01048}$ given all else is equal.

Assess the significance of the coefficients

```
stats.p = [0.1899, 0.0350, 0.0000, 0.0118]
```

The $p$-values of $0.0350$, $0.0000$, and $0.0118$ for engine displacement, horsepower, and weight of a car, respectively, indicate that these factors are significant on the odds of MPG of a car being less than or equal to a certain value versus being greater than that value.

In five lines of code we have determined that the MPG is most influenced (smaller $p$-value) by engine displacement, horsepower, and weight, while acceleration is a by-product and thus less influential.