

## Logistic Regression Explanation

We start with a random variable  $Y$  following a Bernoulli distribution

$$Y \mid X = x \sim \text{Bernoulli}(p(x)), \quad Y \in \{0, 1\} \quad (1)$$

Alternatively,

$$\begin{aligned} P(Y = 1 \mid X = x) &= p(x) \\ P(Y = 0 \mid X = x) &= 1 - p(x) \end{aligned}$$

Now compute the conditional expectation,

$$\begin{aligned} E(Y \mid X = x) &= \sum y \cdot P(Y = y \mid X = x) \\ E(Y \mid X = x) &= 1 \cdot P(Y = 1 \mid X = x) + 0 \cdot P(Y = 0 \mid X = x) \\ E(Y \mid X = x) &= P(Y = 1 \mid X = x) \end{aligned}$$

Now we define:

$$P(x) = P(Y = 1 \mid X = x)$$

$$\boxed{P(x) = E(Y \mid X = x)} \quad (2)$$

Intuitively, because  $Y \in \{0, 1\}$ , average is equivalent to probability.

$E(Y \mid X = x)$  is telling us if we look at all patients with feature  $x$ , what is the average value of  $Y$  among them. Suppose 70 are malignant ( $Y = 1$ ) and 30 are benign ( $Y = 0$ ), the average value of  $Y = 0.7$ , but that is equivalent to  $P(Y = 1 \mid X = x)$  because 70 out of 100 were malignant. Because  $Y$  only takes values 0 and 1, When taking the average of 0's and 1's: The average equals the fraction of 1's. And the fraction of 1's is the probability of 1.

# Linear Regression

Suppose we have random variables  $Y, X_1, X_2, \dots, X_p$ . We define 2 criteria for a linear model, namely: (note, this idea is taken from [math.stackexchange](#))

1.  $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon$
2.  $E(Y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$

Where  $\beta_0, \beta_1, \dots, \beta_p$  are constants.

For criteria 2, we enforce this linear relationship on average because real-world data contains randomness. We can imagine if it were exact in the sense  $Y = \beta_0 + \beta_1 X$  with no error term, it will not make sense of what we observe in reality. Two people studying the same amount of hours, say 5 hours may not score the same in an exam.

Take the conditional expectation of 1:

$$E(Y \mid X_1 = x_1, \dots, X_p = x_p) = E(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon \mid X_1 = x_1, \dots, X_p = x_p) \quad (3)$$

Since we condition on  $X_1 = x_1, \dots, X_p = x_p$ , the random variables  $X_i$  are "fixed" at those values.

$$E(Y \mid X_1 = x_1, \dots, X_p = x_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + E(\epsilon \mid X_1 = x_1, \dots, X_p = x_p)$$

Plugging in criteria 2:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + E(\epsilon \mid X_1 = x_1, \dots, X_p = x_p)$$

$$E(\epsilon \mid X_1 = x_1, \dots, X_p = x_p) = 0 \quad (4)$$

The statistical definition of a linear regression model is defined as:

$$\boxed{E(Y \mid X_1 = x_1, \dots, X_p = x_p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} \quad (5)$$

Remember we did not derive this, because that line of reasoning is circular from criteria 2, we however did derive the "Zero Conditional Mean Assumption" in (4).

## Back to Logistic Regression

In Linear Regression, the random variable  $Y$  is continuous, in logistic regression, it is binary. Most importantly, notice (2) and (5) are both modelling the conditional expectation. Let us now try to define the logistic regression, keeping in mind a constraint:

$$0 < P(x) < 1 \quad (6)$$

Looking at (2) and (5), a natural first attempt is:

$$P(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

However, this linear function is unbounded and contradicts (6). It's very important to understand we want some probability related quantity to equal a linear function, so we attempt to do a transformation on  $P(x)$  to match the unboundedness of the linear function. In other words we want some transformation  $g$  such that:

$$g : (0, 1) \rightarrow (-\infty, \infty)$$

We shall define the odds to remove the upper bound of 1.

$$odds = \frac{P}{1-P}, \quad \text{bound is now } (0, \infty)$$

$odds > 1$  if  $Y = 1$  is more likely,  $odds < 1$  if  $Y = 0$  is more likely. Now take the  $\ln$  to remove the lower bound.

$$\ln \left( \frac{P(x)}{1-P(x)} \right), \quad \text{bound is now } (\infty, \infty)$$

Therefore, the logistic regression model is defined as the Logit:

$$\boxed{Logit = \ln(odds) = \ln \left( \frac{P(x)}{1-P(x)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p} \quad (7)$$

Note in lecture notes,  $P(x)$  is referred to as  $P(Y = 1)$ . The bigger the *Logit* is, the bigger  $P(Y = 1)$  is.

Because we want to predict probability which is  $P(x)$ , we now solve for  $p(x)$  by multiplying by exponent on both sides.

$$e^{\ln\left(\frac{P}{1-P}\right)} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

$$\frac{P}{1-P} = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

$$\begin{aligned} P &= e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p} - P e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p} \\ &= P \left[ \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{P} - e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p} \right] \\ 1 &= \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{P} - e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p} \end{aligned}$$

$$P(1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}) = e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}$$

$$P = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p}}$$

Therefore, we arrive at the Sigmoid function:

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p)}} \quad (8)$$

In R, the function `predict()` has the following syntax:

```
predict(object from "glm", newdata = NULL,
        type = c("link", "response", "terms"),
        se.fit = FALSE, dispersion = NULL, ...)
```

and returns  $P$ , the Sigmoid function when `type = "response"`

## Application (Graduate Admissions Logistic Regression supplementary notes)

In R, when we performed:

```
grad$rank <- factor(grad$rank, levels = c(1, 2, 3, 4))
logit_model <- glm(admit ~ gre + gpa + rank,
                  data = train_data,
                  family = binomial)
```

R now understands `rank` as categorical, not numeric, so R creates a dummy variable where `rank 1` is the baseline. The logistic regression model now becomes:

$$\ln(odds) = \beta_0 + \beta_1 gre + \beta_2 gpa + (-0.89)rank2 + (-1.48)rank3 + (-1.80)rank4$$

Suppose we want to compare the difference in odds between rank 1.

For rank 1 (all dummy variables = 0):

$$\ln(odds_1) = \beta_0 + \beta_1 gre + \beta_2 gpa$$

For rank 2 (only rank 2 = 1):

$$\ln(odds_2) = \beta_0 + \beta_1 gre + \beta_2 gpa + (-0.89)$$

$$\ln(odds_2) - \ln(odds_1) = -0.89$$

$$\ln\left(\frac{odds_2}{odds_1}\right) = -0.89$$

$$\frac{odds_2}{odds_1} = e^{-0.89} \approx 0.41$$

$$odds_2 = 0.41 odds_1$$

$$1 - 0.41 = 0.59$$

Therefore, rank 2 odds are 59 percent lower than rank 1, meaning applicants from rank 2 institutions have a 59 percent lower odds of admission than those from rank 1 institutions.