37 2025-05-05 | Week 16 | Lecture 37

37.1 Odds

Suppose A is an event with probability p. The *odds* or *odds ratio* is the fraction

$$\frac{p}{1-p}.$$

The idea of *odds* (which can be any positive number) is older than that of probability (which can only be betwen 0 and 1). When outcomes are equally likely, the *probability* p is understood to be

$$\frac{\text{number of favorable outcomes}}{\text{total number of possible outcomes}},$$

whereas *odds* is

$$\frac{\text{number of favorable outcomes}}{\text{number of unfavorable outcomes}}$$

For example, the probability of rolling a 6 on a dice is 1/6, and the odds is 1/5. In some sense knowing the odds gives you the same amount of information as knowing the probability (since if you know one, you can compute the other), but the interpretation is different. For example, odds of 1/5 tell us that failure is 5 times more likely than success. We would say the odds are "5-to-1 against", or "1-to-5 against".

Odds are commonly used in betting situations. Odds is useful in betting situations because it gives a measure of the player's advantage. The first mathematical text on probability was Girolamo Cardano's 15-page text "The Book on Games of Chance" written in 1564 (though first published only a century later). Much of the work is in terms of "odds", and he appears to be the first to formulate probability in terms of the ratio of favorable outcomes to total outcomes.

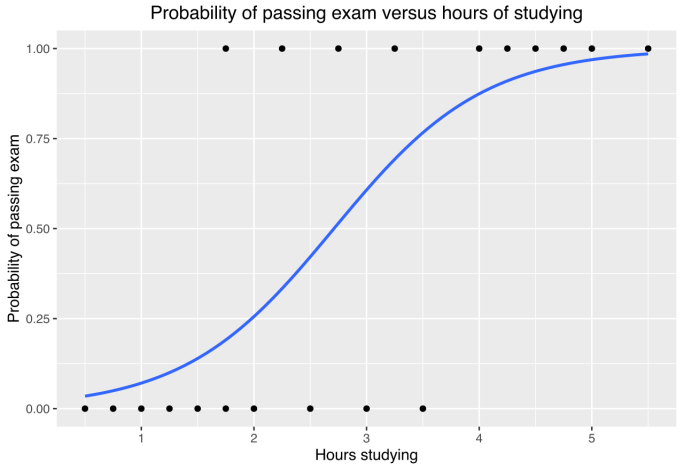
37.2 Logistic Regression

Suppose we have a numerical predictor variable x, which can take a range of values, and we are interested in a binary response variable Y, which can take only values 0 and 1. That is Y provides a binary classification. For example,

$$x = \text{ mileage of car} \quad \text{and} \quad Y = \begin{cases} 1 & : \text{ if the car needs maintenance} \\ 0 & : \text{ if the car doesn't need maintenance} \end{cases}$$

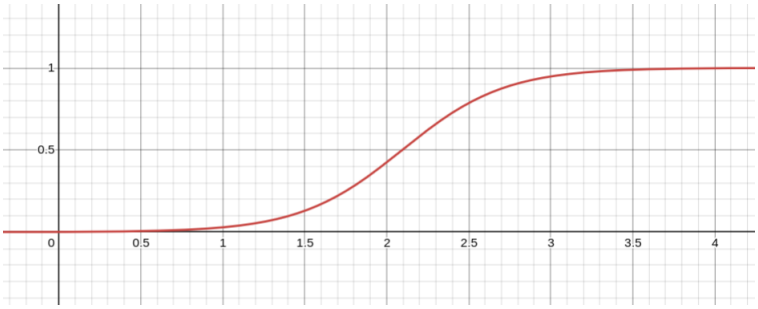
For the wiki problem `https://en.wikipedia.org/wiki/Logistic_regression#Example`, we have data

Hours (x_k)	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass (y_k)	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



For this setting, it doesn't make sense to try to fit a best line, since there are only 2 possible y-values: 0 and 1. instead, we consider the probability $p(x) = \mathbb{P}[Y = 1]$, which increases from 0 to 1 as the mileage x increases.

The *logistic function* is the function $f(x) = \frac{e^{C+Dx}}{1+e^{C+Dx}}$, where $C, D \in \mathbb{R}$. An example when $D > 0$ is the following increasing function



In *logistic regression*, we make the **assumption** that the logarithm of the odds is approximately linear, i.e, that

$$\log\left(\frac{p(x)}{1-p(x)}\right) \approx C + Dx \tag{37}$$

for some choice of C and D. This is equivalent to assuming that the probability of, say, engine failure is approximated by the logistic function, so we assume that

$$p(x) = \frac{e^{C+Dx}}{1+e^{C+Dx}} \tag{38}$$

for some choice of real numbers C and D.

When we found a "best fit line" for some data, recall that we looked for values of C and D which minimized the magnitude of the error vector

$$e = (e_1, e_2, \ldots, e_n)$$

where $e_i = y_i - (C + Dx_i)$ is the vertical distance between our observed data and what is predicted by the line $y = C + Dx$.

For logistic regression, we do something similar, but instead of minimizing the sum of squares $\|e\|^2 = e_1^2 + e_2^2 + \ldots + e_n^2$, we instead seek to minimize something much more complicated: *the surprisal*.

Consider the single data point (x_i, y_i). Here y_i is either 0 or 1, but $x_i \in \mathbb{R}$.

According to the model given in Eq. (38), for any choice of model parametesr C and D, the predicted probability that $y_i = 1$ is

$$p(x_i) = \frac{e^{C+Dx_i}}{1+e^{C+Dx_i}}$$

Let's call this number p_i and observe that $0 < p_i < 1$. The *surprisal* (or "*log loss*") at the data point (x_i, y_i) is the quantity

$$\ell_i := -\log\left(p_i^{y_i}(1-p_i)^{1-y_i}\right) = \begin{cases} \log(\frac{1}{p_i}) & : \quad y_i = 1 \\ \log\left(\frac{1}{1-p_i}\right) & : \quad y_i = 0 \end{cases}$$

Observations

- $\ell_i > 0$
- ℓ_i really does measure "surprise" in some sense. For example, if $p_i \approx 0$ and $y_i = 0$, then $\ell_i \approx \log(1) = 0$. But if $p_i \approx 1$ and $y_i = 0$, then $\ell_i \approx \log(\text{big number}) = \text{big}$. Bigger values means there is more discrepancy between what is predicted by the model and what is observed.

The *total loss* is the sum of all the surprisals:

$$T(C, D) = \ell_1 + \ell_2 + \ldots + \ell_n$$

Observations

- T is a complicated function of the parameters C and D.
- Each choice of C, D correspond to some model from the family Eq. (38)
- When $T(C, D)$ is large, the observed data $(x_1, y_1), \ldots, (x_n, y_n)$ would be unlikely to occur for that choice of C and D. But if $T(C, D)$ is close to zero, then the model would be more likely to produce data like what was observed.
- In this way the quantity $T(C, D)$ may be regarded as a "distance" between your observed data and the model with parameters C, D. The goal is to find the values of C, D which minimize this distance—these will give you the "best" model for your data.

In logistic regression, the goal is to try to find the values of C, D which minimize the $T(C, D)$. This cannot reasonably be done by hand, as the function T is complicated enough that we don't have closed form solutions. Instead, more typically, one will use software to check many values of C and D and use whichever appears to give you the smallest $T(C, D)$. Actual implementations use sophisticated heuristics to do this.

For the wiki problem, the best C and D we get are $C \approx -4.1$ and $D \approx 1.5$, this gives the "best model" to be

$$p_{\text{best}}(x) = \frac{e^{-4.1+1.5x}}{1 + e^{-4.1+1.5x}}$$

and which gives the curve graphed above.

This model can then be used to make predictions.For example, if you study 3 hours, your probability of passing is predicted to be

$$p_{\text{best}}(3) = \frac{e^{-4.1+1.5(3)}}{1 + e^{-4.1+1.5(3)}} \approx 0.6.$$

The example we have worked has assumed that there is only one explanatory variable x. This restriction isn't essential to the theory. Suppose instead that we have 2 explanatory variables, x and w. Now, instead of Eq. (37) we have

$$\log\left(\frac{p(x)}{1 - p(x)}\right) \approx C + Dx + Ew$$

We have added a new parameter E. So now we must minimize the total loss over all C, D, and E, rather than just over all C and D. Nothing else changes. You can add as many explanatory variables as you like.

If you have a dataset with lots of possible explanatory variables, how do you know which ones to include and which to exclude from a logistic regression analysis? For example: Y indicates whether a surgery patient experiences complications due to surgery (assume same type of surgery). Possible explanatory variables

- patient age, sex
- surgical device type
- year that surgery was performed
- attending physician
- preexisting conditions
- innumerable other risk factors

If you include all of them, you get results that may be overfitted and may be hard to interpret. One option is to repeatedly do many analyes with different subsets of your explanatory variables, with some penalty applied so that models with fewer parameters are preferred (see, e.g. the Bayesian Information Criterion (BIC) score). These are related to goodness of fit tests like the ones we did for the skittle distribution.