Week11

November 8, 2021

1 Week11

Up until this point in lecture we discussed how to model the conditional distribution of a random variable with a normal distribution—linear regression. Linear regression can be used to understand how a set of covariates $(x_1, x_2, \cdots, x_p)$ relate linearly to a random variable Y. We assume that Y is a normally distributed variable. Because we assume Y can be modeled as a normally distributed random variable we expect that our Y data is some set of negative, positive decimal data (i.e. -2.12, 3.14, 78.2, etc).

But what if our Y data is instead a set of 0s and 1s that represent the absence and presence of some phenomena? Our plan is to explore logistic regression.

The goal of logistic regression is to model a set of Y data with binary responses (Yes/No), (Presence/Absence).

1.1 Data setup

Suppose we have a dataset of p covariates and a single target variable y. Then our dataset can be represented as

$$\mathcal{D} = [(x_1^1, x_1^2, \cdots, x_1^p, y^1), (x_2^1, x_2^2, \cdots, x_2^p, y^2), \cdots, (x_N^1, x_N^2, \cdots, x_N^p, y^N)] \tag{1}$$

where x_a^b corresponds to the ath covariate from the bth datapoint. Note that our setup above is identical to our setup for multivairate linear regression (MLR).

The difference between MLR and logistic regression (LR) is that in LR each y_i is either the value 0 or 1.

1.2 A model for Y

Lets start to model Y by first searching or a random variable that generates the value 0 or 1. We know that a Bernoulli distributed random variable with parameter θ will generate the value 1 with probability θ and the value 0 with probability $1 - \theta$.

That is, if Y is a Bernoulli distributed random variable then

$$\begin{aligned} Y &\sim \text{Bern}(\theta) && (2)\\ p(Y=1) &= \theta && (3)\\ p(Y=0) &= 1-\theta && (4)\\ p(Y=y) &= \theta^{y}(1-\theta)^{1-y} && (5) \end{aligned}$$

The expected value of Y is $\mathbb{E}(Y) = \theta$ and the variance of Y is $\mathbb{V}(Y) = \theta(1-\theta)$.

When we explore MLR, we started by assuming our Y had a normal distriubution $\mathcal{N}(\mu, \sigma^2)$ and then modified μ so that it was a function that depended on parameters β and x data. We chose to modify μ because the expected value of Y was μ.

Let us take the same approach with our Y above and model the conditional distribtuion of Y given parameters β and x data as

$$\begin{aligned} Y_i|\beta_0, \beta_1, x_i &\sim \text{Bern}(\theta(x)) && (6)\\ \theta(x) &= \beta_0 + \beta_1 x && (7)\\ & && (8) \end{aligned}$$

But we have a problem. The parameter θ is constrained to be a value between 0 and 1, yet the quantity $\beta_0 + \beta_1 x$ can take any value from negative to positive infinity. We need a way to constrain our values for θ to be between 0 and 1.

1.2.1 The logistic function

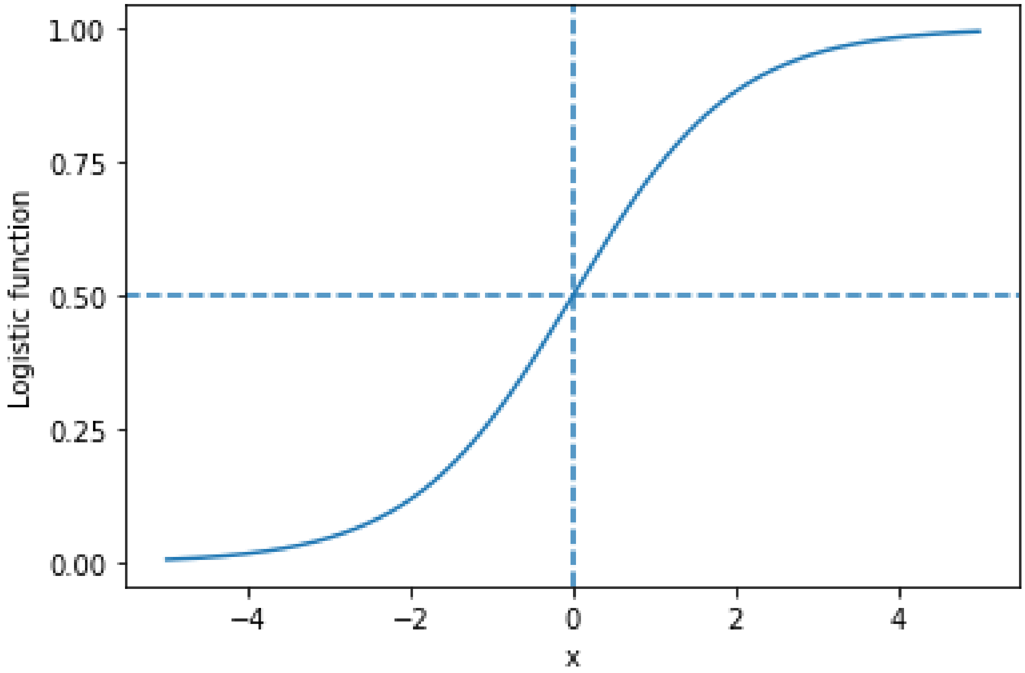
One method to constrain θ to be between 0 and 1 is to use the logistic function. The logistic function f is

$$f(x) = \frac{e^x}{1+e^x} \tag{9}$$

For x values that approach positive infinity the logistic function approaches the value 1. For x values that approach negative infinity the logistic function approaches the value 0.

[8]:

```python
def logistic(x):
    import numpy as np
    e = np.exp(x)
    return e/(1+e)

fig,ax = plt.subplots()
domain = np.linspace(-5,5,10**3)

ax.plot(domain,logistic(domain))

ax.set_xlabel("x")
ax.set_ylabel("Logistic function")

ax.axvline(0,ls="--")
ax.axhline(0.5,ls="--")

ax.set_yticks([0,0.25,0.50,0.75,1.0])

plt.show()
```



Because logistic function maps values from negative infinity to positive infintiy to numbers between 0 and 1, we can use this function to give us valid θ values.

$$Y_i|\beta_0, \beta_1, x_i \sim \text{Bern}(\theta(x)) \tag{10}$$
$$\theta(x) = L(\beta_0 + \beta_1 x) \tag{11}$$
$$\tag{12}$$

where L is the logistic functon. In other words,

$$Y_i|\beta_0, \beta_1, x_i \sim \text{Bern}(\theta(x)) \tag{13}$$
$$\theta(x) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{14}$$
$$\tag{15}$$

1.3 A way to write logistic regression that looks more familiar

At times it can be convienant to rewrite our logistic regression model above to look more similar to the more familar multivariate linear regression. For multivariate linear regression, the expedcted value of our target variable was equal to $\beta_0 + \beta_1 x$.

or logistic regression our expected value is

$$\mathbb{E}(Y) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{16}$$

Let's rearrange the above equation to isolate $\beta_0 + \beta_1 x$.

$$\begin{align}
\theta = \mathbb{E}(Y) &= \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}} \tag{17}\\
1 + e^{\beta_0+\beta_1 x}(\theta) &= e^{\beta_0+\beta_1 x} \tag{18}\\
\theta + \theta e^{\beta_0+\beta_1 x} &= e^{\beta_0+\beta_1 x} \tag{19}\\
\theta &= e^{\beta_0+\beta_1 x} - \theta e^{\beta_0+\beta_1 x} \tag{20}\\
\theta &= e^{\beta_0+\beta_1 x}(1-\theta) \tag{21}\\
\frac{\theta}{1-\theta} &= e^{\beta_0+\beta_1 x} \tag{22}\\
\log\left(\frac{\theta}{1-\theta}\right) &= \beta_0 + \beta_1 x \tag{23}
\end{align}$$

The ratio $\frac{\theta}{1-\theta}$ is called the **odds** and so the quantity $\log\left(\frac{\theta}{1-\theta}\right)$ is called the **log odds**.

1.4 A 1-unit change in X

Just like in MLR, we can look at how