Logistic Regression

We consider a specific kind of response where $Y \in \{0, 1\}$

- We can write this as $Y|X = x \sim \text{Bernoulli}$
- The p parameter of the Bernoulli is dependent on x
  - $p = f(x)$
- The regression function is then:

$$\mathbb{E}\left[Y|X = x\right] = f(x)$$

  - By definition, this cannot be a linear function because those eventually go to infinity
  - We need $f(x)$ to be constrained between 0 and 1
- We use the σ function to squish values between 0 and 1

$$\sigma(t) = \frac{e^t}{1 + e^t}$$

  - We therefore use the model $f(x) = \sigma(x^T \beta^*)$ for logistic regression

Another choice is $\Phi(t)$, where Φ is the Gaussian cdf

- This leads to a **probit regression model**
- It has the representation where:

$$Y = 1(X^T \beta^* + Z > 0)$$

- We can actually choose the cdf of any random variable as long as it is symmetric about 0

MLE

The reason logistic regression is commonly used is because it has a concave log likelihood, making it possible to use gradient ascent to optimize

$$\ell(\beta) = \sum_{i=1}^{n} \left(Y_i X_i^T \beta - \log\left(1 + e^{X_i^T \beta}\right)\right)$$

Multiclass

A common extension is when $Y \in \{0, \ldots, M\}$

An initial model might be to use $M + 1$ coefficient vectors and assume:

$$p_j(x) = \frac{e^{x^T \beta_j^*}}{\sum_l e^{x^T \beta_l^*}}$$

- This doesn't work because when $M = 1$, we don't reduce down to our logistic regression case
- This has $M + 1$ degrees of freedom while we really only have M

Instead, let's look at what $x^T \beta$ represented in the original 2 class regression problem:

$$x^T \beta = \log \frac{f(x)}{1 - f(x)} = \log \frac{p_1(x)}{p_0(x)}$$

- We can generalize this by setting $x^T \beta_j = \log \frac{p_j(x)}{p_0(x)}$ for $j \in 1, \ldots, M$
- This gives us the model:

$$p_j(x) = \frac{e^{x^T \beta_j^*}}{1 + \sum_{l=1}^{M} e^{x^T \beta_l^*}}$$

- This does reduce to the model when $M = 1$

- The log likelihood is more complicated, and turns out to be the negative cross entropy loss from ML