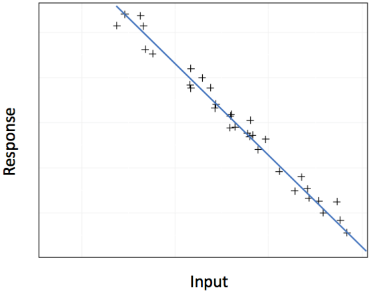
Logistic Regression

Clayton W. Schupp

Galvanize

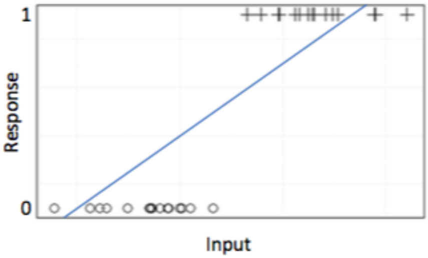
Review of Linear Regression



- Models a continuous response as a function of one or more input variables
- Assumes the response is a linear function of the input
- Finds the linear function that gives the best fit (minimizes residual error)

Classification

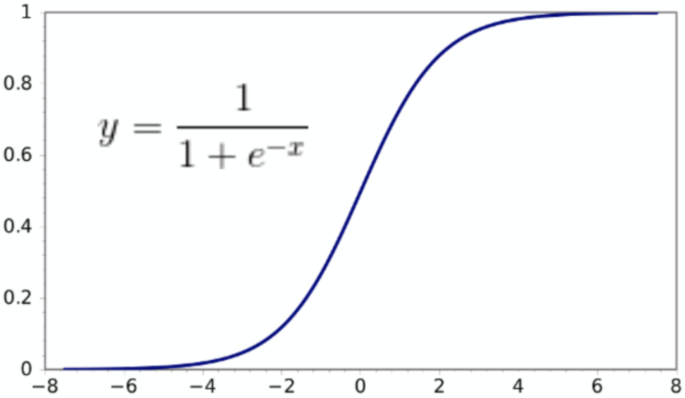
In the classification setting, the input stays the same, but now response is discrete $\longrightarrow$ we could still use linear regression, but it doesn't really fit the data well



We would prefer a function with the following properties:

- Can take a continuous input $(-\inf, \inf)$
- Output should be between 0 and 1
- Output should not 'waste much time' transitioning between 0 and 1

The Logistic (Sigmoid) Function



Logistic Regression: Output

The logistic regression model is a generalized linear model where the response variable is **binary**.

$$Y_i = \begin{cases} 1, & \text{if an event occurs} \\ 0, & \text{if it doesn't} \end{cases}$$

We are interested in the probability that an event occurs given a subject's profile

$$\pi_i = P(Y_i = 1 | X = x_i)$$

x_i represents the vector of feature values for the i^{th} subject

The distribution of $Y_i | X = x_i$ is **Bernoulli(π_i)**

Logistic Regression: Input

Just as with linear regression, the linear predictor is

$$X\beta = \beta_0 x_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where X is your design matrix: x_0 is a column vector of 1's and $x_1, \ldots, x_p$ are the feature column vectors

Note: we have p features and $p + 1$ parameters

The explanatory variables may be quantitative, categorical, or mixed

Logistic Regression: Linking Input to Output

The function that connects the linear predictor to the desired output is the logistic function

$$y = \frac{1}{1 + e^{-x}}$$

Giving us the logistic regression model

$$\pi_i = \frac{1}{1 + e^{-X\beta}} = \frac{e^{X\beta}}{1 + e^{X\beta}}$$

An Aside: Odds and Probabilities

Given a probability (p) of an event occurring, the odds of that event are

$$odds = \frac{p}{1-p}$$

Similarly, given the odds, you can calculate the probability

$$p = \frac{odds}{1+odds}$$

An Aside: Probabilities and Odds

Odds are commonly used in gambling, especially horse-racing

- Even odds (1:1) $\longrightarrow p = \frac{1}{1+1} = 0.5$
- Odds are 3:1 for an event $\longrightarrow p = \frac{3}{1+3} = 0.75$
- Long shot: 20:1 against

$$\longrightarrow 1 - p = 1 - \frac{20}{21} = \frac{1}{21} = 0.0476$$

Anybody ready for Vegas???

Back to the Logistic Regression Model

The logistic model can be rewritten in terms of odds via the logit function

$$logit(\pi) = log\left(\frac{\pi_i}{1-\pi}\right) = logodds$$

Giving us a nice framework that seems familiar

$$logit(\pi_i) = X\beta$$

Estimating the Parameters

Since $Y_i|X$ is Bernoulli(π_i), the likelihood of a single data point (y_i) is

$$\pi_i^{y_i} \cdot (1 - \pi_i)^{1-y_i}$$

Therefore the likelihood of all the data is

$$\begin{aligned} L(\beta) &= \prod_{i=1}^{n} \pi_i^{y_i} \cdot (1 - \pi_i)^{1-y_i} \\ &= \prod_{i=1}^{n} \left(\frac{e^{X\beta}}{1 + e^{X\beta}} \right)^{y_i} \cdot \left(\frac{1}{1 + e^{X\beta}} \right)^{1-y_i} \end{aligned}$$

Estimating the Parameters

The regression coefficients can be estimated using maximum likelihood estimation

Unlike linear regression, no closed form solution exists, therefore an iterative method such as Newton-Rhapson or Gradient Descent is needed

Reasons that the model may not reach convergence

- A large number of features relative to subjects $\longrightarrow$ rule of thumb is at least 10 cases for each explanatory variable
- Multicollinearity
- Sparseness, specifically low cell counts for categorical predictors

Evaluating Goodness-of-Fit

With linear regression, we could assess the fit using the R^2 which is essentially a transformation of the residual sum of squares

Deviance is analogous in the logistic regression setting

$$D = -2ln(likelihood) \sim \chi^2_{df}$$

We can now calculate the deviance for a fitted model D_{fitted} and a null intercept-only model (D_{null}) which will allow us to calculate the pseudo-R^2

$$R^2_L = \frac{D_{null} - D_{fitted}}{D_{null}}$$

Note: unlike in linear regression, this cannot be interpreted as proportion of variance explained

Comparing Models

If you want to compare two nested models:

$$H_0 : \text{reduced model} \qquad\qquad H_1 : \text{full model}$$

Then you can calculate the following test statistic

$$D_{red} - D_{full} \sim \chi^2_{df}$$

where df = number of parameters removed

If you want to compare two non-nested models, you can pick the model that minimizes Akaike's Criterion (AIC)

An Aside: Odds Ratio

Given the definition of odds above, the odds ratio is

$$OR = \frac{Odds_1}{Odds_2} = \frac{(p_1/(1-p_1)}{(p_2/(1-p_2)}$$

For example, say the probability of a disease in individuals with a certain genetic trait is $p_1 = 0.05$ while in the general population its $p_2 = 0.001$ the resulting odds ratio would be

$$OR = \frac{0.05/0.95}{0.001/0.999} \approx 53$$

This represents a measure of relative risk such that an individual with the genetic trait is 53 time more likely to develop the disease than a randomly chosen person

Model Interpretation

In linear regression, the $\hat{\beta}$ coefficients can be interpreted directly as the change in y for a 1-unit increase in the explanatory variable

In logistic regression, however, this would represent the change in logit value for a 1-unit increase in the explanatory variable, which is not interpretable

We can however convert the $\hat{\beta}$ coefficient to an estimate of Odds Ratio for a 1-unit increase in the explanatory variable

$$\widehat{OR} = e^{\hat{\beta}}$$

Making Predictions

Once the $\hat{\beta}$ coefficients have been calculated, we can estimate the probabilities of the event occurring $(Y = 1)$ for a specific covariate profile (X)

$$\hat{\pi} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}}$$

$\hat{\pi}$ is a vector of probabilities for the entire sample.

To find a specific probability π_i, you would find the dot product of the i^{th} row of the X matrix and the vector of $\hat{\beta}$ coefficients

Uses of Logistic Regression

- To **model the probabilities** of certain conditions or states as a function of explanatory variables $\longrightarrow$ identify "Risk" factors for certain conditions (i.e. disease, divorce, etc)
- To describe differences between subjects from different groups
- To adjust for the "bias" in comparing 2 groups in an observation study $\longrightarrow$ propensity scores

Uses of Logistic Regression

- To **predict probabilities** that subjects fall into one of 2 categories on a dichotomous response variable
- To **classify** subjects into one of 2 categories
  $\longrightarrow$ one of our main focuses
- Lots of other possibilities

Classification

General Method

1. For each unclassified subject, you would calculate the probability the subject falls in a specific class using the fitted model
2. You would compare that probability to a predetermined decision rule boundary $\longrightarrow$ default is 0.5
3. If the predicted probability is $>$ than 0.5 $\longrightarrow$ classify as 1
4. If the predicted probability is $\leq$ 0.5 $\longrightarrow$ classify as 0