HUDM 6122
Multivariate Analysis

Logistic Regression and Classification

Classification and Logistic Regression

- The linear regression model and classification models discussed so far assume the variables are quantitative (numerical);
- If, instead, the response variable is *qualitative*, (or *categorical*), the task of predicting responses is aka *classification;*
- Logistic regression is an approach for classification when the response variable Y is restricted to two values (dichotomous);
- For example, Y may be recorded as “male” or “female”, or ”employed” and “unemployed”;
- Such variables are coded as 0 and 1;
- The parameter of interest is $p = \mathrm{P}(Y = 1 \mid \boldsymbol{X})$, the proportion of the population coded as 1;
- Note that the mean of Y is also $p = \mathrm{E}(Y) = 0\times(1\text{-}p) + 1\times p$

Linear Regression Approach

- With a dichotomous (0 or 1) outcome *Y*, one could use a linear probability model (LPM) with the usual assumptions.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad iid$$

- The predicted values for the linear probability model are

$$\hat{Y}_i = E(Y_i \mid X_i = x_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip}$$

- For a dichotomous outcome,

$$\mathrm{E}(Y_i \mid X_i = x_i) = \Pr(Y_i = 1 \mid X_i = x_i)$$

Why Not Linear Regression?

- Predicted values from linear regression are not constrained to the interval [0,1], even though probabilities should be.
- Residuals from the linear model are not normally distributed because they are dichotomous for each $X = x$:

$$\varepsilon_i = \begin{cases} -p(X = x) & \text{if} \quad Y_i = 0 \\ 1 - p(X = x) & \text{if} \quad Y_i = 1 \end{cases}$$

- Also, the variance of ε_i is $p(X = x)[1 - p(X = x)]$, which means it is not constant.
- Furthermore, if the outcome has more than two categories, linear regression becomes impossible unless they are ordered, and we assume the distances between all categories are identical.
- We need a method that deals with the above-mentioned deficiencies.

Logistic Regression: Foundation

- Assume the outcome Y is coded as 0/1. Our goal is to specify a model for

$$p(X) = \Pr(Y = 1 \mid X = x)$$

- In linear regression we use a linear model in the covariates $X_1, \ldots, X_p$:

$$p(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- The model above, however, has a range equal to $(-\infty, \infty)$. Therefore, instead of modeling p we will model the odds ratio:

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

which is the ratio of the probability of 1 to the probability of 0.

- Note that, unlike probability, the odds ratio can be greater than 1 (indeed, could be equal to any positive number).

Logistic Regression: Foundation

- Since the odds can't be negative, we need a further transformation.
- We will model the natural log of the odds, called the *logit*:

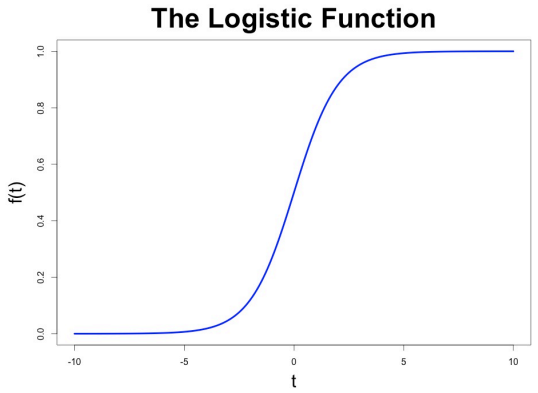
$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right)$$

- Note that if we solve for p we obtain:

$$p = \frac{e^{\text{logit}}}{1 + e^{\text{logit}}}$$

and the solution is restricted to (0,1). That is, we used a transformation function f:(-∞, ∞) → (0,1). It is called the *logistic function*

$$f(t) = \frac{\exp(t)}{1+\exp(t)}.$$



Simple Logistic Regression

- Assume there is only one predictor z
- Then:

$$\text{logit}(p) = \ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 z$$

- In other words, the log odds are *linear* in z.
- You can write the same model in terms of *probabilities*:

$$\frac{p(z)}{1-p(z)} = e^{\beta_0+\beta_1 z}$$

and finally solve for $p(z)$:

$$p(z) = \frac{e^{\beta_0+\beta_1 z}}{1+e^{\beta_0+\beta_1 z}}$$

The last equation follows the logistic curve from previous slide.

Other Sigmoidal Response Functions

- The *logistic function* is a prime candidate:

$$f(t) = \frac{\exp(t)}{1+\exp(t)}.$$

- An alternative is the *probit function*:

$$\Phi(t)$$

where Φ is the cdf of the standard normal distribution.

- Theoretically, *any cumulative distribution function* can serve as a response function.

Logistic Regression: Foundation

- Applying the logistic function to the linear transformation of the predictors gives three equivalent formulations.

$$p(X) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)}$$ probability

or

$$\frac{p(X)}{1 - p(X)} = \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)$$ odds

or

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$ log(odds) or "logit"

- The logit is linear in $X_1, \ldots, X_p$.

Logistic Regression: Estimating Betas

- Just as in simple regression, the coefficients β_0 and β_1 are unknown and must be estimated using available data. *Maximum likelihood* is the most commonly used method for this problem.
- The likelihood that needs to be maximized is:

$$\mathrm{L}(\beta_0, \beta_1, \ldots, \beta_p) = \frac{\prod_{i=1}^{n} e^{y_i(\beta_0+\beta_1 z_{i1}+\cdots+\beta_p z_{ip})}}{\prod_{i=1}^{n}\left(1 + e^{(\beta_0+\beta_1 z_{i1}+\cdots+\beta_p z_{ip})}\right)^{1-y_i}}$$

- Unlike the closed-form analytical solution (i.e., the normal equations) available for linear regression, the score function for logistic regression is transcendental. That is, there is no closed-form solution so we must solve with numerical optimization methods such as Newton's method (IRLS) which is done in several iterations.