Chapter 7: Logistic Regression and Generalized Linear Models

An Online Course

Sponsored by The Georgia R School

Presented by Geoffrey S. Hubona

Binary and Count Response Variables

- Ordinary Least Squares models assume the response variable to be (approximately) normally distributed. However, many experiments require an assessment of the relationship between covariates and a binary response variable.
  - A variable measured **at only two levels** or with **counts**.
- Generalized linear models provide a framework to estimate regression models with non-normal response variables.
- The regression relationship between the covariates and the response is modelled by a linear combination of the covariates.

ESR and Plasma Proteins

- Erythrocyte sedimentation rate (ESR) is the rate at which red blood cells settle out of suspension in blood plasma.
    - ESR < 20mm/hr indicates a ‘healthy’ individual
- IF ESR increases when the level of certain blood plasma proteins rise in association with particular medical conditions, it might be useful to screen blood donors for these conditions.
- Question of interest is: ***Whether there is any association between the probability of an ESR reading > 20mm/hr and the levels of the two plasma proteins?***

ESR and Plasma Proteins

- 32 observations, 3 Data Variables:
  - **Fibrinogen:** numeric;
  - **Globulin:** numeric;
  - **ESR:** factor (levels are ***< 20mm/hr*** and ***> 20mm/hr***).

Women's Role in Society

- In a survey carried out in 1974/1975, respondents were asked if he or she agreed or disagreed with the statement "Women should take care of running their homes and leave running the country up to men".
- Questions of interest are ***whether the responses of men and women differ*** and ***how years of education affect the response***.

Women's Role in Society

- 40 Observations, 4 Data Variables:
  - **Education:** years of education (integer);
  - **Gender:** factor (levels are ***male*** and ***female***);
  - **Agree:** number of subjects in agreement with the statement.
  - **Disagree:** number of subjects in disagreement with the statement.

Colonic Polyps

- Data from a placebo controlled trial of a non-steroidal anti-inflammatory drug in the treatment of familial andenomatous polyposis.
- Questions of interest is ***whether the number of polyps is related to treatment and/or age of patients***.

Colonic Polyps

- 20 Observations, 3 Data Variables:
  - **Number:** number of colonic polyps at 12 months;
  - **Treat:** factor (levels are ***placebo*** and ***drug***);
  - **Age:** age of the patient.

Driving and Back Pain

- Study to investigate whether driving a car is a risk factor for low back pain resulting from acute herniated lumbar intervertebral discs (AHLID).
- A *case-control* study was used with cases selected from people diagnosed with AHLID.
    - 217 matched pairs (128 males, 89 female).
- Cases were all from the same admitting hospital and were also matched on *age* and *gender*.

Driving and Back Pain

- 217 Matched Pairs, 4 Data Variables:
  - **ID:** factor which identifies match pairs;
  - **Status:** factor (levels are ***case*** and ***control***);
  - **Driver:** factor (levels are ***no*** and ***yes***);
  - **Suburban:** factor (levels ***no*** and ***yes***) indicating suburban residency.

Logistic Regression

The ordinary multiple regression model is described as $y \sim \mathcal{N}(\mu, \sigma^2)$ where $\mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$.

This makes it clear that this model is suitable for continuous response variables with, conditional on the values of the explanatory variables, a normal distribution with constant variance.

So clearly the model would not be suitable for applying to the erythrocyte sedimentation rate since the response variable is binary.

Logistic Regression

For modelling the expected value of the response directly as a linear function of explanatory variables, a suitable transformation is modelled. In this case the most suitable transformation is the *logistic* or *logit* function of $\pi = P(y = 1)$ leading to the model

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

The logit of a probability is simply the log of the odds of the response taking the value one.

Logistic Regression

The logit function can take any real value, but the associated probability always lies in the required $[0, 1]$ interval. In a logistic regression model, the parameter β_j associated with explanatory variable x_j is such that $\exp(\beta_j)$ is the odds that the response variable takes the value one when x_j increases by one, conditional on the other explanatory variables remaining constant. The parameters of the logistic regression model (the vector of regression coefficients β) are estimated by maximum likelihood.

Generalized Linear Model (GLM)

Essentially GLMs consist of three main features;

1. An *error distribution* giving the distribution of the response around its mean.
2. A *link function*, g, that shows how the linear function of the explanatory variables is related to the expected value of the response

$$g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q.$$

3. The *variance function* that captures how the variance of the response variable depends on the mean.

Estimation of the parameters in a GLM is usually achieved through a maximum likelihood approach.

Summary

- **Generalized linear models** provide a very powerful and flexible framework for the application of regression models to a variety of non-normal response variables, for example, **logistic regression** to binary responses and **Poisson regression** to count data.