Generalized linear models

Regression Models

Brian Caffo, Jeff Leek, Roger Peng
Johns Hopkins Bloomberg School of Public Health

Linear models

- Linear models are the most useful applied statistical technique. However, they are not without their limitations.
  - Additive response models don't make much sense if the response is discrete, or stricly positive.
  - Additive error models often don't make sense, for example if the outcome has to be positive.
  - Transformations are often hard to interpret.
  - There's value in modeling the data on the scale that it was collected.
  - Particularly interpetable transformations, natural logarithms in specific, aren't applicable for negative or zero values.

Generalized linear models

- Introduced in a 1972 RSSB paper by Nelder and Wedderburn.
- Involves three components
  - An *exponential family* model for the response.
  - A systematic component via a linear predictor.
  - A link function that connects the means of the response to the linear predictor.

Example, linear models

- Assume that $Y_i \sim N(\mu_i, \sigma^2)$ (the Gaussian distribution is an exponential family distribution.)
- Define the linear predictor to be $\eta_i = \sum_{k=1}^p X_{ik} \beta_k$.
- The link function as g so that $g(\mu) = \eta$.
  - For linear models $g(\mu) = \mu$ so that $\mu_i = \eta_i$
- This yields the same likelihood model as our additive error Gaussian linear model

$$Y_i = \sum_{k=1}^p X_{ik} \beta_k + \epsilon_i$$

where $\epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$

Example, logistic regression

- Assume that $Y_i \sim \text{Bernoulli}(\mu_i)$ so that $E[Y_i] = \mu_i$ where $0 \le \mu_i \le 1$.
- Linear predictor $\eta_i = \sum_{k=1}^{p} X_{ik}\beta_k$
- Link function $g(\mu) = \eta = \log\left(\frac{\mu}{1-\mu}\right)$ g is the (natural) log odds, referred to as the **logit**.
- Note then we can invert the logit function as

$$\mu_i = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \quad \text{and} \quad 1 - \mu_i = \frac{1}{1 + \exp(\eta_i)}$$

Thus the likelihood is

$$\prod_{i=1}^{n} \mu_i^{y_i}(1 - \mu_i)^{1-y_i} = \exp\left(\sum_{i=1}^{n} y_i\eta_i\right)\prod_{i=1}^{n}(1 + \eta_i)^{-1}$$

Example, Poisson regression

- Assume that $Y_i \sim \text{Poisson}(\mu_i)$ so that $E[Y_i] = \mu_i$ where $0 \leq \mu_i$
- Linear predictor $\eta_i = \sum_{k=1}^{p} X_{ik} \beta_k$
- Link function $g(\mu) = \eta = \log(\mu)$
- Recall that e^x is the inverse of $\log(x)$ so that

$$\mu_i = e^{\eta_i}$$

Thus, the likelihood is

$$\prod_{i=1}^{n} (y_i!)^{-1} \mu_i^{y_i} e^{-\mu_i} \propto \exp\left(\sum_{i=1}^{n} y_i \eta_i - \sum_{i=1}^{n} \mu_i \right)$$

Some things to note

- In each case, the only way in which the likelihood depends on the data is through

$$\sum_{i=1}^{n} y_i \eta_i = \sum_{i=1}^{n} y_i \sum_{k=1}^{p} X_{ik} \beta_k = \sum_{k=1}^{p} \beta_k \sum_{i=1}^{n} X_{ik} y_i$$

  Thus if we don't need the full data, only $\sum_{i=1}^{n} X_{ik} y_i$. This simplification is a consequence of chosing so-called 'canonical' link functions.

- (This has to be derived). All models acheive their maximum at the root of the so called normal equations

$$0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{Var(Y_i)} W_i$$

  where W_i are the derivative of the inverse of the link function.

About variances

$$0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\mathrm{Var}(Y_i)} W_i$$

- For the linear model $\mathrm{Var}(Y_i) = \sigma^2$ is constant.
- For Bernoulli case $\mathrm{Var}(Y_i) = \mu_i(1 - \mu_i)$
- For the Poisson case $\mathrm{Var}(Y_i) = \mu_i$.
- In the latter cases, it is often relevant to have a more flexible variance model, even if it doesn't correspond to an actual likelihood

$$0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\phi\mu_i(1 - \mu_i)} W_i \quad \text{and} \quad 0 = \sum_{i=1}^{n} \frac{(Y_i - \mu_i)}{\phi\mu_i} W_i$$

- These are called 'quasi-likelihood' normal equations

Odds and ends

- The normal equations have to be solved iteratively. Resulting in $\hat{\beta}_k$ and, if included, $\hat{\phi}$.
- Predicted linear predictor responses can be obtained as $\hat{\eta} = \sum_{k=1}^{p} X_k \hat{\beta}_k$
- Predicted mean responses as $\hat{\mu} = g^{-1}(\hat{\eta})$
- Coefficients are interpretted as

$$g(E[Y|X_k = x_k + 1, X_{\sim k} = x_{\sim k}]) - g(E[Y|X_k = x_k, X_{\sim k} = x_{\sim k}]) = \beta_k$$

  or the change in the link function of the expected response per unit change in X_k holding other regressors constant.
- Variations on Newon/Raphson's algorithm are used to do it.
- Asymptotics are used for inference usually.
- Many of the ideas from linear models can be brought over to GLMs.