Lecture 22

November 5, 2018

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- GLM makes the following assumptions:

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- GLM makes the following assumptions:
- $y_1, \ldots, y_n$ are independent with y_i having the pmf or pdf of the form

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{1}$$

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- GLM makes the following assumptions:
- $y_1, \ldots, y_n$ are independent with y_i having the pmf or pdf of the form

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \qquad (1)$$

- Here θ_i and ϕ_i are parameters.

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- GLM makes the following assumptions:
- $y_1, \ldots, y_n$ are independent with y_i having the pmf or pdf of the form

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{1}$$

- Here θ_i and ϕ_i are parameters.
- θ_i is the canonical parameter and ϕ_i is called the dispersion parameter.

Generalized Linear Models

- We have n observations on a response variable $y_1, \ldots, y_n$ and on each of p explanatory variables x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.
- GLM makes the following assumptions:
- $y_1, \ldots, y_n$ are independent with y_i having the pmf or pdf of the form

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \qquad (1)$$

- Here θ_i and ϕ_i are parameters.
- θ_i is the canonical parameter and ϕ_i is called the dispersion parameter.
- One often assumes that ϕ_i is the same for all i.

- Let $\mu_i = \mathbb{E}(y_i)$. For an increasing function g, we model $g(\mu_i)$ as
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- Let $\mu_i = \mathbb{E}(y_i)$. For an increasing function g, we model $g(\mu_i)$ as
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- Because $\mu_i = b'(\theta_i)$, we can write $\theta_i = (b')^{-1}(\mu_i)$ where $(b')^{-1}$ is the inverse function of b'.

- Let $\mu_i = \mathbb{E}(y_i)$. For an increasing function g, we model $g(\mu_i)$ as
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- Because $\mu_i = b'(\theta_i)$, we can write $\theta_i = (b')^{-1}(\mu_i)$ where $(b')^{-1}$ is the inverse function of b'.
- The link function $g = (b')^{-1}$ is known as the canonical link.

- Let $\mu_i = \mathbb{E}(y_i)$. For an increasing function g, we model $g(\mu_i)$ as
$$g(\mu_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$
- Because $\mu_i = b'(\theta_i)$, we can write $\theta_i = (b')^{-1}(\mu_i)$ where $(b')^{-1}$ is the inverse function of b'.
- The link function $g = (b')^{-1}$ is known as the canonical link.
- The resulting GLM is called the canonical GLM. This is given by
$$(b')^{-1}(\mu_i) = \theta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Example One - Normal Linear Model

- Here y_i has the normal distribution $N(\mu_i, \sigma^2)$.

Example One - Normal Linear Model

- Here y_i has the normal distribution $N(\mu_i, \sigma^2)$.
- The pdf of y_i can be written in the form (1) with $\theta_i = \mu_i$ and $\phi_i = \sigma^2$ and $a(\phi_i) = \phi_i$.

Example One - Normal Linear Model

- Here y_i has the normal distribution $N(\mu_i, \sigma^2)$.
- The pdf of y_i can be written in the form (1) with $\theta_i = \mu_i$ and $\phi_i = \sigma^2$ and $a(\phi_i) = \phi_i$.
- The link function used is the identity link function $g(\mu_i) = \mu_i$. This is the canonical link here.

Example Two - Binary Regression including Logistic and Probit Models

- Here y_i has the Bernoulli distribution with parameter p_i.

Example Two - Binary Regression including Logistic and Probit Models

- Here y_i has the Bernoulli distribution with parameter p_i.
- The pmf of y_i can be written in the form (1) with $\theta_i = \log(p_i/(1-p_i))$ and $\phi_i = 1$ and $b(\theta_i) = \log(1+e^{\theta_i})$.

Example Two - Binary Regression including Logistic and Probit Models

- Here y_i has the Bernoulli distribution with parameter p_i.
- The pmf of y_i can be written in the form (1) with $\theta_i = \log(p_i/(1-p_i))$ and $\phi_i = 1$ and $b(\theta_i) = \log(1+e^{\theta_i})$.
- The canonical GLM is therefore

$$\theta_i = \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Example Two - Binary Regression including Logistic and Probit Models

- Here y_i has the Bernoulli distribution with parameter p_i.
- The pmf of y_i can be written in the form (1) with $\theta_i = \log(p_i/(1-p_i))$ and $\phi_i = 1$ and $b(\theta_i) = \log(1+e^{\theta_i})$.
- The canonical GLM is therefore

$$\theta_i = \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- This is the *Logistic Regression Model*. $p_i/(1-p_i)$ denotes the odds of the event that $y_i = 1$.

Example Two - Binary Regression including Logistic and Probit Models

- Here y_i has the Bernoulli distribution with parameter p_i.
- The pmf of y_i can be written in the form (1) with $\theta_i = \log(p_i/(1-p_i))$ and $\phi_i = 1$ and $b(\theta_i) = \log(1+e^{\theta_i})$.
- The canonical GLM is therefore

$$\theta_i = \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- This is the *Logistic Regression Model*. $p_i/(1-p_i)$ denotes the odds of the event that $y_i = 1$.
- The interpretation of β_j is that it represents the increase in log-odds of the event that $y = 1$ for a unit increase in x_j when all other explanatory variables are held constant.

- In other words, e^{β_j} denotes the factor by which the odds of success (response equal to one) change for a unit increase in x_j (all other explanatory variables remaining unchanged).

- In other words, e^{β_j} denotes the factor by which the odds of success (response equal to one) change for a unit increase in x_j (all other explanatory variables remaining unchanged).
- The function $p \mapsto \log(p/(1-p))$ is the link function in the Logistic Model and is called the logit function.

- In other words, e^{β_j} denotes the factor by which the odds of success (response equal to one) change for a unit increase in x_j (all other explanatory variables remaining unchanged).
- The function $p \mapsto \log(p/(1-p))$ is the link function in the Logistic Model and is called the logit function.
- This is the most popular link function for Bernoulli data.

Probit model

- Another link function is the probit link: $g(x) = \Phi^{-1}(x)$ where Φ is the cdf of the standard normal density.

Probit model

- Another link function is the probit link: $g(x) = \Phi^{-1}(x)$ where Φ is the cdf of the standard normal density.
- This leads to the probit model:

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Fitting GLMs to data

- Suppose we decide to fit a GLM to the data. How then do we estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$?

Fitting GLMs to data

- Suppose we decide to fit a GLM to the data. How then do we estimate the parameters $\beta_0, \beta_1, \ldots, \beta_p$?
- We shall first work this out in the case of the Logistic Regression Model. The general case will be dealt with later.

Fitting the Logistic Regression Model to Data

- How to estimate $\beta_0, \beta_1, \ldots, \beta_p$ from the model:

  $$y_i \sim^{independent} Ber(p_i) \;\text{ where }\; \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for $i = 1, \ldots, n$.

Fitting the Logistic Regression Model to Data

- How to estimate $\beta_0, \beta_1, \ldots, \beta_p$ from the model:

  $$y_i \sim^{independent} Ber(p_i) \;\text{ where }\; \log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for $i = 1, \ldots, n$.
- The data is $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Fitting the Logistic Regression Model to Data

- How to estimate $\beta_0, \beta_1, \dots, \beta_p$ from the model:

  $$y_i \sim^{independent} Ber(p_i) \text{ where } \log \frac{p_i}{1 - p_i} = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$$

  for $i = 1, \dots, n$.
- The data is $y_1, \dots, y_n$ and x_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$.
- The model can alternatively be written as

  $$y_i \sim Ber\left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}\right)$$

  with $y_1, \dots, y_n$ being independent.

- How to estimate $\beta_0, \ldots, \beta_p$?

- How to estimate $\beta_0, \ldots, \beta_p$?
- One simply uses Maximum Likelihood.

- How to estimate $\beta_0, \dots, \beta_p$?
- One simply uses Maximum Likelihood.
- The likelihood of $y_1, \dots, y_n$ is

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with} \quad p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}.$$

- How to estimate $\beta_0, \ldots, \beta_p$?
- One simply uses Maximum Likelihood.
- The likelihood of $y_1, \ldots, y_n$ is

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with} \quad p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}.$$

- This likelihood is simply a function of $\beta_0, \ldots, \beta_p$ and so it can be maximized to yield estimates of $\beta_0, \ldots, \beta_p$.

- How to estimate $\beta_0, \ldots, \beta_p$?
- One simply uses Maximum Likelihood.
- The likelihood of $y_1, \ldots, y_n$ is

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i} \quad \text{with} \quad p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}.$$

- This likelihood is simply a function of $\beta_0, \ldots, \beta_p$ and so it can be maximized to yield estimates of $\beta_0, \ldots, \beta_p$.
- It is easier to work with the log-likelihood.

- The log-likelihood is given by

- The log-likelihood is given by

$$\ell(\beta) \quad = \quad \sum_{i=1}^{n} \left(y_i \log p_i + (1 - y_i) \log(1 - p_i) \right)$$

- The log-likelihood is given by

$$\begin{aligned}
\ell(\beta) &= \textstyle\sum_{i=1}^{n} \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \\
&= \textstyle\sum_{i=1}^{n} \Big[y_i(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}) - \\
&\quad \log(1 + \exp\left(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}\right)) \Big].
\end{aligned}$$

- The log-likelihood is given by

$$
\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{n} \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \\
&= \sum_{i=1}^{n} \Big[y_i(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}) - \\
&\quad \log(1 + \exp\left(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}\right)) \Big].
\end{aligned}
$$

- Unfortunately, one cannot write down the minimizer for $\ell(\beta)$ in closed form. One therefore uses Newton's method.

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.
- Then we can use Taylor's approximation of $\ell(\beta)$ around $\beta^{(m)}$:

$$\tilde{\ell}(\beta)$$
$$:= \ell(\beta^{(m)}) + \nabla\ell(\beta^{(m)})^T(\beta - \beta^{(m)}) + \frac{(\beta - \beta^{(m)})^T H\ell(\beta^{(m)})(\beta - \beta^{(m)})}{2}$$

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.
- Then we can use Taylor's approximation of $\ell(\beta)$ around $\beta^{(m)}$:

$$\tilde{\ell}(\beta)$$
$$:= \ell(\beta^{(m)}) + \nabla\ell(\beta^{(m)})^T(\beta - \beta^{(m)}) + \frac{(\beta-\beta^{(m)})^T H\ell(\beta^{(m)})(\beta-\beta^{(m)})}{2}$$

- So instead of maximizing $\ell(\beta)$ we can maximize $\tilde{\ell}(\beta)$ which has gradient

$$\nabla\tilde{\ell}(\beta) = \nabla\ell(\beta^{(m)}) + H\ell(\beta^{(m)})(\beta - \beta^{(m)})$$

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.
- Then we can use Taylor's approximation of $\ell(\beta)$ around $\beta^{(m)}$:

$$\tilde{\ell}(\beta)$$
$$:= \ell(\beta^{(m)}) + \nabla\ell(\beta^{(m)})^T(\beta - \beta^{(m)}) + \frac{(\beta-\beta^{(m)})^T H\ell(\beta^{(m)})(\beta-\beta^{(m)})}{2}$$

- So instead of maximizing $\ell(\beta)$ we can maximize $\tilde{\ell}(\beta)$ which has gradient

$$\nabla\tilde{\ell}(\beta) = \nabla\ell(\beta^{(m)}) + H\ell(\beta^{(m)})(\beta - \beta^{(m)})$$

- And so $\nabla\tilde{\ell}(\beta) = 0$ if only if

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.
- Then we can use Taylor's approximation of $\ell(\beta)$ around $\beta^{(m)}$:

$$\tilde{\ell}(\beta) := \ell(\beta^{(m)}) + \nabla\ell(\beta^{(m)})^T(\beta - \beta^{(m)}) + \frac{(\beta-\beta^{(m)})^T H\ell(\beta^{(m)})(\beta-\beta^{(m)})}{2}$$

- So instead of maximizing $\ell(\beta)$ we can maximize $\tilde{\ell}(\beta)$ which has gradient

$$\nabla\tilde{\ell}(\beta) = \nabla\ell(\beta^{(m)}) + H\ell(\beta^{(m)})(\beta - \beta^{(m)})$$

- And so $\nabla\tilde{\ell}(\beta) = 0$ if only if

$$\beta = H\ell(\beta^{(m)})^{-1}\left(H\ell(\beta^{(m)})\beta^{(m)} - \nabla\ell(\beta^{(m)})\right)$$

Newton's method

- Suppose we have a candidate solution $\beta^{(m)}$.
- Then we can use Taylor's approximation of $\ell(\beta)$ around $\beta^{(m)}$:

$$\tilde{\ell}(\beta) := \ell(\beta^{(m)}) + \nabla\ell(\beta^{(m)})^T(\beta - \beta^{(m)}) + \frac{(\beta-\beta^{(m)})^T H\ell(\beta^{(m)})(\beta-\beta^{(m)})}{2}$$

- So instead of maximizing $\ell(\beta)$ we can maximize $\tilde{\ell}(\beta)$ which has gradient

$$\nabla\tilde{\ell}(\beta) = \nabla\ell(\beta^{(m)}) + H\ell(\beta^{(m)})(\beta - \beta^{(m)})$$

- And so $\nabla\tilde{\ell}(\beta) = 0$ if only if

$$\begin{aligned}\beta &= H\ell(\beta^{(m)})^{-1}\left(H\ell(\beta^{(m)})\beta^{(m)} - \nabla\ell(\beta^{(m)})\right)\\ &= \beta^{(m)} - H\ell(\beta^{(m)})^{-1}\nabla\ell(\beta^{(m)}).\end{aligned}$$

- Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla\ell(\beta^{(m)}) \tag{2}$$

where $\nabla\ell(\beta)$ and $H\ell(\beta)$ denote the gradient and Hessian of the function $\ell(\beta)$ respectively:

- Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla\ell(\beta^{(m)}) \tag{2}$$

  where $\nabla\ell(\beta)$ and $H\ell(\beta)$ denote the gradient and Hessian of the function $\ell(\beta)$ respectively:
- $\nabla\ell(\beta) := (\partial\ell(\beta)/\partial\beta_0, \ldots, \partial\ell(\beta)/\partial\ell(\beta_p))^T$ and $H\ell(\beta)$ is the $(p+1) \times (p+1)$ matrix whose entries are second order derivatives of $\ell(\beta)$.

- Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla\ell(\beta^{(m)}) \tag{2}$$

  where $\nabla\ell(\beta)$ and $H\ell(\beta)$ denote the gradient and Hessian of the function $\ell(\beta)$ respectively:
- $\nabla\ell(\beta) := (\partial\ell(\beta)/\partial\beta_0, \ldots, \partial\ell(\beta)/\partial\ell(\beta_p))^T$ and $H\ell(\beta)$ is the $(p+1) \times (p+1)$ matrix whose entries are second order derivatives of $\ell(\beta)$.
- For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.

- It is quite easy to write down $\nabla\ell(\beta)$ and $H\ell(\beta)$.

- It is quite easy to write down $\nabla\ell(\beta)$ and $H\ell(\beta)$.
- Check that

$$\nabla\ell(\beta) = \sum_{i=1}^{n}(y_i - p_i)(1, x_{i1}, \ldots, x_{ip})^T$$

- It is quite easy to write down $\nabla\ell(\beta)$ and $H\ell(\beta)$.
- Check that

$$\nabla\ell(\beta) = \sum_{i=1}^{n}(y_i - p_i)(1, x_{i1}, \ldots, x_{ip})^T$$

- and

$$H\ell(\beta) = -\sum_{i=1}^{n} p_i(1 - p_i)(1, x_{i1}, \ldots, x_{ip})^T(1, x_{i1}, \ldots, x_{ip}).$$

- It is quite easy to write down $\nabla \ell(\beta)$ and $H\ell(\beta)$.
- Check that

$$\nabla \ell(\beta) = \sum_{i=1}^{n} (y_i - p_i)(1, x_{i1}, \dots, x_{ip})^T$$

- and

$$H\ell(\beta) = -\sum_{i=1}^{n} p_i(1 - p_i)(1, x_{i1}, \dots, x_{ip})^T (1, x_{i1}, \dots, x_{ip}).$$

- These expression look much nicer in matrix notation. As before, Y denotes the vector of response values $(y_1, \dots, y_n)^T$ and X denotes the $n \times (p+1)$ matrix whose first column is 1 and the remaining columns correspond to the explanatory variables.

- Let β denote the vector $(\beta_0, \ldots, \beta_p)^T$. Let p denote the vector $(p_1, \ldots, p_n)^T$ and let W denote the $n \times n$ diagonal matrix whose ith diagonal element is $p_i(1 - p_i)$.

- Let β denote the vector $(\beta_0, \ldots, \beta_p)^T$. Let p denote the vector $(p_1, \ldots, p_n)^T$ and let W denote the $n \times n$ diagonal matrix whose ith diagonal element is $p_i(1 - p_i)$.
- Check that

$$\nabla \ell(\beta) = X^T(Y - p)$$

- Let β denote the vector $(\beta_0, \ldots, \beta_p)^T$. Let p denote the vector $(p_1, \ldots, p_n)^T$ and let W denote the $n \times n$ diagonal matrix whose ith diagonal element is $p_i(1 - p_i)$.
- Check that

$$\nabla \ell(\beta) = X^T(Y - p)$$

- and

$$H\ell(\beta) = -X^T W X.$$

- Let β denote the vector $(\beta_0, \ldots, \beta_p)^T$. Let p denote the vector $(p_1, \ldots, p_n)^T$ and let W denote the $n \times n$ diagonal matrix whose ith diagonal element is $p_i(1 - p_i)$.
- Check that

$$\nabla \ell(\beta) = X^T(Y - p)$$

- and

$$H\ell(\beta) = -X^T WX.$$

- The iterative scheme (2) therefore becomes

$$\beta^{(m+1)} = \beta^{(m)} + (X^T WX)^{-1} X^T (Y - p).$$

- This can be rewritten as

$$\beta^{(m+1)} = (X^T W X)^{-1} X^T W Z \tag{3}$$

where

$$Z = X\beta^{(m)} + W^{-1}(Y - p).$$