Generalized Linear Models 1

October 28, 2018

GLMs

- We have so far studied 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 $j = 1, \ldots, p$.

GLMs

- We have so far studied 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 $j = 1, \ldots, p$.
- The linear model that we have seen models

$$\mu_i := \mathbb{E} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

GLMs

- We have so far studied 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 $j = 1, \ldots, p$.
- The linear model that we have seen models

$$\mu_i := \mathbb{E} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- What this model implies is that when there is a unit increase in the explanatory variable x_j, the mean of the response variable changes by the amount β_j.

GLMs

- We have so far studied 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 $j = 1, \ldots, p$.
- The linear model that we have seen models

$$\mu_i := \mathbb{E} y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

- What this model implies is that when there is a unit increase in the explanatory variable x_j, the mean of the response variable changes by the amount β_j.
- This may not always be a reasonable assumption.

GLMs

- For example, if the response y_i is a binary variable, then its mean μ_i is a probability which is always constrained to stay between 0 and 1.

GLMs

- For example, if the response y_i is a binary variable, then its mean μ_i is a probability which is always constrained to stay between 0 and 1.
- Therefore, the amount by which μ_i changes per unit change in x_j would now depend on the value of μ_i (for example, the change when $\mu_i = 0.9$ may not be the same as when $\mu_i = 0.5$).

GLMs

- For example, if the response y_i is a binary variable, then its mean μ_i is a probability which is always constrained to stay between 0 and 1.
- Therefore, the amount by which μ_i changes per unit change in x_j would now depend on the value of μ_i (for example, the change when $\mu_i = 0.9$ may not be the same as when $\mu_i = 0.5$).
- Therefore, modeling μ_i as a linear combination of $x_1, \dots, x_p$ may not be the best idea always.

GLMs

- For example, if the response y_i is a binary variable, then its mean μ_i is a probability which is always constrained to stay between 0 and 1.
- Therefore, the amount by which μ_i changes per unit change in x_j would now depend on the value of μ_i (for example, the change when $\mu_i = 0.9$ may not be the same as when $\mu_i = 0.5$).
- Therefore, modeling μ_i as a linear combination of $x_1, \dots, x_p$ may not be the best idea always.
- A more general model might be

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{1}$$

for a function g that is not necessarily the identity function.

- Another feature of the linear model that people do not always like is that some aspects of the theory are tied to the normal distribution.

- Another feature of the linear model that people do not always like is that some aspects of the theory are tied to the normal distribution.
- Indeed, most of the results on hypothesis testing rely on the assumption of normality.

- Another feature of the linear model that people do not always like is that some aspects of the theory are tied to the normal distribution.
- Indeed, most of the results on hypothesis testing rely on the assumption of normality.
- The assumption that $y_1, \ldots, y_n$ are normal may not always be appropriate. Examples arise when $y_1, \ldots, y_n$ are binary or when they represent counts.

- Another feature of the linear model that people do not always like is that some aspects of the theory are tied to the normal distribution.
- Indeed, most of the results on hypothesis testing rely on the assumption of normality.
- The assumption that $y_1, \dots, y_n$ are normal may not always be appropriate. Examples arise when $y_1, \dots, y_n$ are binary or when they represent counts.
- It might therefore be nice to generalize the theory of linear models to include these other distributional assumptions for the response values.

- Another feature of the linear model that people do not always like is that some aspects of the theory are tied to the normal distribution.
- Indeed, most of the results on hypothesis testing rely on the assumption of normality.
- The assumption that $y_1, \ldots, y_n$ are normal may not always be appropriate. Examples arise when $y_1, \ldots, y_n$ are binary or when they represent counts.
- It might therefore be nice to generalize the theory of linear models to include these other distributional assumptions for the response values.
- Generalized Linear Models (GLM) generalize linear models by including both of the above features.

- They allow more general distributional assumptions for $y_1, \ldots, y_n$ and they also allow (1).

Distributional Assumptions in GLMs

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs).

Distributional Assumptions in GLMs

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs).
- It is assumed that $y_1, \ldots, y_n$ are independent.

Distributional Assumptions in GLMs

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs).
- It is assumed that $y_1, \ldots, y_n$ are independent.
- We also assume that the pmf or pdf of y_i can be modelled by two parameters θ_i and ϕ_i and can be written as

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{2}$$

Distributional Assumptions in GLMs

- In GLMs, the response variables $y_1, \ldots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs).
- It is assumed that $y_1, \ldots, y_n$ are independent.
- We also assume that the pmf or pdf of y_i can be modelled by two parameters θ_i and ϕ_i and can be written as

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{2}$$

- θ_i is the main parameter (also called the canonical parameter).

Distributional Assumptions in GLMs

- In GLMs, the response variables $y_1, \dots, y_n$ can be either discrete (have pmfs) or continuous (have pdfs).
- It is assumed that $y_1, \dots, y_n$ are independent.
- We also assume that the pmf or pdf of y_i can be modelled by two parameters θ_i and ϕ_i and can be written as

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \tag{2}$$

- θ_i is the main parameter (also called the canonical parameter).
- ϕ_i is called the dispersion parameter and one often assumes that ϕ_i is the same for all i.

- The function $b(\theta_i)$ is called the cumulant function.

- The function $b(\theta_i)$ is called the cumulant function.
- This distributional form includes the normal density assumption used in the classical linear models. In classical linear models, we assume that $y_i \sim N(\mu_i, \sigma^2)$.

- The function $b(\theta_i)$ is called the cumulant function.
- This distributional form includes the normal density assumption used in the classical linear models. In classical linear models, we assume that $y_i \sim N(\mu_i, \sigma^2)$.
- The density of y_i can then be written as

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma^2} \right)$$

- The function $b(\theta_i)$ is called the cumulant function.
- This distributional form includes the normal density assumption used in the classical linear models. In classical linear models, we assume that $y_i \sim N(\mu_i, \sigma^2)$.
- The density of y_i can then be written as

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma^2}\right)$$

- This can be rewritten as

$$f(x) := \frac{\exp(-y^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} \exp\left(\frac{y\mu_i - \mu_i^2/2}{\sigma^2}\right).$$

- The function $b(\theta_i)$ is called the cumulant function.
- This distributional form includes the normal density assumption used in the classical linear models. In classical linear models, we assume that $y_i \sim N(\mu_i, \sigma^2)$.
- The density of y_i can then be written as

$$f(x) := \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-(x-\mu_i)^2}{2\sigma^2}\right)$$

- This can be rewritten as

$$f(x) := \frac{\exp(-y^2/(2\sigma^2))}{\sqrt{2\pi}\sigma} \exp\left(\frac{y\mu_i - \mu_i^2/2}{\sigma^2}\right).$$

- This is clearly in the form (2) with $\theta_i = \mu_i, \phi_i = \sigma^2, a(\phi_i) = \phi_i$ and $b(\theta_i) = \theta_i^2/2$.

- Note that ϕ_i here does not depend on i.

- Note that ϕ_i here does not depend on i.
- Other distributions that are of the form (2) are:

- Note that ϕ_i here does not depend on i.
- Other distributions that are of the form (2) are:
  1. **Bernoulli**: Here y_i takes the value 0 with probablity $1 - p_i$ and 1 with probability p_i.

- Note that ϕ_i here does not depend on i.
- Other distributions that are of the form (2) are:
  1. **Bernoulli**: Here y_i takes the value 0 with probablity $1 - p_i$ and 1 with probability p_i.
  2. The pmf of y_i is

$$f(x) = p_i^x (1 - p_i)^{1-x} = \exp\left(x \log \frac{p_i}{1 - p_i} + \log(1 - p_i) \right).$$

- Note that ϕ_i here does not depend on i.
- Other distributions that are of the form (2) are:
  1. **Bernoulli**: Here y_i takes the value 0 with probablity $1 - p_i$ and 1 with probability p_i.
  2. The pmf of y_i is

$$f(x) = p_i^x (1 - p_i)^{1-x} = \exp\left(x \log \frac{p_i}{1 - p_i} + \log(1 - p_i) \right).$$

  3. If we take $\theta_i := \log(p_i/(1 - p_i))$ and $b(\theta_i) = \log(1 + e^{\theta_i})$ and $\phi_i = 1$, then this is in the form (2).

Binomial

- **Binomial**: Suppose $n_i y_i \sim Bin(n_i, p_i)$ i.e., y_i denotes the proportion of successes in n_i tosses of a coin with probability of success p_i.

Binomial

- **Binomial**: Suppose $n_i y_i \sim Bin(n_i, p_i)$ i.e., y_i denotes the proportion of successes in n_i tosses of a coin with probability of success p_i.
- Check then that the pmf of y_i is

$$f(x) = \binom{n_i}{n_i x} \exp\left(\frac{x \log(p_i/(1-p_i)) + \log(1-p_i)}{1/n_i}\right).$$

Binomial

- **Binomial**: Suppose $n_i y_i \sim Bin(n_i, p_i)$ i.e., y_i denotes the proportion of successes in n_i tosses of a coin with probability of success p_i.
- Check then that the pmf of y_i is

$$f(x) = \binom{n_i}{n_i x} \exp\left(\frac{x \log(p_i/(1-p_i)) + \log(1-p_i)}{1/n_i}\right).$$

- This is of the form (2) with $\theta_i = \log(p_i/(1-p_i))$ and $\phi_i = 1/n_i$ and $b(\theta_i) = \log(1 + e^{\theta_i})$.

Poisson

- **Poisson**: Suppose $y_i \sim Poi(\lambda_i)$.

Poisson

- **Poisson**: Suppose $y_i \sim Poi(\lambda_i)$.
- Then its pmf is

$$f(x) := e^{-\lambda_i} \frac{\lambda_i^x}{x!} = \frac{1}{x!} \exp\left(x \log \lambda_i - \lambda_i\right)$$

which is again of the form (2) with $\theta_i = \log \lambda_i$ and $b(\theta_i) = e^{\theta_i}$ and $\phi_i = 1$.

Poisson

- **Poisson**: Suppose $y_i \sim Poi(\lambda_i)$.
- Then its pmf is

  $$f(x) := e^{-\lambda_i} \frac{\lambda_i^x}{x!} = \frac{1}{x!} \exp\left(x \log \lambda_i - \lambda_i\right)$$

  which is again of the form (2) with $\theta_i = \log \lambda_i$ and $b(\theta_i) = e^{\theta_i}$ and $\phi_i = 1$.
- There are other examples too such as the Gamma distribution but we will mainly deal with the ones above.

- The mean and variance of y_i when y_i has the pmf or pdf (2) can be easily computed by a simple trick.

- The mean and variance of y_i when y_i has the pmf or pdf (2) can be easily computed by a simple trick.
- We will illustrate this below in the case when y_i has a pmf; the case of pdf is exactly identical (just replace sums by integrals). Because $f(x; \theta_i, \phi_i)$ is a density, we have

$$\sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) = 1$$

- The mean and variance of y_i when y_i has the pmf or pdf (2) can be easily computed by a simple trick.
- We will illustrate this below in the case when y_i has a pmf; the case of pdf is exactly identical (just replace sums by integrals). Because $f(x; \theta_i, \phi_i)$ is a density, we have

$$\sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) = 1$$

- Differentiating both sides with respect to θ_i, we get

$$\sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) \frac{x - b'(\theta_i)}{a(\phi_i)} = 0 \qquad (3)$$

$$f(x; \theta_i, \phi_i) := h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \qquad (2)$$

- This means that

$$\sum_x xh(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) =$$
$$b'(\theta_i) \sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right)$$

- This means that

$$\sum_x x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) =$$
$$b'(\theta_i) \sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right)$$

- The left hand side above is simply $\mathbb{E}(y_i)$ and the right hand side is just $b'(\theta_i)$.

Because $f(x; \theta_i, \phi_i)$ is a density, we have

$$\sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) = 1$$

- This means that

$$\sum_x xh(x,\phi_i)\exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) = \\ b'(\theta_i)\sum_x h(x,\phi_i)\exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right)$$

- The left hand side above is simply $\mathbb{E}(y_i)$ and the right hand side is just $b'(\theta_i)$.
- We therefore have

$$\mu_i := \mathbb{E}y_i = b'(\theta_i).$$

- This means that

$$\sum_x x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) =$$
$$b'(\theta_i) \sum_x h(x, \phi_i) \exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right)$$

- The left hand side above is simply $\mathbb{E}(y_i)$ and the right hand side is just $b'(\theta_i)$.
- We therefore have

$$\mu_i := \mathbb{E} y_i = b'(\theta_i).$$

- This can be rewritten as $\theta_i = (b')^{-1}(\mu_i)$ where by $(b')^{-1}$ we mean the inverse function of b'.

- This means that

$$\sum_x xh(x,\phi_i)\exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right) = b'(\theta_i)\sum_x h(x,\phi_i)\exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right)$$

because the right hand side is summed to 1

- The left hand side above is simply $\mathbb{E}(y_i)$ and the right hand side is just $b'(\theta_i)$.
- We therefore have

$$f(x;\theta_i,\phi_i) := h(x,\phi_i)\exp\left(\frac{x\theta_i - b(\theta_i)}{a(\phi_i)}\right). \quad (2)$$

$$\mu_i := \mathbb{E}y_i = b'(\theta_i).$$

- This can be rewritten as $\theta_i = (b')^{-1}(\mu_i)$ where by $(b')^{-1}$ we mean the inverse function of b'.
- Differentiating (3) again with respect to θ_i, it is easy to show that

$$var(y_i) = b''(\theta_i)a(\phi_i).$$

Generalized Linear Models

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (2). $(b')^{-1}$ we mean the inverse function of b'.

Generalized Linear Models

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (2). $(b')^{-1}$ we mean the inverse function of b'.
- We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

for an increasing function g.

Generalized Linear Models

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (2). $(b')^{-1}$ we mean the inverse function of b'.
- We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

for an increasing function g.
- This g is called the *link function*. In classical linear models, $g(\mu_i) = \mu_i$ which means that we have the identity link.

Generalized Linear Models

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (2). $(b')^{-1}$ we mean the inverse function of b'.
- We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for an increasing function g.
- This g is called the *link function*. In classical linear models, $g(\mu_i) = \mu_i$ which means that we have the identity link.
- The link function $g = (b')^{-1}$ is called the *canonical link function*.

Generalized Linear Models

- In GLM, we assume that $y_1, \ldots, y_n$ are independent with pmf or pdf of the form (2). $(b')^{-1}$ we mean the inverse function of b'.
- We then write

$$g(\mu_i) := \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

  for an increasing function g.
- This g is called the *link function*. In classical linear models, $g(\mu_i) = \mu_i$ which means that we have the identity link.
- The link function $g = (b')^{-1}$ is called the *canonical link function*.
- Recall that $(b')^{-1}(\mu_i) = \theta_i$. Thus GLM with the canonical link function models the canonical parameter θ_i as a linear function of the explanatory variables.