Introduction to Generalized Linear Models

Generalized Linear Models

Three components:

- Response variable y
- Explanatory variables $x_1, x_2, x_3, \ldots$
- Link Function g

1. Response variable y

- a.k.a. random component
- Assume independent observations $y_1, \ldots, y_n$ from a particular distribution
  - Model: $\mu_X = E(Y_X)$
  - i.e. how response depends on explanatory variables

2. Explanatory variables (the x's)

- a.k.a. systematic component
- "covariates," "explanatory variables," "features"
- Linear combination of your covariates: $x_1, \ldots, x_p$

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

2. The link function

- The relationship between your response variable, y, and the linear combination of your covariates: x_1, …, x_p

2. The link function

- So if you have:
    1. Response variable Y:

       Model: $\mu_X = E(Y_X)$

    2. Explanatory variables $x_1, \ldots, x_p$:

       $\eta_X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

2. The link function

- So if you have:
  1. Response variable Y:

     Model: $\mu_X = E(Y_X)$
  2. Explanatory variables $x_1, \ldots, x_p$:

     $\eta_X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

- The link function g connects $E(Y_X)$ *to* $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

2. The link function

- So if you have:
  1. Response variable Y:

     Model: $\mu_X = E(Y_X)$
  2. Explanatory variables $x_1, \ldots, x_p$:

     $\eta_X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

- The link function g connects $E(Y_X)$ *to* $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- So the link function is:

$$g(\mu_X) = \eta_X$$

2. The link function

- So if you have:
  1. Response variable Y:

     Model: $\mu_X = E(Y_X)$
  2. Explanatory variables $x_1, \ldots, x_p$:

     $\eta_X = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$

- The link function g connects $E(Y_X)$ *to* $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$
- So the link function is:

$$g(\mu_X) = \eta_X$$

$$g(E(Y_X)) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLM for Linear Regression

GLM for Linear Regression

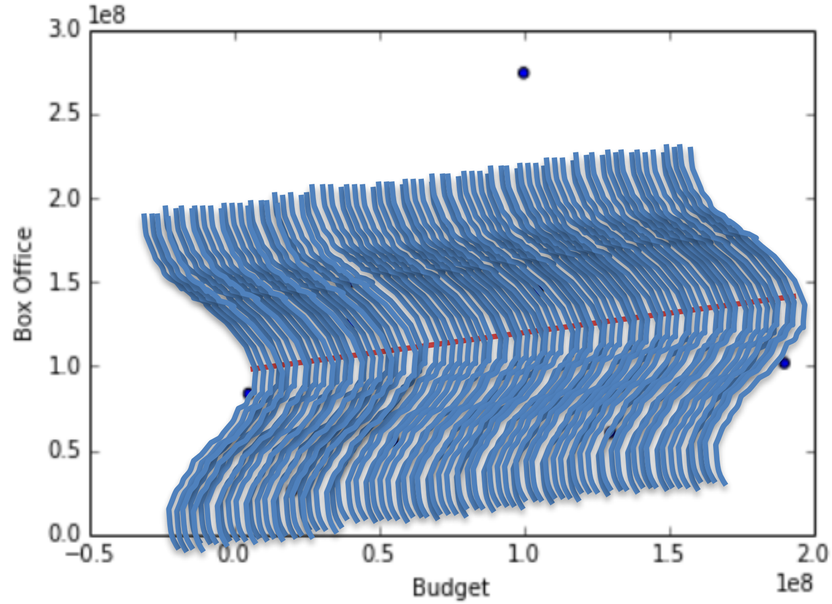
1. Response variable, Y
2. The covariates: $x_1, \ldots, x_p$
3. The link function between the Y and the x's

GLM for Linear Regression

1. Response variable

Y is normally distributed (conditioned on the covariates) with $E(Y_X) = \mu_X$ and constant variance $E(Y) = \sigma^2$:

→ $\varepsilon \sim N(0, \sigma^2)$
→ $Y \sim N(\mu_X, \sigma^2)$



GLM for Linear Regression

2. The covariates: $x_1, \ldots, x_p$

The covariates produce a linear predictor η given by:

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLM for Linear Regression

3. Link function g:

   How do we connect $E(Y_X)$ to $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$?

GLM for Linear Regression

3. Link function g:

   How do we connect $E(Y_X)$ to $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$?

$$E(Y_X) \xrightarrow{g} \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

GLM for Linear Regression

3. Link function g:

   How do we connect $E(Y_X)$ to $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$?

$$E(Y_X) \xrightarrow{g} \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

   With the identity function: $g(x) = x$

$$g(E(Y_X)) = \eta$$

$$E(Y_X) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

In this formulation of classical linear models in the GLM framework, the random component Y_X has a normal distribution, and the link function is the identify function.

GLM allows for 2 extensions of OLS

1. The distribution of Y can come from any exponential family distribution
2. The link function g can be any (monotonic differentiable) function.

Note: the same Maximum Likelihood fitting procedure applies to all GLMs.

GLMs for binary data

GLMs for binary data

1. What is the response?

GLMs for binary data

1. What is the response?

Y is 1 or 0.

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

GLMs for binary data

1. What is the response?

Y is 1 or 0.

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

This is a Bernoulli trial.

GLMs for binary data

1. What is the response?

Y is 1 or 0.

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

This is a Bernoulli trial.

Equivalently, Y follows a Binomial distribution:

$$Y \sim Binomial\ (n = 1,\ P_{success} = \pi)$$

GLMs for binary data

1. What is the response?

Y is 1 or 0.

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

This is a Bernoulli trial.

Equivalently, Y follows a Binomial distribution:

$$Y \sim Binomial\,(n = 1, P_{success} = \pi)$$

$$E(Y) = \pi$$
$$Var\,(Y) = \pi\,(1 - \pi)$$

GLMs for binary data

1. What is the response?

Y is 1 or 0.

$$P(Y = 1) = \pi$$
$$P(Y = 0) = 1 - \pi$$

This is a Bernoulli trial.

Equivalently, Y follows a Binomial distribution:

$$Y \sim Binomial\ (n = 1, P_{success} = \pi)$$

$E(Y) = \pi$

$Var\ (Y) = \pi\ (1 - \pi)$

Notice: $Var\ (Y)$ changes as π changes!

GLMs for binary data

2. What is the systematic component (the x's part)?

GLMs for binary data

2. What is the systematic component (the x's part)?

Once again, it's just a linear combination of the covariates $x_1, \ldots, x_p$

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0$ or 1) to the linear combination of the covariates $x_1, \ldots, x_p$:

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0$ or 1) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$E(Y_X) \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0 \text{ or } 1$) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$E(Y_X) \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The input of this function is

$$E(Y_X) = \mu_X = \pi \; ; 0 \leq \pi \leq 1$$

The output of this function is

$$\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0$ or 1) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$\pi_X \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0 \text{ or } 1$) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$\pi_X \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$g(\pi_X) = \log\left(\frac{\pi_X}{1 - \pi_X}\right)$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0$ or 1) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$\pi_X \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$g(\pi_X) = log(\frac{\pi_X}{1-\pi_X})$$

the logit function

$$log(\frac{\pi_X}{1-\pi_X}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

GLMs for binary data

3. What is the link function?

i.e.: the function g that connects Y (where $Y = 0$ *or* 1) to the linear combination of the covariates $x_1, \ldots, x_p$:

$$\pi_X \xrightarrow{g} \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

$$g(\pi_X) = \log\left(\frac{\pi_X}{1-\pi_X}\right)$$

the logit function

$$\log\left(\frac{\pi_X}{1-\pi_X}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

The link function for logistic regression is the logit function.

GLMs for everything!