Advanced Section #5: Generalized Linear Models: Logistic Regression and Beyond

Nick Stern

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader, and Chris Tanner

Outline

1. Motivation
    - Limitations of linear regression
2. Anatomy
    - Exponential Dispersion Family (EDF)
    - Link function
3. Maximum Likelihood Estimation for GLM’s
    - Fischer Scoring

Motivation

Motivation

Linear regression framework:

$$y_i = x_i^T \beta + \epsilon_i$$

Assumptions:

1. **Linearity:** Linear relationship between expected value and predictors
2. **Normality:** Residuals are normally distributed about expected value
3. **Homoskedasticity:** Residuals have constant variance σ^2
4. **Independence:** Observations are independent of one another

Motivation

Expressed mathematically…

- Linearity

$$\mathbb{E}[y_i] = x_i^T \beta$$

- Normality

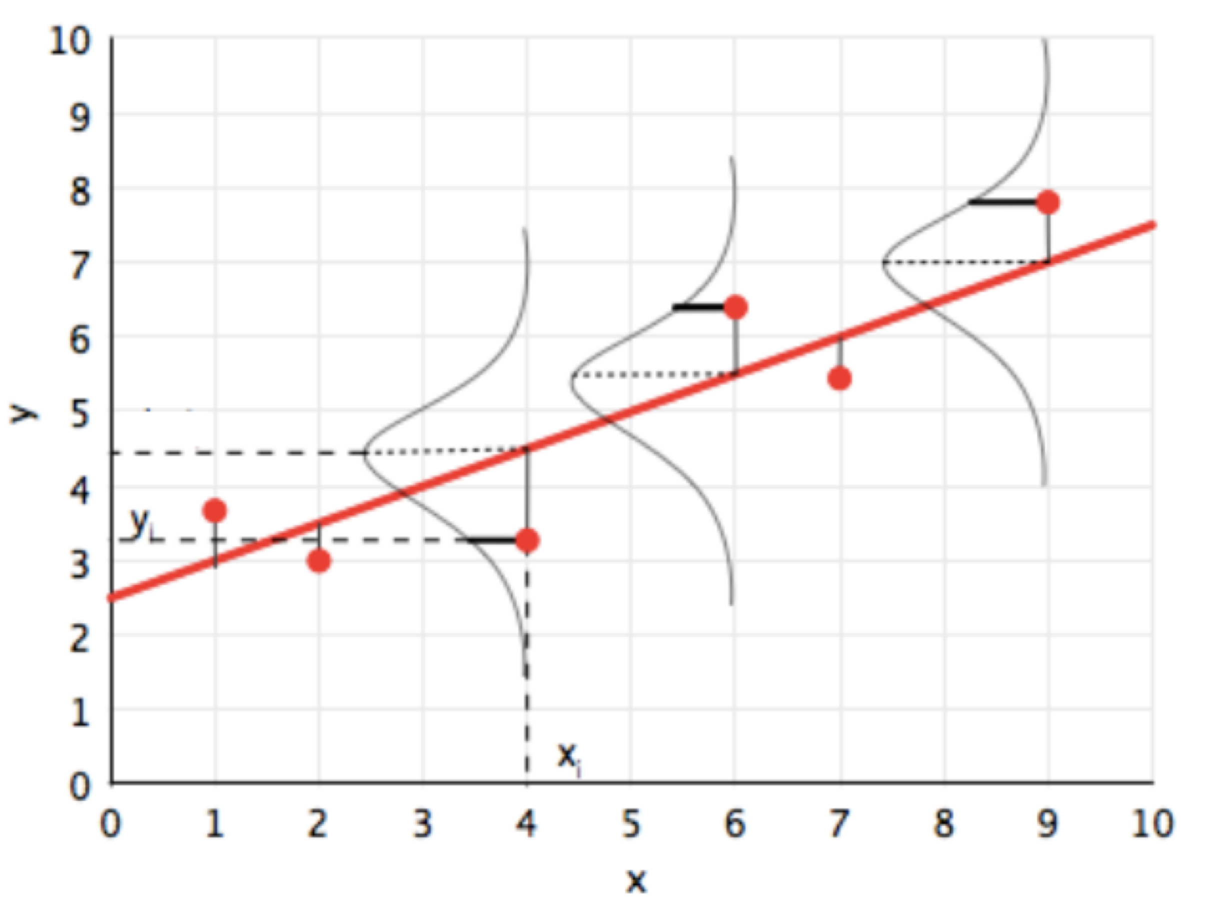
$$y_i \sim \mathcal{N}(x_i^T \beta, \sigma^2)$$

- Homoskedasticity

$$\sigma^2 \text{ (instead of) } \sigma_i^2$$

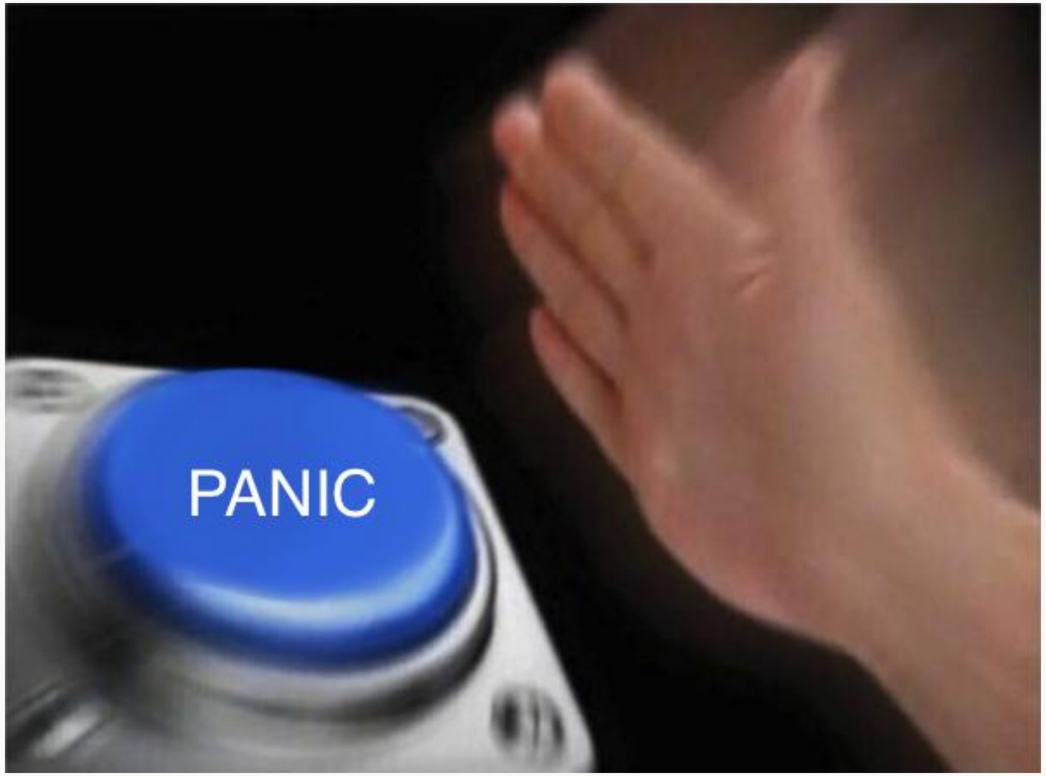
- Independence

$$p(y_i | y_j) = p(y_i) \text{ for } i \neq j$$



Motivation

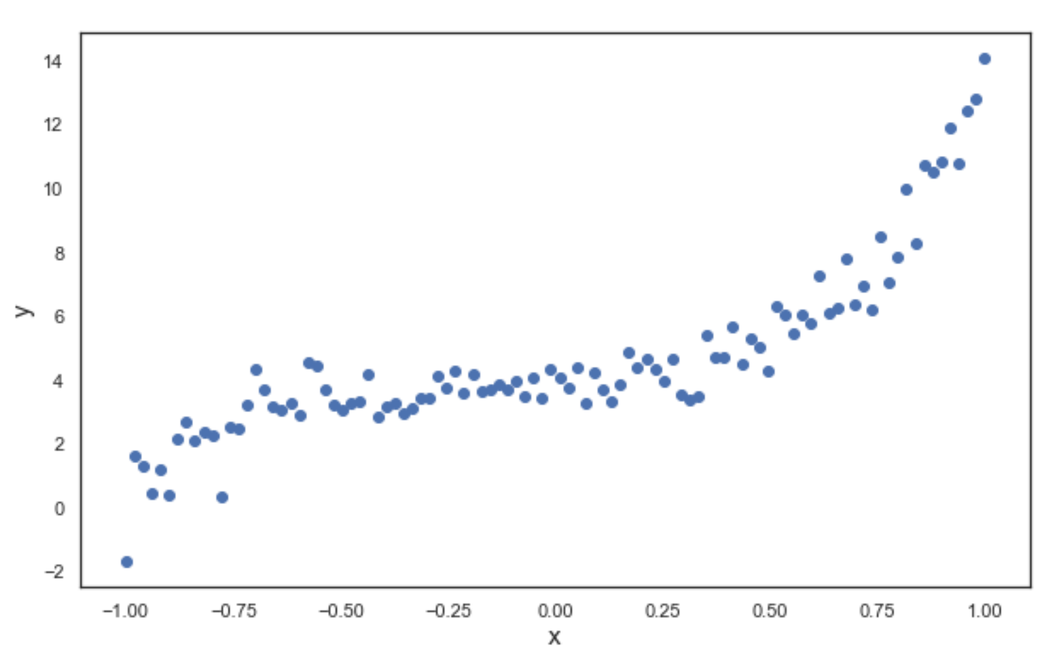
What happens when our assumptions break down?



Motivation

We have options within the framework of linear regression

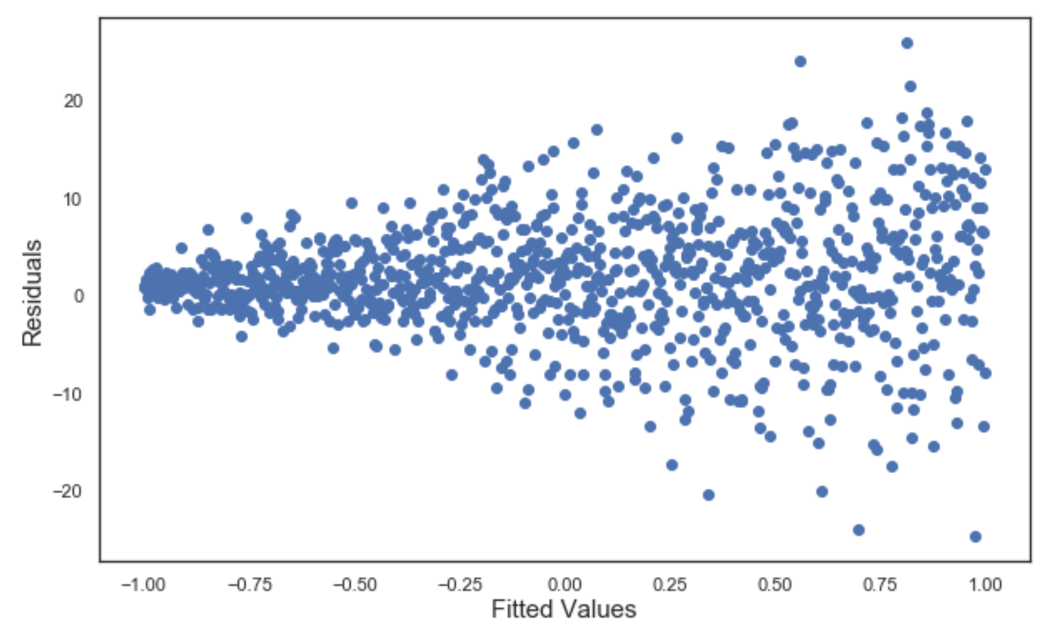
Nonlinearity



Transform X or Y

(Polynomial Regression)

Heteroskedasticity



Weight observations

(WLS Regression)

Motivation

But assuming Normality can be pretty limiting...

Consider modeling the following random variables:

- Whether a coin flip is heads or tails (Bernoulli)
- Counts of species in a given area (Poisson)
- Time between stochastic events that occur w/ constant rate (gamma)
- Vote counts for multiple candidates in a poll (multinomial)

Motivation

We can extend the framework for linear regression.

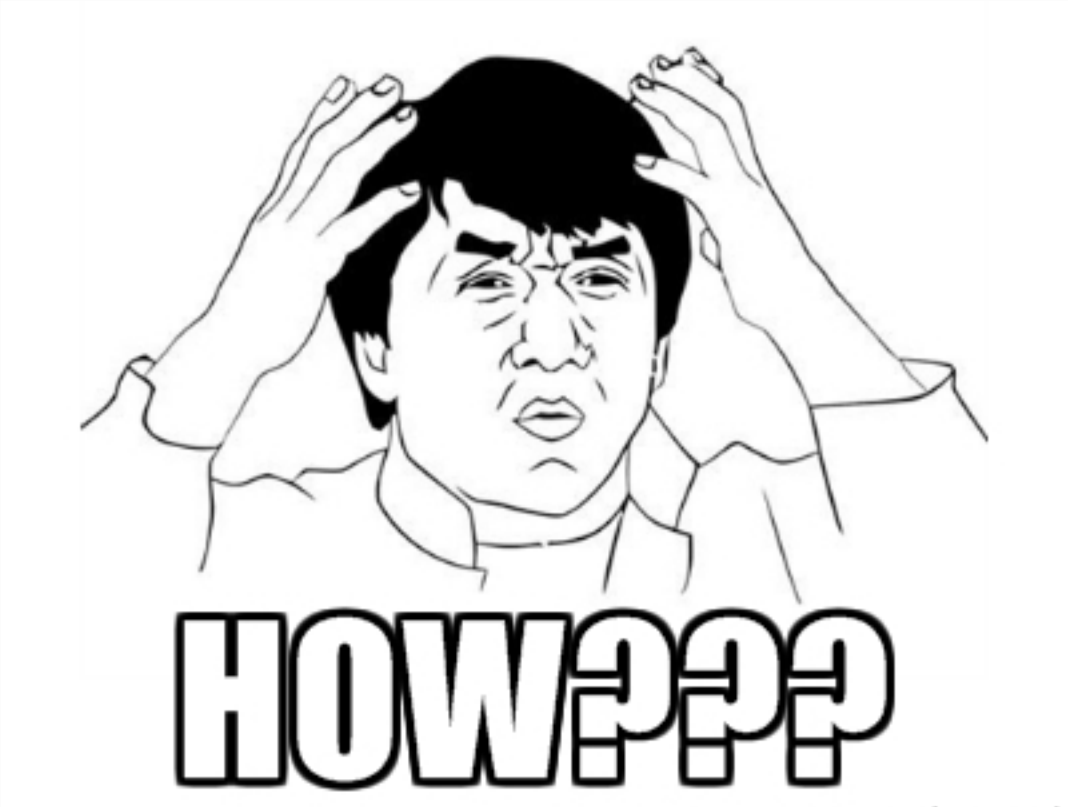
Enter:

Generalized Linear Models

Relaxes:

- Normality assumption
- Homoskedasticity assumption

Motivation



Anatomy

Anatomy

Two adjustments must be made to turn LM into GLM

1. Assume response variable comes from a family of distributions called the **exponential dispersion family (EDF)**.

2. The relationship between expected value and predictors is expressed through a **link function**.

Anatomy – EDF Family

The EDF family contains: Normal, Poisson, gamma, and more!

The probability density function looks like this:

$$f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right)$$

Where

θ - “canonical parameter”

ϕ - “dispersion parameter”

$b(\theta)$ - “cumulant function”

$c(y, \phi)$ - “normalization factor”

Anatomy – EDF Family

Example: representing Bernoulli distribution in EDF form.

PDF of a Bernoulli random variable:

$$f(y_i | p_i) = p_i^{y_i}(1 - p_i)^{1 - y_i}$$

Taking the log and then exponentiating (to cancel each other out) gives:

$$f(y_i | p_i) = \exp(y_i \log(p_i) + (1 - y_i)\log(1 - p_i))$$

Rearranging terms…

$$f(y_i | p_i) = \exp\left(y_i \log\left(\frac{p_i}{1 - p_i}\right) + \log(1 - p_i)\right)$$

Anatomy – EDF Family

Comparing:

$$f(y_i|\, p_i) \;= \exp\left(y_i \log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right) \quad \text{vs.} \quad f(y_i|\theta_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right)$$

Choosing:

$$\theta_i = \log\left(\frac{p_i}{1-p_i}\right)$$

$$\phi_i = 1$$

$\longrightarrow$

$$b(\theta_i) = \log\left(1 + e^{\theta_i}\right)$$

$$c(y_i, \phi_i) = 0$$

And we recover the EDF form of the Bernoulli distribution

Anatomy – EDF Family

The EDF family has some useful properties. Namely:

1. $\mathbb{E}[y_i] \equiv \mu_i = b'(\theta_i)$

2. $Var[y_i] = \phi_i b''(\theta_i)$

(the proofs for these identities are in the notes)

Plugging in the values we obtained for Bernoulli, we get back:

$$\mathbb{E}[y_i] = p_i, \;\; Var[y_i] = p_i(1 - p_i)$$

Anatomy – Link Function

Time to talk about the link function

Anatomy – Link Function

Recall from linear regression that:

$$\mu_i = x_i^T \beta$$

Does this work for the Bernoulli distribution?

$$\mu_i = p_i = x_i^T \beta$$

Solution: wrap the *expectation* in a function called the **link function**:

$$g(\mu_i) = x_i^T \beta \equiv \eta_i$$

*For the Bernoulli distribution, the link function is the "logit" function (hence "logistic" regression)

Anatomy – Link Function

Link functions are a *choice*, not a property. A good choice is:

1. Differentiable (implies "smoothness")
2. Monotonic (guarantees invertibility)
   1. Typically increasing so that μ increases w/ η
3. Expands the range of μ to the entire real line

Example: Logit function for Bernoulli

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

Anatomy – Link Function

Logit function for Bernoulli looks familiar…

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

Choosing the link function by setting $\theta_i = \eta_i$ gives us what is known as the "**canonical link function**." Note:

$$\mu_i = b'(\theta_i) \rightarrow \theta_i = b'^{-1}(\mu_i)$$

(derivative of cumulant function must be invertible)

This choice of link, while not always effective, has some nice properties. Take STAT 149 to find out more!

Anatomy – Link Function

Here are some more examples (fun exercises at home)

Distribution $f(y_i\|\theta_i)$	Mean Function $\mu_i = b'(\theta_i)$	Canonical Link $\theta_i = g(\mu_i)$
Normal	θ_i	μ_i
Bernoulli/Binomial	$\frac{e^{\theta_i}}{1+e^{\theta_i}}$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$
Poisson	e^{θ_i}	$\log(\mu_i)$
Gamma	$\frac{-1}{\theta_i}$	$\frac{-1}{\mu_i}$
Inverse Gaussian	$(-2\theta_i)^{-\frac{1}{2}}$	$\frac{-1}{2\mu_i^2}$

Maximum Likelihood Estimation

Maximum Likelihood Estimation

Recall from linear regression – we can estimate our parameters, θ, by choosing those that maximize the likelihood, $L(y|\theta)$, of the data, where:

$$L(y|\theta) = \prod_{i}^{N} p(y_i|\theta_i)$$

In words: likelihood is the probability of observing a set of "N" independent datapoints, given our assumptions about the generative process.

Maximum Likelihood Estimation

For GLM's we can plug in the PDF of the EDF family:

$$L(y|\theta) = \prod_{i=1}^{N} \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i)\right)$$

How do we maximize this? Differentiate w.r.t. θ and set equal to 0. Taking the log first simplifies our life:

$$\ell(y|\theta) = \sum_{i=1}^{N} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{N} c(y_i, \phi_i)$$

Maximum Likelihood Estimation

Through lots of calculus & algebra (see notes), we can obtain the following form for the derivative of the log-likelihood:

$$\ell'(y|\theta) = \sum_{i=1}^{N} \frac{1}{Var[y_i]} \frac{\partial \mu_i}{\partial \beta} (y_i - \mu_i)$$

Setting this sum equal to 0 gives us the **generalized estimating equations:**

$$\sum_{i=1}^{N} \frac{1}{Var[y_i]} \frac{\partial \mu_i}{\partial \beta} (y_i - \mu_i) = 0$$

Maximum Likelihood Estimation

When we use the canonical link, this simplifies to the **normal equations**:

$$\sum_{i=1}^{N} \frac{(y_i - \mu_i)x_i^T}{\phi_i} = 0$$

Let's attempt to solve the normal equations for the Bernoulli distribution. Plugging in μ_i and ϕ_i we get:

$$\sum_{i=1}^{N} \left(y_i - \frac{e^{x_i^T \beta}}{1 - e^{x_i^T \beta}} \right) x_i^T = 0$$

Maximum Likelihood Estimation

Sad news: we can't isolate β analytically.

Maximum Likelihood Estimation

Good news: we can approximate it numerically. One choice of algorithm is the **Fisher Scoring** algorithm.

In order to find the θ that maximizes the log-likelihood, $\ell(y|\theta)$:

1. Pick a starting value for our parameter, θ_0.

2. Iteratively update this value as follows:

$$\theta_{i+1} = \theta_i - \frac{\ell'(\theta_i)}{\mathbb{E}[\ell''(\theta_i)]}$$

In words: perform gradient ascent with a learning rate inversely proportional to the expected curvature of the function at that point.

Maximum Likelihood Estimation

Here are the results of implementing the Fisher Scoring algorithm for simple logistic regression in python:

DEMO

Questions?