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

Marios Mattheakis and Pavlos Protopapas

CS109A Introduction to Data Science

Pavlos Protopapas and Kevin Rader

# Outline

1. Generalized Linear Models (GLMs):
    a. Motivation.
    b. Linear Regression Model (Recap): jumping-off point
    c. Generalize the Linear Model:
        i. Generalization of random component (Error Distribution).
        ii. Generalization of systematic component (Link Function).

2. Maximum Likelihood Estimation in this General Framework:
    a. Canonical Links.
    b. General Links.

# Motivation

Ordinary Linear Regression (OLS) is a great model … but cannot describe all the situations.

OLS assumes:

- ➢ **Normal** distributed observations.
- ➢ **Expectation** that **linearly** depends on predictors.

Many real-world observations do not follow these assumptions, e.g.:

- ➢ Binary data: *Bernoulli* or *Binomial* distributions.
- ➢ Positive data: *Exponential* or *Gamma* distributions.

# GLMs formulations: Overview

**Error distribution:**

Normal

Poisson

Bernoulli

...more

→ Exponential Family Distributions

**Regression Model**

$\mu_i \sim \mathbf{x}_i$

$\mu_i \sim e^{\mathbf{x}_i}$

$\mu_i \sim \frac{e^{\mathbf{x}_i}}{1+e^{\mathbf{x}_i}}$

...more

→ Link Function

$g(\mu_i) \sim \mathbf{x}_i$

Generalized Linear Models

# Regression Models

Suppose a dataset with *n* training points

$$\{y_i, \mathbf{x}_i\} \quad (i = 1, ..., n)$$

$$y_i \in \mathbb{R}$$

$$\mathbf{x}_i \in \mathbb{R}^{p+1}$$

In a Regression model we are looking for:

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

- ➢ $f$ is some fixed but unknown function.
- ➢ $\epsilon_i$ a random error term.

# Linear Regression Model

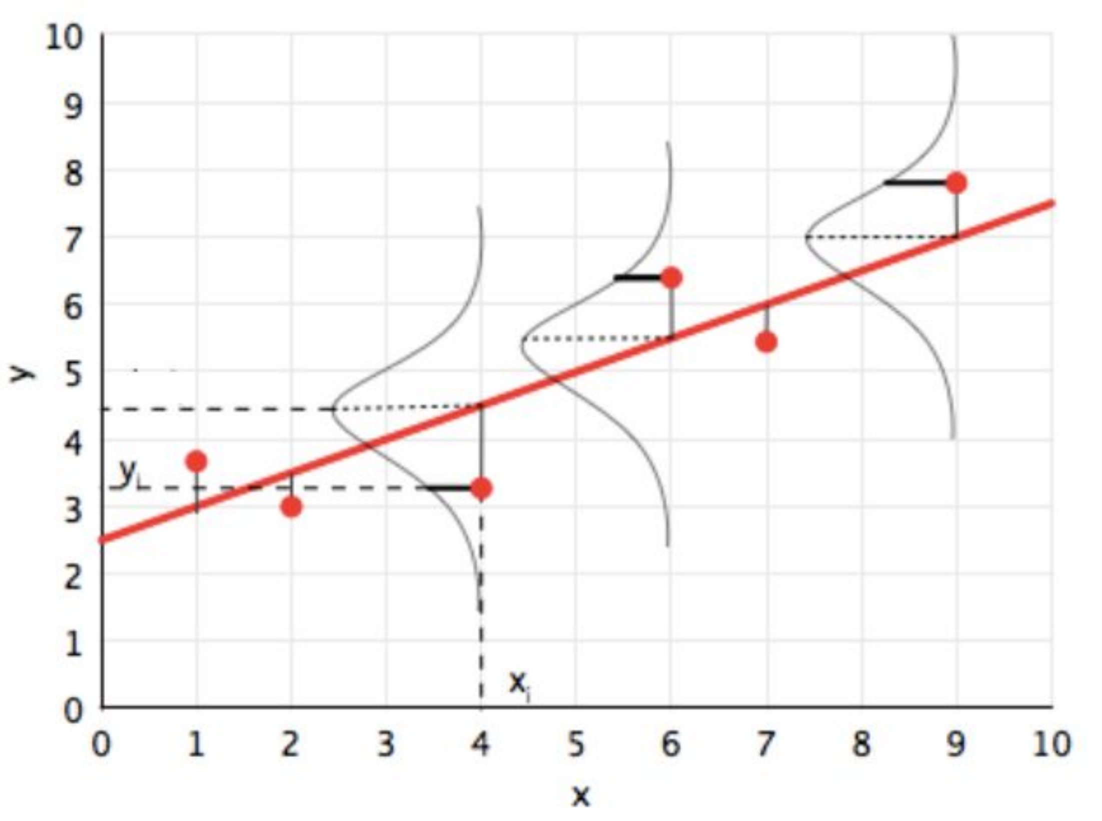
The observations are independently distributed about:

A linear predictor

$$y_i \sim \mathbf{x}_i$$

with a Normal distribution.

$$\epsilon_i \sim \mathcal{N}(0, \sigma^2)$$




Linear Model:

$$y_i = \mathbf{x}_i^T \beta + \epsilon_i \qquad (\beta \in \mathbb{R}^{p+1})$$

# Linear Regression Model

The conditional on the predictor distribution:

$$p(y_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i^T\beta, \sigma^2)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\beta)^2}{2\sigma^2}\right)$$

$$\mu_i = \mathbb{E}[y_i] = \mathbf{x}_i^T\beta$$

$$\text{var}[y_i] = \mathbb{E}[(y_i - \mu_i)^2] = \sigma^2$$

$$y_i = \mathbf{x}_i^T\beta + \epsilon_i$$

# GLMs formulation

# GLMs formulation

This will be a two-step generalization of simple linear regression.

1. Random Component:

$$p(y_i|\mathbf{x}_i) = \mathcal{N}(\mathbf{x}_i^T\beta, \sigma^2) \longrightarrow p(y_i|\mathbf{x}_i) = \text{Exponential Family}$$

2. Systematic Component:

$$\mu_i = \mathbf{x}_i^T\beta \longrightarrow g(\mu_i) = \mathbf{x}_i^T\beta$$

# Exponential Family of Distributions

A wide range of distributions that includes a special cases the Normal, exponential, Gamma, Poisson, Bernoulli, binomial, and many others.

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

$\theta_i$ : ***canonical parameter*** and is the parameter of interest.

$\phi_i$ : ***dispersion parameter*** and is a scale parameter relative to variance.

$b(\theta_i)$ : ***cumulant function*** and completely characterizes the distribution.

$c(y_i, \phi_i)$ : **normalization** factor.

# Likelihood and Score function

Likelihood:

$$L(y_i|\theta_i) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$$

log-likelihood:

$$\ell(y_i|\theta_i) = \sum_{i=1}^{n} \log f_{\theta_i}(y_i)$$

easier and numerically more stable

Score function:

$$S(\theta_i) = \frac{\partial \ell(y_i|\theta_i)}{\partial \theta_i}$$

# Two General Identities

$$\mathbb{E}[S(\theta_i)] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0$$

$$I(\theta_i) \equiv -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \mathrm{var}(S(\theta_i))$$

$I(\theta_i)$ is the called Fisher information matrix.

$\mathbb{E}[.]^{\nu}$ denotes the v moment.

# Some derivatives before the proofs

First derivative of log-likelihood:

$$S(\theta_i) = \frac{\partial \ell}{\partial \theta_i} = \frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i}$$

Second derivative of log-likelihood:

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = \frac{1}{f_{\theta_i}^2} \left( f_{\theta_i} \frac{\partial^2 f_{\theta_i}}{\partial \theta_i^2} - \left( \frac{\partial f_{\theta_i}}{\partial \theta_i} \right)^2 \right)$$

# Some useful relations before the proofs

The ν moment of an arbitrary function:

$$\mathbb{E}[h]^{\nu} = \int_{y_i} h^{\nu} f_{\theta_i}(y_i) dy_i$$

Since the observations are assumed independent of each other:

$$\text{var}[h] = \mathbb{E}\left[(h - \mathbb{E}[h])^2\right] = \mathbb{E}[h^2] - \mathbb{E}[h]^2$$

For a well defined probability density:

$$\int_{y_i} f_{\theta_i}(y_i) dy_i = 1$$

# Proof of Identity I

$$\mathbb{E}[S(\theta_i)] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0$$

$$S(\theta_i) = \frac{\partial \ell}{\partial \theta_i} = \frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i}$$

Proof:

$$\mathbb{E}\left[S(\theta_i)\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = \int_{y_i} \frac{1}{f_{\theta_i}} \frac{\partial f_{\theta_i}}{\partial \theta_i} f_{\theta_i} dy_i$$

$$= \frac{\partial}{\partial \theta_i} \int_{y_i} f_{\theta_i}\, dy_i = 0$$

the regularity condition takes the derivative out of the integral.

# Proof of Identity II

$$I(\theta_i) \equiv -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \text{var}(S(\theta_i))$$

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = \frac{1}{f_{\theta_i}^2}\left(f_{\theta_i}\frac{\partial^2 f_{\theta_i}}{\partial \theta_i^2} - \left(\frac{\partial f_{\theta_i}}{\partial \theta_i}\right)^2\right)$$

## Proof

$$\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] = \mathbb{E}\left[\frac{1}{f_{\theta_i}}\frac{\partial^2 f_{\theta_i}}{\partial \theta_i^2}\right] - \mathbb{E}\left[\left(\frac{1}{f_{\theta_i}}\frac{\partial f_{\theta_i}}{\partial \theta_i}\right)^2\right]$$

1st term:

$$\mathbb{E}\left[\frac{1}{f_{\theta_i}}\frac{\partial^2 f_{\theta_i}}{\partial \theta_i^2}\right] = \int_{y_i}\frac{1}{f_{\theta_i}}\frac{\partial^2 f_{\theta_i}}{\partial \theta_i^2} f_{\theta_i} dy_i = \frac{\partial^2}{\partial \theta_i^2}\int_{y_i} f_{\theta_i} dy_i = 0$$

2nd term:

$$\mathbb{E}\left[\left(\frac{1}{f_{\theta_i}}\frac{\partial f_{\theta_i}}{\partial \theta_i}\right)^2\right] = \mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i} - \underbrace{\mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]}_{=0}\right)^2\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \text{var}(S(\theta_i))$$

# Mean & Variance Formulas in the Exponential Family

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

$$\text{var}[y_i] = \mathbb{E}\left[(y_i - \mu_i)^2\right] = \phi_i b''(\theta_i)$$

where primes denote derivatives w.r.t. canonical parameter $\theta_i$

$b(\theta_i)$ is the **cumulant** function of the distribution, since it completely determines the first two moments.

# Some derivatives before the proofs

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

$$\ell = \log f_{\theta_i} = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + \sum_{i=1}^{n} c(y_i, \phi_i)$$

$$\frac{\partial \ell}{\partial \theta_i} = \sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}$$

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\sum_{i=1}^{n} \frac{b''(\theta_i)}{\phi_i}$$

# Proof of mean formula

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

Proof

$$\mathbb{E}[S(\theta_i)] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = 0$$

$$\mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right] = \mathbb{E}\left[\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}\right] = \sum_{i=1}^{n} \mathbb{E}\left[\frac{y_i - b'(\theta_i)}{\phi_i}\right]$$

$$= \sum_{i=1}^{n} \frac{1}{\phi_i}\mathbb{E}[y_i] - \sum_{i=1}^{n} \frac{1}{\phi_i}\mathbb{E}[b'(\theta_i)] = 0$$

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

# Proof of Variance formula

$$\mathrm{var}[y_i] = \mathbb{E}\left[(y_i - \mu_i)^2\right] = \phi_i b''(\theta_i)$$

Proof

$$I(\theta_i) \equiv -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] = \mathbb{E}\left[\frac{\partial \ell}{\partial \theta_i}\right]^2 = \mathrm{var}(S(\theta_i))$$

$$\mathbb{E}\left[\left(\frac{\partial \ell}{\partial \theta_i}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ell}{\partial \theta_i^2}\right] \Rightarrow \mathbb{E}\left[\left(\sum_{i=1}^{n} \frac{y_i - b'(\theta_i)}{\phi_i}\right)^2\right] = -\mathbb{E}\left[-\sum_{i=1}^{n} \frac{b''(\theta_i)}{\phi_i}\right]$$

$$\Rightarrow \sum_{i=1}^{n} \frac{1}{\phi_i^2}\mathbb{E}\left[(y_i - \mu_i)^2\right] = \sum_{i=1}^{n} \frac{1}{\phi_i}\mathbb{E}\left[b''(\theta_i)\right]$$

$$\Rightarrow \mathrm{var}[y_i] \equiv \mathbb{E}\left[(y_i - \mu_i)^2\right] = \phi_i b''(\theta)$$

# Normal Distribution: Example

Probability density in Normal distribution:

$$f(y_i|\bar{y},\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i-\bar{y})^2}{2\sigma^2}\right)$$

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

$$f(y_i|\bar{y},\sigma^2) = \exp\left(\frac{y_i\bar{y} - \frac{1}{2}\bar{y}^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)$$

$$\theta_i = \bar{y} \quad \text{and} \quad b(\theta_i) = \theta_i^2/2 \quad \text{and} \quad \phi_i = \sigma^2$$

$$\mathbb{E}[y_i] = b' = \theta_i = \bar{y} \qquad\qquad \text{var}[y_i] = \phi_i b'' = \sigma^2$$

# Bernoulli distribution: Example

It is a discrete probability distribution of a random binary variable:

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

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

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

$$\theta_i = \log\frac{p}{1-p} \longrightarrow f(y_i|\theta_i) = \exp\left(y_i\theta_i - \log(1+e_i^{\theta})\right)$$

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

$$\mathbb{E}[y_i] = \frac{e^{\theta_i}}{(1+e^{\theta_i})} = p$$

$$\mathrm{var}[y_i] = \frac{e^{\theta_i}}{(1+e^{\theta_i})^2} = p(1-p)$$

# Second step of GLMs formulation: Link Function

Systematic Component:

$$\mu_i = \mathbf{x}_i^T \beta \longrightarrow g(\mu_i) = \mathbf{x}_i^T \beta$$

# Link Function

A *link function* $g(.)$ is a one-to-one differentiable transformation that transforms the expectation values to be linear with the predictors

$$\eta_i = g(\mu_i) = \mathbf{x}_i^T \beta$$

$\eta_i$ is called *linear predictor.*

One-to-one function, so we can invert to get

$$\mu_i = g^{-1}(\mathbf{x}_i^T \beta)$$

The link transforms the expectation **NOT** the observations.
For instance, for the link

$$g(.) = \log(.)$$

$$\log(\mu_i) = \mathbf{x}_i^T \beta \quad ✓$$

$$\log(y_i) = \mathbf{x}_i^T \beta \quad ✗$$

# Canonical Links

A *Canonical Link* makes the linear predictor equal to the *canonical parameter*

$$\eta_i = g(\mu_i) = \theta_i$$

A *Canonical Transformation* is relative to the cumulant function

$$\begin{aligned} g(\mu_i) &= \theta_i \Rightarrow \\ \mu_i &= g^{-1}(\theta_i) \Rightarrow \\ b'(\theta_i) &= g^{-1}(\theta_i) \end{aligned}$$

So, the cumulant function must be invertible

# Normal and Bernoulli distributions: Examples

**Normal Distribution:**

We found earlier:

$$\theta_i = \mu_i$$

Hence,

$$\theta_i = g(\mu_i) = \mu_i$$

$$g = \text{Identity}$$

**Bernoulli Distribution:**

We found earlier:

$$\theta_i = \log\left[\frac{\mu_i}{1-\mu_i}\right]$$

Hence,

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

$$g = \text{Logit}$$

# Data Distribution and Canonical Links

| Distribution: $f_{\theta_i}$ | Mean Function: $\mu = b'(\theta)$ | Canonical Link: $\theta = g(\mu)$ |
|---|---|---|
| Normal | $\theta$ | $\mu$ |
| Bernoulli/Binomial | $e^{\theta}/(1+e^{\theta})$ | $\log(\mu/(1-\mu))$ |
| Poisson | $e^{\theta}$ | $\log \mu$ |
| Gamma | $-1/\theta$ | $-1/\mu$ |
| Inverse Gaussian | $(-2\theta)^{-1/2}$ | $-1/(2\mu^2)$ |

# GLMs: A general framework

We found that linear, logistic and other regression models are special cases of the GMLs.

Working in such a general framework is a great advantage. There is general theory that can be applied afterwards in any specific distribution and regression model.

For instance, we have the general Likelihood and we can derive to general equations that Maximize the Likelihood.

# Maximum Likelihood Estimation (MLE)

# Maximum Likelihood Estimation (MLE)

Likelihood in the Exponential Family:

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

Log-likelihood in the Exponential Family:

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

# log-likelihood is a strictly concave function

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

$$\frac{\partial^2 \ell}{\partial^2 \beta^2} = \sum_{i=1}^{n} 0 - \frac{b''(\mathbf{x}_i^T \beta)\mathbf{x}_i^2}{\phi_i} = -\sum_{i=1}^{n} \frac{1}{\phi_i^2} \mathrm{var}[y_i]\mathbf{x}_i^2 < 0$$

hence, it can be maximized.

# MLE for Canonical Links

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^{n} \frac{1}{\phi_i} \left( y_i - b' \left( \mathbf{x}_i^T \beta \right) \right) \mathbf{x}_i^T$$

Normal Equations for MLE

$$\sum_{i=1}^{n} \frac{1}{\phi_i} \left( y_i - \mu_i \right) \mathbf{x}_i^T = 0$$

Solving Normal Equations we estimate the coefficients

$$\mu_i = b'(\mathbf{x}_i^T \beta) = g^{-1}(\mathbf{x}_i^T \beta)$$

# MLE Examples

$$\sum_{i=1}^{n} \frac{1}{\phi_i} (y_i - \mu_i) \mathbf{x}_i^T = 0$$

Normal Distribution: Link = Identity

$$\mu_i = \mathbf{x}_i^T \beta \quad \longrightarrow \quad \sum_{i=1}^{n} \frac{1}{\phi_i} \left(y_i - \mathbf{x}_i^T \beta\right) \mathbf{x}_i^T = 0$$

Bernoulli Distribution: Link = Logit

$$\mu_i = \frac{e^{\mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}} \quad \longrightarrow \quad \sum_{i=1}^{n} \frac{1}{\phi_i} \left(y_i - \frac{e^{\mathbf{x}_i^T \beta}}{1 + e^{\mathbf{x}_i^T \beta}}\right) \mathbf{x}_i^T = 0$$

# MLE for General Links

Sometimes we may use non-Canonical links. For instance, for algorithmic purposes such in the Bayesian probit regression.

$$g(\mu_i) = \mathbf{x}_i^T \beta \neq \theta_i$$

Generalizing Estimating Equations:

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

# Summary

- **Generalized Linear Models:**
  1. Motivation: OLS cannot describe everything. Good jumping-off.
  2. Formulation:
     - ➢ Generalization of Random Component (error distribution).
     - ➢ Generalization of Systematic Component (Link function).
  3. Normal & Bernoulli distributions: Examples.

- **Maximum Likelihood Estimation (MLE)**
  1. General Framework: One theory for many regression models.
  2. Normal Equations for MLE (Canonical Links).
     - ➢ Linear & Logistic Regression examples.
  3. Generalized Estimating Equations (General Links).

# Questions ??

Office hours for Adv. Sec.

Monday 6:00-7:30 pm

Tuesday 6:30-8:00 pm

# General Equations: Proof

$$\frac{\partial}{\partial \beta_j}\ell(y_i|\theta_i) = \sum_{i=1}^{n} \frac{1}{\phi_i}\left(y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \beta_j}\right)$$

$$= \sum_{i=1}^{n} \frac{1}{\phi_i}\left(y_i \frac{\partial \theta_i}{\partial \beta_j} - \frac{\partial b(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}\right)$$

$$= \sum_{i=1}^{n} \frac{1}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j}(y_i - \mu_i)$$

Using the chain rule:

$$\frac{\partial \mu_i}{\partial \beta_j} = \frac{\partial b'(\theta_i)}{\partial \beta_j} = \frac{\partial b'(\theta_i)}{\partial \theta_i}\frac{\partial \theta_i}{\partial \beta_j}$$

$$= b''(\theta_i)\frac{\partial \theta_i}{\partial \beta_j} = \frac{\text{var}[y_i]}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j}$$

hence $\frac{1}{\phi_i}\frac{\partial \theta_i}{\partial \beta_j} = \frac{1}{\text{var}[y_i]}\frac{\partial \mu_i}{\partial \beta_j}$

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