# Generalized Linear Models

Greg Strabel

June 9, 2021

## 1 The Overdispersed Exponential Family

The overdispersed exponential family is the class of probability measures that have density functions of the form

$$f_Y\left(y|\theta,\tau\right)=h\left(y,\tau\right)\exp\left(\frac{b\left(\theta\right)^T T\left(y\right)-A\left(\theta\right)}{d\left(\tau\right)}\right) \tag{1}$$

Differentiating $f$ with respect to $\theta_i$

$$\frac{\partial f_Y}{\partial\theta_i}=f_Y\left(y|\theta,\tau\right)\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial\theta_i}T_j\left(y\right)-\frac{\partial A}{\partial\theta_i}\right] \tag{2}$$

so that

$$0=\int\frac{\partial f_Y}{\partial\theta_i}dy=\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial\theta_i}\mathbb{E}T_j\left(y\right)-\frac{\partial A}{\partial\theta_i}\right] \tag{3}$$

from which it follows that

$$\frac{\partial A}{\partial\theta_i}=\sum_{j=1}^{m}\frac{\partial b_j}{\partial\theta_i}\mathbb{E}T_j\left(y\right) \tag{4}$$

Taking the second partial derivative of $f$ with respect to $\theta_i$ yields

$$\begin{aligned}\frac{\partial^2 f_Y}{\partial\theta_i^2}=&\frac{\partial f_Y}{\partial\theta_i}\left(y|\theta,\tau\right)\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial\theta_i}T_j\left(y\right)-\frac{\partial A}{\partial\theta_i}\right]\\&+f_Y\left(y|\theta,\tau\right)\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial^2 b_j}{\partial\theta_i^2}T_j\left(y\right)-\frac{\partial^2 A}{\partial\theta_i^2}\right]\end{aligned} \tag{5}$$

from which it follows that

$$
\begin{aligned}
0 = \int \frac{\partial^2 f_Y}{\partial \theta_i^2} dy = \int f_Y\left(y|\theta,\tau\right)\left(\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial \theta_i}T_j\left(y\right)-\frac{\partial A}{\partial \theta_i}\right]\right)^2 dy \\
+\int f_Y\left(y|\theta,\tau\right)\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial^2 b_j}{\partial \theta_i^2}T_j\left(y\right)-\frac{\partial^2 A}{\partial \theta_i^2}\right] dy \\
= Var\left(\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial \theta_i}T_j\left(y\right)\right]\right)+\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial^2 b_j}{\partial \theta_i^2}\mathbb{E}T_j\left(y\right)-\frac{\partial^2 A}{\partial \theta_i^2}\right]
\end{aligned} \tag{6}
$$

Hence,

$$
Var\left(\frac{1}{d\left(\tau\right)}\left[\sum_{j=1}^{m}\frac{\partial b_j}{\partial \theta_i}T_j\left(y\right)\right]\right)=\frac{1}{d\left(\tau\right)}\left[\frac{\partial^2 A}{\partial \theta_i^2}-\sum_{j=1}^{m}\frac{\partial^2 b_j}{\partial \theta_i^2}\mathbb{E}T_j\left(y\right)\right] \tag{7}
$$

When $b$ is the identity function, (4) implies that

$$
\mathbb{E}T_i\left(y\right)=\frac{\partial A}{\partial \theta_i} \tag{8}
$$

and (7) implies

$$
Var\left(T_i\left(y\right)\right)=d\left(\tau\right)\frac{\partial^2 A}{\partial \theta_i^2} \tag{9}
$$

If both $b$ and $T$ are the identity function, then the model is in *canonical* form and $\theta$ is the *canonical parameter*, in which case $\mathbb{E}y_i = \frac{\partial A}{\partial \theta_i}\left(\theta\right)$.

## 2 Tweedie Distribution

If both $b$ and $T$ are the identity function, $\sigma^2 \equiv d\left(\tau\right)$ and

$$
A\left(\theta\right)=\begin{cases}\frac{1}{2-p}\left(\theta\left(1-p\right)\right)^{\frac{p-2}{p-1}} & \text{p} \in (-\infty,0]\cup(1,2)\cup(2,\infty)\\ -\log\left(-\theta\right) & \text{p=2}\\ e^{\theta} & \text{p=1}\end{cases} \tag{10}
$$

then $Y \sim Tw_p\left(\mu,\sigma^2\right)$, where

$$
\mu \equiv \mathbb{E}Y=\frac{\partial A}{\partial \theta}=\begin{cases}\left(\theta\left(1-p\right)\right)^{\frac{1}{1-p}} & \text{p} \in (-\infty,0]\cup(1,2)\cup(2,\infty)\\ -\theta^{-1} & \text{p=2}\\ e^{\theta} & \text{p=1}\end{cases} \tag{11}
$$

Note that

$$\frac{\partial^2 A}{\partial \theta^2} = \begin{cases} (\theta (1-p))^{\frac{p}{1-p}} & \text{p} \in (-\infty, 0] \cup (1,2) \cup (2, \infty) \\ \theta^{-2} & \text{p=2} \\ e^{\theta} & \text{p=1} \end{cases}$$
$$= \mu^p \tag{12}$$

so that

$$Var(Y) = \sigma^2 \mu^p \tag{13}$$

When $p \in (1,2)$, the Tweedie distribution is the marginal distribution of a compound Poisson Gamma distribution. One can see this by comparing the characteristic functions of the two distributions. First we calculate the characteristic function of the Tweedie distribution:

$$\begin{aligned} \mathbb{E}e^{itY} &= \int e^{ity} \exp\left\{\frac{\theta y - A(\theta)}{\sigma^2}\right\} h(y,\tau)\, dy \\ &= \int \exp\left\{\frac{\left(\theta + it\sigma^2\right) y - A\left(\theta + it\sigma^2\right)}{\sigma^2}\right\} \exp\left\{\frac{A\left(\theta + it\sigma^2\right) - A(\theta)}{\sigma^2}\right\} h(y,\tau)\, dy \\ &= \exp\left\{\frac{A\left(\theta + it\sigma^2\right) - A(\theta)}{\sigma^2}\right\} \end{aligned} \tag{14}$$

Now

$$\begin{aligned} A\left(\theta + it\sigma^2\right) - A(\theta) &= \frac{1}{2-p}\left[\left(\left(\theta + it\sigma^2\right)(1-p)\right)^{\frac{p-2}{p-1}} - (\theta(1-p))^{\frac{p-2}{p-1}}\right] \\ &= \frac{1}{2-p}\left[\left(\left(\frac{\mu^{1-p}}{1-p} + it\sigma^2\right)(1-p)\right)^{\frac{p-2}{p-1}} - \left(\frac{\mu^{1-p}}{1-p}(1-p)\right)^{\frac{p-2}{p-1}}\right] \\ &= \frac{1}{2-p}\left[\left(\mu^{1-p} + it\sigma^2(1-p)\right)^{\frac{p-2}{p-1}} - \left(\mu^{1-p}\right)^{\frac{p-2}{p-1}}\right] \\ &= \frac{\mu^{2-p}}{2-p}\left[\left(1 - it\sigma^2 \frac{p-1}{\mu^{1-p}}\right)^{\frac{p-2}{p-1}} - 1\right] \end{aligned} \tag{15}$$

It follows that

$$\mathbb{E}e^{itY} = \exp\left\{\frac{\mu^{2-p}}{(2-p)\sigma^2}\left[\left(1 - it\sigma^2\frac{p-1}{\mu^{1-p}}\right)^{\frac{p-2}{p-1}} - 1\right]\right\} \tag{16}$$

Now consider the case of a compound Poisson Gamma distribution $C = \sum_{i=1}^{N} Z_i$ where $N \sim Poisson\left(\lambda\right)$ and $Z_i \sim Gamma\left(\alpha, \beta\right)$. The characteristic function of $C$ is

$$\begin{aligned}
\mathbb{E}e^{itC} &= \mathbb{E}_N \left(\mathbb{E}_X e^{itX}\right)^N \\
&= \mathbb{E}_N \left(1 - \frac{it}{\beta}\right)^{-\alpha N} \\
&= \exp\left\{\lambda \left[\left(1 - \frac{it}{\beta}\right)^{-\alpha} - 1\right]\right\}
\end{aligned} \tag{17}$$

Setting

$$\lambda = \frac{\mu^{2-p}}{\left(2-p\right)\sigma^2} \tag{18}$$

$$\alpha = \frac{2-p}{p-1} \tag{19}$$

and

$$\beta = \frac{\mu^{1-p}}{\left(p-1\right)\sigma^2} \tag{20}$$

we find that

$$\mathbb{E}e^{itC} = \exp\left\{\frac{\mu^{2-p}}{\left(2-p\right)\sigma^2}\left[\left(1 - it\sigma^2\frac{p-1}{\mu^{1-p}}\right)^{\frac{p-2}{p-1}} - 1\right]\right\} \tag{21}$$

Thus, when $p \in (1,2)$, the Tweedie distribution is the marginal distribution of a compound Poisson Gamma distribution. Importantly, this implies that when $p \in (1,2)$, the Tweedie distribution has a point mass at zero and

$$\mathbb{P}\left\{Y = 0\right\} = \exp\left\{\frac{\mu^{2-p}}{\left(p-2\right)\sigma^2}\right\} \tag{22}$$

# 3 Generalized Linear Models

In generalized linear models, we assume that the response $Y \in \mathbb{R}^K$ comes from the overdispersed exponential family, that $\theta$ is the canonical parameter, $T$ is the identity function and that $\theta = f\left(\eta\right)$ where $\eta = \left(\mathbb{I}_M \otimes X\right)^T \beta$, $f : \mathbb{R}^M \to \mathbb{R}^K$, $X \in \mathbb{R}^P$ is a set of exogenous predictors and $\beta \in \mathbb{R}^{PM}$. The log-likelihood of a single observation becomes:

$$L = \log h\left(Y, \tau\right) + \frac{1}{d\left(\tau\right)}\left[Y^T f\left(\left(\mathbb{I}_M \otimes X\right)^T \beta\right) - A\left(f\left(\left(\mathbb{I}_M \otimes X\right)^T \beta\right)\right)\right] \tag{23}$$

Note that

$$\frac{\partial L}{\partial \eta} = \frac{1}{d\left(\tau\right)}\left[Y^T - \frac{\partial A}{\partial \theta}\right]\frac{\partial f}{\partial \eta}, \tag{24}$$

$$\frac{\partial L}{\partial \beta} = \frac{1}{d(\tau)} \left[ Y^T - \frac{\partial A}{\partial \theta} \right] \frac{\partial f}{\partial \eta} \left( \mathbb{I}_M \otimes X \right)^T \tag{25}$$

and

$$\frac{\partial^2 L}{\partial \beta \partial \beta'} = \frac{\partial \eta}{\partial \beta}^T \frac{\partial^2 L}{\partial \eta \partial \eta'} \frac{\partial \eta}{\partial \beta} + \frac{\partial L}{\partial \eta} \frac{\partial^2 \eta}{\partial \beta \partial \beta'} \tag{26}$$

Since $\mathbb{E}\left[\frac{\partial L}{\partial \eta} \mid X\right] = 0$ and

$$\frac{\partial^2 L}{\partial \eta \partial \eta'} = -\frac{1}{d(\tau)} \left[ \frac{\partial f}{\partial \eta}^T \frac{\partial^2 A}{\partial \theta \partial \theta'} \frac{\partial f}{\partial \eta} \right] + \frac{1}{d(\tau)} \sum_{k=1}^{K} \left[ Y_k - \frac{\partial A}{\partial \theta_k} \right] \frac{\partial^2 f_k}{\partial \eta \partial \eta'} \tag{27}$$

which implies that

$$\mathbb{E}\left[ \frac{\partial^2 L}{\partial \eta \partial \eta'} \mid X \right] = -\frac{1}{d(\tau)} \frac{\partial f}{\partial \eta}^T \frac{\partial^2 A}{\partial \theta \partial \theta'} \frac{\partial f}{\partial \eta} \tag{28}$$

we have

$$-\mathbb{E}\left[ \frac{\partial^2 L}{\partial \beta \partial \beta'} \mid X \right] = \frac{1}{d(\tau)} \frac{\partial \eta}{\partial \beta}^T \frac{\partial f}{\partial \eta}^T \frac{\partial^2 A}{\partial \theta \partial \theta'} \frac{\partial f}{\partial \eta} \frac{\partial \eta}{\partial \beta} \tag{29}$$

Given a dataset $\{x_i, y_i\}_{i=1\dots n}$, the Newton-Raphson algorithm suggests solving

$$\sum_{i=1}^{n} \frac{1}{d(\tau)} \frac{\partial \eta_i}{\partial \beta}^T \frac{\partial f_i}{\partial \eta}^T \frac{\partial^2 A_i}{\partial \theta \partial \theta'} \frac{\partial f_i}{\partial \eta} \frac{\partial \eta_i}{\partial \beta} \left( \beta^* - \beta \right) = \sum_{i=1}^{n} \frac{1}{d(\tau)} \frac{\partial \eta_i}{\partial \beta}^T \frac{\partial f_i}{\partial \eta}^T \left[ Y_i - \frac{\partial A_i}{\partial \theta}^T \right] \tag{30}$$

for $\beta^*$. Setting

$$w_i = \frac{\partial f_i}{\partial \eta}^T \frac{\partial^2 A_i}{\partial \theta \partial \theta'} \frac{\partial f_i}{\partial \eta} \tag{31}$$

this becomes

$$\sum_{i=1}^{n} \left( \mathbb{I}_M \otimes x_i \right) w_i \left( \mathbb{I}_M \otimes x_i \right)^T \beta^* = \sum_{i=1}^{n} \left( \mathbb{I}_M \otimes x_i \right) w_i \left[ \eta_i + w_i^{-1} \frac{\partial f_i}{\partial \eta}^T \left[ y_i - \frac{\partial A_i}{\partial \theta}^T \right] \right] \tag{32}$$

which is of the form of a weighted least squares regression where the regressors are $(\mathbb{I}_M \otimes x_i)$, the regressands are $\eta_i + w_i^{-1} \frac{\partial f_i}{\partial \eta}^T \left[ y_i - \frac{\partial A_i}{\partial \theta}^T \right]$ and the weights are $w_i$. Hence, the maximum likelihood estimator $\widehat{\beta}$ can be found using iteratively reweighted least squares. In the preceding presentation

$$\mathbb{E}\left[ Y | X \right] = \frac{\partial A}{\partial \theta}^T \left( f \left( \left( \mathbb{I}_M \otimes X \right)^T \beta \right) \right) \tag{33}$$

If we define $g$ implicitly by the equation

$$g^{-1} := \frac{\partial A}{\partial \theta}^T \circ f \tag{34}$$

then we recover the common presentation of generalized models that start with a *link function* $g$ such that

$$g\left( \mathbb{E}\left[ Y \mid X \right] \right) = \left( \mathbb{I}_M \otimes X \right)^T \beta \tag{35}$$