Applied Statistical Methods II

Chapter #14

Logistic Regression, Poisson Regression, and Generalized Linear Models

Part I

Looking forward:

- Look at regression models for binary outcomes.
- Focus on logistic regression.

Linear Regression Model

Recall that

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} + \epsilon_i.$$

- We usually assumed that ϵ_i are iid $N(0, \sigma^2)$.
- This implies that $Y_i \sim N(X_i^T \beta, \sigma^2)$.

Example of Binary

- A bank wants to determine how age affects the probability of a person defaulting on their loan.
- They consider all loans that were outstanding during the year 2010.
- The data are:
  - $Y_i = 1$ if a person defaulted on their loan in 2010 and 0 otherwise.
  - X_i the person's age.

Binomial Distribution

- Intuitively, we can model a person defaulting in 2010 as a Bernoulli random variable:
  - $Y_i \sim Bernoulli(\pi_i)$
  - $Y_i = 1$ with probability π_i.
  - $Y_i = 0$ with probability $1 - \pi_i$.
- Want to determine how π_i depends on X_i.
  - In this example, we only look at one covariate X_i.
  - As in the linear regression model, we can have multiple covariates.

Mean and Variance

$$\begin{aligned} E(Y_i) &= 1 \times P(Y_i = 1) + 0 \times Pr(Y_i = 0) \\ &= 1 \times \pi_i + 0 \times (1 - \pi_i) \\ &= \pi_i. \end{aligned}$$

$$\begin{aligned} Var(Y_i) &= E(Y_i^2) - [E(Y_i)]^2 \\ &= \pi_i - \pi_i^2 \\ &= \pi_i (1 - \pi_i). \end{aligned}$$

The variance is a function of the mean! What could go wrong if we model Y_i with a standard non-linear or linear model?

Regression model for Binary response

- We model $\pi_i = \beta_0 + \beta_1 X_i$.
- Note that $0 \leq \pi_i \leq 1$.
- But the fitted $\hat{\beta}_0 + \hat{\beta}_1 X_i$ might be outside $[0, 1]$ for some X_i.
- Solution: model $g(\pi_i) = \beta_0 + \beta_1 X_i$ for some function g that maps $[0, 1]$ to $[-\infty, +\infty]$.
  - g is called the link function.
  - In general, we require g to be strictly increasing or decreasing and differentiable.
  - This make g invertible.
- Model $Y_i \sim Bernoulli(\pi_i)$ and $\pi_i = g^{-1}(\beta_0 + \beta_1 X_i)$
  - Note that g^{-1} exists and is differentiable by strict monotonicity/differentiability requirements.

Regression model for Binary response

- We model $\pi_i = \beta_0 + \beta_1 X_i$.
- Note that $0 \leq \pi_i \leq 1$.
- But the fitted $\hat{\beta}_0 + \hat{\beta}_1 X_i$ might be outside $[0, 1]$ for some X_i.
- Solution: model $g(\pi_i) = \beta_0 + \beta_1 X_i$ for some function g that maps $[0, 1]$ to $[-\infty, +\infty]$.
  - g is called the link function.
  - In general, we require g to be strictly increasing or decreasing and differentiable.
  - This make g invertible.
- Model $Y_i \sim Bernoulli(\pi_i)$ and $\pi_i = g^{-1}(\beta_0 + \beta_1 X_i)$
  - Note that g^{-1} exists and is differentiable by strict monotonicity/differentiability requirements.

Possible link function g

To determine potential link functions, it is useful to look at log-likelihood of Bernoulli:

- If $y \sim \text{Ber}(\pi)$, then likelihood is $L(\pi \mid y) = \pi^y (1-\pi)^{1-y}$ and log-likelihood is

$$\ell(\pi \mid y) = y \log(\pi) + (1-y)\log(1-\pi) = y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)$$

- Setting $\eta = \log\left(\frac{\pi}{1-\pi}\right)$, we get $\ell(\eta; y) = y\eta - \log(1+e^{\eta}) = y\eta - b(\eta)$.
- $\eta = \log\left(\frac{\pi}{1-\pi}\right)$ is called the **natural parameter**.
  - rth **cumulant** of y is $b^{(r)}(\eta) \Rightarrow E(y) = b'(\eta)$ and $\text{Var}(y) = b''(\eta)$.
  - $\ell(\eta \mid y)$ is concave in η.
- $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ is called the **logit function**, and satisfies definition of a link.

Possible link function g

To determine potential link functions, it is useful to look at log-likelihood of Bernoulli:

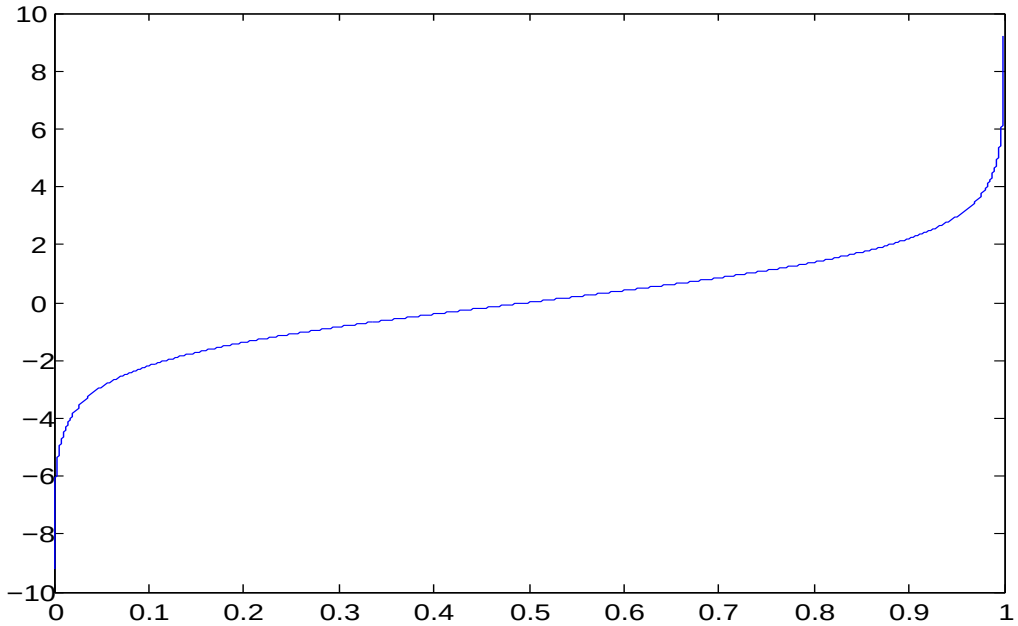
- If $y \sim \text{Ber}(\pi)$, then likelihood is $L(\pi \mid y) = \pi^y (1-\pi)^{1-y}$ and log-likelihood is

$$\ell(\pi \mid y) = y \log(\pi) + (1-y)\log(1-\pi) = y \log\left(\frac{\pi}{1-\pi}\right) + \log(1-\pi)$$

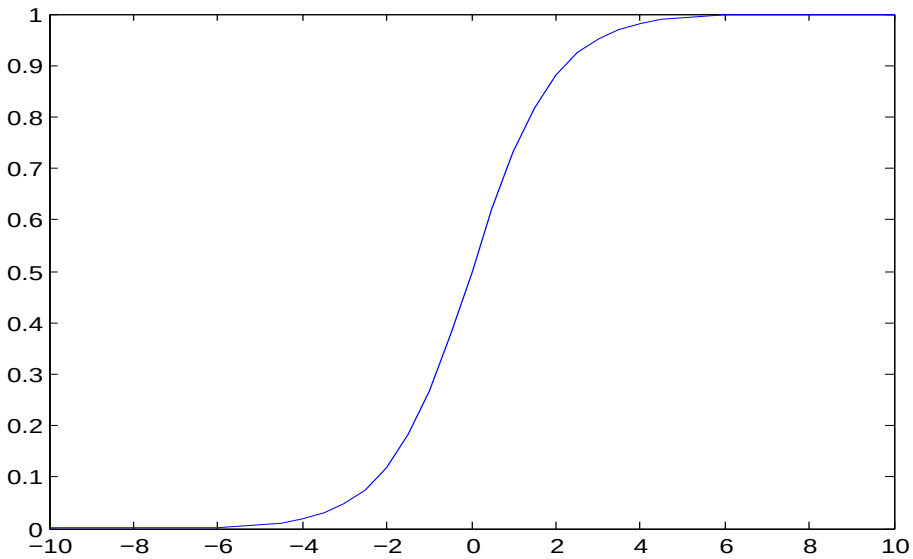
- Setting $\eta = \log\left(\frac{\pi}{1-\pi}\right)$, we get $\ell(\eta; y) = y\eta - \log(1+e^{\eta}) = y\eta - b(\eta)$.
- $\eta = \log\left(\frac{\pi}{1-\pi}\right)$ is called the **natural parameter**.
  - rth **cumulant** of y is $b^{(r)}(\eta) \Rightarrow E(y) = b'(\eta)$ and $\text{Var}(y) = b''(\eta)$.
  - $\ell(\eta \mid y)$ is concave in η.
- $g(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$ is called the **logit function**, and satisfies definition of a link.

Logit function for g

- By far the most popular link function for Bernoulli data is the logit function.
- $\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \beta_0 + \beta_1 X_i$



- The quantity $\frac{\pi_i}{1-\pi_i}$ is known as the **odds** of Y_i. So we model the log odds as a linear function of X_i.
- Note that there is a one-to-one correspondence between the odds and probability of Y_i.
- The inverse of the logit is the expit:
  - $\pi_i = \text{expit}(\eta_i) = \frac{\exp(\eta_i)}{1+\exp(\eta_i)}$



Other Link Functions

- Other link functions are also used for Bernoulli data.
- Probit function: $\eta_i = \Phi^{-1}(\pi_i)$
- Inverse T_ν-distribution: $\eta_i = F_\nu^{-1}(\pi_i)$
- Complementary log-log: $\eta_i = \log(-\log(1-\pi_i))$
- Even identity link is OK in some applications.
- Logit: Its coefficients have nice interpretations as odds ratios, and it is the natural link function derived from the canonical form of the exponential famliy (will see later).

Putting It Together: Regression Models for Binary Data

- $Y_i \mid X_i \sim Ber(\pi_i)$ and are **independent**.
- $g(\pi_i) = X_i^T \beta$. Note: there is NO transformation of response Y_i, only a transformation of its mean.
- likelihood function:

$$
\begin{aligned}
& L(\beta | Y_1, \ldots, Y_n) \qquad (1) \\
= & \prod_i \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i} \\
= & \prod_i \left\{ g^{-1}(X_i^T \beta) \right\}^{Y_i} \left\{ 1 - g^{-1}(X_i^T \beta) \right\}^{1 - Y_i}
\end{aligned}
$$

- We will talk about MLE.
- The model on page 563, $Y_i = E(Y_i) + \epsilon_i$, is not intuitive.

Putting It Together: Regression Models for Binary Data

- $Y_i \mid X_i \sim Ber(\pi_i)$ and are **independent**.
- $g(\pi_i) = X_i^T \beta$. Note: there is NO transformation of response Y_i, only a transformation of its mean.
- likelihood function:

$$
\begin{aligned}
& L(\beta | Y_1, \ldots, Y_n) \qquad (1) \\
= & \prod_i \pi_i^{Y_i} (1 - \pi_i)^{1 - Y_i} \\
= & \prod_i \left\{ g^{-1}(X_i^T \beta) \right\}^{Y_i} \left\{ 1 - g^{-1}(X_i^T \beta) \right\}^{1 - Y_i}
\end{aligned}
$$

- We will talk about MLE.
- The model on page 563, $Y_i = E(Y_i) + \epsilon_i$, is not intuitive.

Simple Logistic Regression Model

- The simple logistic regression model is:
  - $Y_i \mid X_i$ are independent Bernoulli(π_i) random variables.
  - $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_i$.
  - Recall: $\text{logit}(\pi) = \log[\pi/(1-\pi)]$
- We will use maximum likelihood to estimate β_0 and β_1, and do inference.
- For now we will focus on interpretations of the parameters β_0 and β_1.

Thinking In Terms of Odds

- Notation: let $\pi(X_i) = P(Y_i = 1 \mid X_i) = E(Y_i \mid X_i)$.
  - In the loan example, $\pi(X_i)$ is the probability that a person aged X_i will default.
  - The odds of a person aged X_i defaulting is $\text{odds}(X_i) = \frac{\pi(X_i)}{1-\pi(X_i)}$.
- The odds and probability contain the same information but interpretation is different.
- The odds rages from:
  - 0 when $\pi(X_i) = 0$
  - ∞ when $\pi(X_i) = 1$
  - equals 1 when $\pi(X_i) = 0.5$

Interpretation of β_0

$$\begin{aligned}\beta_0 = \beta_0 + \beta_1 \times 0 &= \text{logit}(\pi(X_i = 0)) \\ &= \log \text{odds}(X_i = 0) \\ &= \log\left(\frac{\pi(X_i = 0)}{1 - \pi(X_i = 0)}\right)\end{aligned}$$

- β_0 is the log-odds of $Y_i = 1$ when $X_i = 0$.
- In the loan example, assume that age is centered at 40: X_i = age - 40.
  - β_0 is the log odds of a person of age 40 defaulting.
  - $\exp(\beta_0)$ is the odds of a person of age 40 defaulting.
  - $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability of a person of age 40 defaulting.

Interpretation of β_1

$$\begin{aligned}
\beta_1 &= (\beta_0 + \beta_1(X_i + 1)) - (\beta_0 + \beta_1 X_i) \\
&= \log\left\{\frac{\pi(X_i+1)}{1-\pi(X_i+1)}\right\} - \log\left\{\frac{\pi(X_i)}{1-\pi(X_i)}\right\} \\
&= \log\{\text{odds}(X_i+1)\} - \log\{\text{odds}(X_i)\} \\
&= \log\left\{\frac{\text{odds}(X_i+1)}{\text{odds}(X_i)}\right\}
\end{aligned}$$

- $OR = \frac{\text{odds}(X_i+1)}{\text{odds}(X_i)} = \frac{\pi(X_i+1)/(1-\pi(X_i+1))}{\pi(X_i)/(1-\pi(X_i))}$ is the **odds ratio** between $X_i + 1$ and X_i
- $OR \in (0, \infty)$.
- In hypothesis testing, $H_0 : OR$ between $X_i + \delta_x$ and X_i is 1 $\Leftrightarrow H_0 : \beta_1 = 0$.

Interpretation of β_1

$$
\begin{aligned}
\beta_1 &= (\beta_0 + \beta_1(X_i + 1)) - (\beta_0 + \beta_1 X_i) \\
&= \log\left\{\frac{\pi(X_i + 1)}{1 - \pi(X_i + 1)}\right\} - \log\left\{\frac{\pi(X_i)}{1 - \pi(X_i)}\right\} \\
&= \log\{\text{odds}(X_i + 1)\} - \log\{\text{odds}(X_i)\} \\
&= \log\left\{\frac{\text{odds}(X_i + 1)}{\text{odds}(X_i)}\right\}
\end{aligned}
$$

- $OR = \frac{\text{odds}(X_i+1)}{\text{odds}(X_i)} = \frac{\pi(X_i+1)/(1-\pi(X_i+1))}{\pi(X_i)/(1-\pi(X_i))}$ is the **odds ratio** between $X_i + 1$ and X_i
- $OR \in (0, \infty)$.
- In hypothesis testing, H_0 : OR between $X_i + \delta_x$ and X_i is 1 $\Leftrightarrow$ $H_0 : \beta_1 = 0$.

The loan example

- In the loan example, β_1 is the log odds ratio between a person of age $X_i + 1$ and a person of age X_i.
- $\exp(\beta_1)$ is called the odds ratio for age.
- $OR > 1$ (equivalently, $\beta_1 > 0$) means that the probability of default increases as age increases.
- Testing if $\beta_1 = 0$ is the same as testing if $\exp(\beta_1) = 1$.

When X_i is a factor variable

- In the loan example, what if X_i is the indicator variable for a person being male?
- We model $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_i$.

$$\begin{aligned} \beta_1 &= \log\left\{\frac{\text{odds}(X_i = 1)}{\text{odds}(X_i = 0)}\right\} \\ &= \log\left\{\frac{\pi(X_i = 1)/(1 - \pi(X_i = 1))}{\pi(X_i = 0)/(1 - \pi(X_i = 0))}\right\} \end{aligned}$$

- β_1 is the log odds ratio of a loan default between males and females.
- $\exp(\beta_1)$ is the odds ratio between males and females.

Logistic Regression with Many Predictors

- We can easily incorporate many different variables in the same logistic regression model.
  - Can also include interaction terms.
- Assume that we observe $Y_i = 0$ or $Y_i = 1$ and we think that $\pi_i = P(Y_i = 1)$ depends on the predictors $X_{i1}, \ldots, X_{i(p-1)}$.
- Can form the logistic regression model:
  - $Y_i \mid X_i$ are independent $Ber(\pi_i)$ random variables.
  - $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)} = X_i^T \beta$
- β_j is the log odds ratio for a loan default for an increase in X_{ij} by one unit while holding all other predictors constant.

The Logistic Regression Model

- $Y_i \mid X_i$ are independent for $i = 1, \ldots, n$
- $Y_i \mid X_i \sim Bernoulli(\pi_i)$
- $\text{logit}(\pi_i) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)}$
  - $\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)})}{1 + \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i(p-1)})}$

Maximum Likelihood Estimation

- We will estimate $\beta_0, \ldots, \beta_{p-1}$ with $\hat{\beta}_0, \ldots, \hat{\beta}_{p-1}$.
- We can use the maximum likelihood estimators (MLEs).
- If $Y_i \sim Bernoulli(\pi_i)$, then its probability distribution is:
  - $f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}$ for $Y_i = 0, 1$
- Since Y_i are independent, their joint distribution is:
  - $g(Y_1, \ldots, Y_n) = \Pi_{i=1}^{n} f_i(Y_i)$
- It is easier to maximize the log-distribution as compared to the distribution.

$$\begin{aligned}
\log\left[g(Y_1, \ldots, Y_n)\right] &= \sum_{i=1}^{n} \log\left[f_i(Y_i)\right] \\
&= \sum_{i=1}^{n} [Y_i \log(\pi_i) + (1 - Y_i)\log(1-\pi_i)]
\end{aligned}$$

Log-likelihood

Writing in terms of the parameters

$$\begin{aligned}\ell(\beta_0, \ldots, \beta_{p-1}) &= \sum_{i=1}^{n} \{ Y_i \log(\pi_i) + (1 - Y_i) \log(1 - \pi_i) \} \\ &= \sum_{i=1}^{n} \left\{ Y_i X_i^T \beta - \log \left[1 + \exp \left(X_i^T \beta \right) \right] \right\}\end{aligned}$$

We want to find $\hat{\beta} = \left(\hat{\beta}_0, \ldots, \hat{\beta}_{p-1} \right)^T$ that maximizes ℓ.

Solving the MLE's

- To find the MLE's, we first compute the p score functions:
  - $U_j(\beta_0, \ldots, \beta_{p-1}) = \frac{\partial \ell}{\partial \beta_j}$
- We want to solve the system of equations: $U_j = 0$ for $j = 0, \ldots, p-1$.
- This has no closed form solution.
- Two popular computation routines for solving score equation (in general) are Newton-Raphson and Fisher's Scoring.
  - For logistic regression, they are the same thing.

Numerical MLE notations

We aim at maximizing $l(\beta)$ w.r. to $\beta \in \mathcal{R}^p$.

Score vector	$U(\beta) = \begin{pmatrix} \partial l/\partial \beta_1 \\ \vdots \\ \partial l/\partial \beta_p \end{pmatrix}$
Hessian matrix	$H(\beta) = (\partial^2 l/\partial \beta_j \partial \beta_k), \ \ 1 \leq j, k \leq p$
Observed information	$I_{\text{obs}}(\beta) = -H(\beta)$
Information	$I(\beta) = E(I_{\text{obs}}(\beta))$
ML estimating equations	$U(\hat{\beta}) = 0, \qquad$ root $\hat{\beta}$ is MLE

Numerical MLE notations

- Taylor expansion for the score function around β near $\hat{\beta}$. $0 = U(\hat{\beta}) = U(\beta) + H(\beta)(\hat{\beta} - \beta) + O(\|\hat{\beta} - \beta\|^2)$
- $\hat{\beta} \approx \beta - H^{-1}(\beta)U(\beta)$ under regularity conditions.
- This motivates the iterative procedure: choose initial value, update, converge.
- Newton-Raphson: $\beta_l = \beta_{l-1} - H^{-1}(\beta_{l-1})U(\beta_{l-1})$
  - Pro: $H(\beta)$ is exact Hessian, better approximation to function.
  - Con: $-H$ may not be positive definite far away from solution.
- Fisher scoring: $\beta_l = \beta_{l-1} + I^{-1}(\beta_{l-1})U(\beta_{l-1}) = \beta_{l-1} + [-E\{H(\beta_{l-1})\}]^{-1} U(\beta_{l-1})$
  - Pro: $I(\beta)$ is always positive definite.
  - Con: $I(\beta)$ may not be a good approximation for Hessian far away from sol'n.

Summary

- Newton-Raphson and Fisher's scoring are both linearizations of the score functions.
- Newton-Raphson uses the observed information and Fisher's scoring using the expected.
- Are not guaranteed to converge.
- Fisher's tends to have better convergence than Newton-Raphson.
- In general, starting values for logistic regression are not as important as those for a non-linear regression using Gauss-Newton.
    - Logistic regression is equivalent to minimizing minus log-likelihood, which is convex in β.

Application to logistic regression

- $$\begin{aligned}\ell(\beta_0,\ldots,\beta_{p-1}) &= \sum_{i=1}^{n}\{Y_i\log(\pi_i)+(1-Y_i)\log(1-\pi_i)\}\\ &= \sum_{i=1}^{n}\left\{Y_iX_i^T\beta-\log\left[1+\exp\left(X_i^T\beta\right)\right]\right\}\end{aligned}$$
- $U=\nabla_\beta\ell(\beta)=\sum_{i=1}^{n}(Y_i-\pi_i)X_i=X^T(Y-\pi)$
- $H=\nabla^2_\beta\ell(\beta)=-\sum_{i=1}^{n}(1-\pi_i)\pi_iX_iX_i^T=$ $-X^T\text{diag}\{\pi_1(1-\pi_1),\ldots,\pi_n(1-\pi_n)\}X.$
- We can see that H does not depend on the data, so that $E(H)=H$ and $I=-H$.
- Fisher's scoring and Newton-Raphson are the same for logistic regression.
- We will show later that these two are the same if one uses the canonical link function.

Application to logistic regression

- $$\begin{aligned}\ell(\beta_0,\ldots,\beta_{p-1}) &= \sum_{i=1}^{n}\{Y_i\log(\pi_i)+(1-Y_i)\log(1-\pi_i)\}\\ &= \sum_{i=1}^{n}\left\{Y_iX_i^T\beta-\log\left[1+\exp\left(X_i^T\beta\right)\right]\right\}\end{aligned}$$
- $U=\nabla_\beta\ell(\beta)=\sum_{i=1}^{n}(Y_i-\pi_i)X_i=X^T(Y-\pi)$
- $H=\nabla^2_\beta\ell(\beta)=-\sum_{i=1}^{n}(1-\pi_i)\pi_iX_iX_i^T=$ $-X^T\text{diag}\{\pi_1(1-\pi_1),\ldots,\pi_n(1-\pi_n)\}X.$
- We can see that H does not depend on the data, so that $E(H)=H$ and $I=-H$.
- Fisher's scoring and Newton-Raphson are the same for logistic regression.
- We will show later that these two are the same if one uses the canonical link function.