# Lecture 4: Generalized Linear Models

## STATS305B: Applied Statistics II

Scott Linderman

January 15, 2025

# Recap

Last time...

- Definition and Examples of Exponential Family Distributions
- The Log Normalizer
- Maximum Likelihood Estimation
- Mean Parameterization
- KL Divergence and Deviance Residuals

# Outline

Today...

- Generalized Linear Models (GLMs)
- Canonical form
- Demo: Poisson GLM
- Non-canonical forms
- Model checking and comparison

# Generalized Linear Models

Logistic regression was a special case of a more general class of models called **generalized linear models** (GLMs).

In a GLM, the conditional distribution $p(Y \mid X = x)$ is modeled as an exponential family distribution whose mean parameter is a function of $X$.

For example, if $Y \in \mathbb{N}$, we could model it with a Poisson GLM; if $Y \in \{1, \ldots, K\}$, we could model it as a categorical GLM.

Many of the nice properties of logistic regression carry over!

# Model

To construct a generalized linear model with exponential family observations, we set

$$\mathbb{E}[y_i \mid \boldsymbol{x}_i] = f(\boldsymbol{\beta}^\top \boldsymbol{x}_i).$$

From above, this implies,

$$\begin{aligned} \nabla A(\eta_i) &= f(\boldsymbol{\beta}^\top \boldsymbol{x}_i) \\ \Rightarrow \eta_i &= [\nabla A]^{-1}\big(f(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\big), \end{aligned}$$

when $\nabla A(\cdot)$ is invertible. (In this case, the exponential family is said to be **minimal**).

The **canonical mean function** is $f(\cdot) = \nabla A(\cdot)$ so that $\eta_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i$.

The (canonical) **link function** is the inverse of the (canonical) mean function.

# Logistic regression revisited

Consider the Bernoulli distribution once more. The gradient of the log normalizer is,

$$\nabla A(\eta) = \nabla \log(1 + e^{\eta}) = \frac{e^{\eta}}{1 + e^{\eta}}$$

This is the logistic function!

Thus, logistic regression is a Bernoulli GLM with the canonical mean function.

# Example: Poisson GLM

Recall that the Poisson distribution can be written in exponential family form as,

$$\mathrm{Po}(y;\lambda) = \frac{1}{y!}\lambda^y e^{-\lambda} = \frac{1}{y!}\exp\{y\log\lambda - \lambda\} = h(y)\exp\left\{\langle t(y),\eta\rangle - A(\eta)\right\}$$

where

- $h(y) = 1/y!$
- $t(y) = y$
- $\eta = \log\lambda$
- $A(\eta) = e^{\eta}$

The canonical mean function is $f(x^\top\beta) = \nabla A(x^\top\beta) = e^{x^\top\beta}$.

The canonical mean function is the exponential.

# Log likelihood: Canonical case

Canonical mean functions lead to nice math. Consider the log joint probability,

$$\begin{aligned}\mathscr{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \langle t(y_i), \eta_i \rangle - A(\eta_i) + c \\ &= \sum_{i=1}^{n} \langle t(y_i), \boldsymbol{\beta}^\top \boldsymbol{x}_i \rangle - A(\boldsymbol{\beta}^\top \boldsymbol{x}_i) + c,\end{aligned}$$

where we have assumed a canonical mean function so $\eta_i = \boldsymbol{\beta}^\top \boldsymbol{x}_i$.

# Gradient of the log likelihood: Canonical case

The gradient is,

$$\begin{aligned}\nabla \mathscr{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \langle t(y_i), \boldsymbol{x}_i \rangle - \langle \nabla A(\boldsymbol{\beta}^\top \boldsymbol{x}_i), \boldsymbol{x}_i \rangle \\ &= \sum_{i=1}^{n} \langle t(y_i) - \mathbb{E}[t(Y); \eta_i], \boldsymbol{x}_i \rangle\end{aligned}$$

where $\eta_i = \boldsymbol{x}_i^\top \boldsymbol{\beta}$ for a GLM with canonical link.

In many cases, $t(y_i) = y_i \in \mathbb{R}$ so

$$\nabla \mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i) \boldsymbol{x}_i.$$

# Hessian of the log likelihood: Canonical case

When $t(y_i) = y_i$, the Hessian is

$$\begin{aligned}\nabla^2_{\boldsymbol{\beta}} \mathscr{L}(\boldsymbol{\beta}) &= -\sum_{i=1}^{n} \nabla^2 A(\boldsymbol{\beta}^\top \boldsymbol{x}_i)\, \boldsymbol{x}_i \boldsymbol{x}_i^\top \\ &= -\sum_{i=1}^{n} \mathrm{Var}[Y; \eta_i]\, \boldsymbol{x}_i \boldsymbol{x}_i^\top\end{aligned}$$

# Newton updates

Now recall the undamped Newton's method updates, written here in terms of the change in weights,

$$\begin{aligned}\Delta\boldsymbol{\beta} &= -[\nabla^2 \mathscr{L}(\boldsymbol{\beta})]^{-1} \nabla \mathscr{L}(\boldsymbol{\beta}) \\ &= \left[\sum_{i=1}^{n} \mathrm{Var}[t(Y); \eta_i] \boldsymbol{x}_i \boldsymbol{x}_i^\top\right]^{-1} \left[\sum_{i=1}^{n} (y_i - \hat{y}_i) \boldsymbol{x}_i\right]\end{aligned}$$

Letting $w_i = \mathrm{Var}[t(Y); \eta_i]$,

$$\begin{aligned}\Delta\boldsymbol{\beta} &= \left[\sum_{i=1}^{n} w_i \boldsymbol{x}_i \boldsymbol{x}_i^\top\right]^{-1} \left[\sum_{i=1}^{n} (y_i - \hat{y}_i) \boldsymbol{x}_i\right] \\ &= (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X})^{-1} [\boldsymbol{X}^\top \boldsymbol{W} \hat{\boldsymbol{z}}]\end{aligned}$$

where $\boldsymbol{W} = \mathrm{diag}([w_1, \ldots, w_i])$ and $\hat{\boldsymbol{z}} = \boldsymbol{W}^{-1}(\boldsymbol{y} - \hat{\boldsymbol{y}})$.

# Iteratively reweighted least squares

This is **iteratively reweighted least squares (IRLS)** with weights $w_i$ and working responses

$$\hat{z}_i = \frac{y_i - \hat{y}_i}{w_i},$$

both of which are functions of the current weights $\boldsymbol{\beta}$.

As in logistic regression, the working responses can be seen as linear approximations to the observed responses mapped back through the link function.

# Demo: Neural Spike Train Analysis

*Go to Colab notebook*

# Non-canonical case

When we choose an arbitrary mean function, the expressions are a bit more complex. Let's focus on the case where $t(y_i) = y_i$ for scalar $y_i$, but allow for arbitrary mean function $f$.

$$\mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \langle y_i, \eta_i \rangle - A(\eta_i) + c,$$

but now $\eta_i = [\nabla A]^{-1} f(\boldsymbol{\beta}^\top \boldsymbol{x}_i)$.

The gradient is,

$$\nabla \mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \hat{y}_i) \frac{\partial \eta_i}{\partial \boldsymbol{\beta}}.$$

# Non-canonical case

Applying the inverse function theorem, as above, yields,

$$\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \text{Var}[Y]^{-1}\boldsymbol{x}_i = \boldsymbol{x}_i / w_i,$$

and

$$\nabla \mathscr{L}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left( \frac{y_i - \hat{y}_i}{w_i} \right) f'(\boldsymbol{\beta}^\top \boldsymbol{x}_i) \boldsymbol{x}_i.$$

With automatic differentiation, it's not a big deal to fit GLMs with non-canonical mean functions using gradient descent or Newton's method. The math just isn't as clean.

# Deviance and Goodness of Fit

We can perform maximum likelihood estimation via Newton's method, but do the resulting parameters $\hat{\boldsymbol{\beta}}_{\text{MLE}}$ provide a good fit to the data?

One way to answer this question is by comparing the fitted model to two reference points: a **saturated** model and a **baseline** model.

For binomial GLMs (including logistic regression) and Poisson GLMs, the saturated model conditions on the mean equaling the observed response.

For example, in a Poisson GLM the saturated model's log probability is,

$$\begin{aligned}\log p_{\text{sat}}(\boldsymbol{y}) &= \sum_{i=1}^{n} \log \text{Po}(y_i; y_i) \\ &= \sum_{i=1}^{n} -\log y_i! + y_i \log y_i - y_i.\end{aligned}$$

# Deviance and Goodness of Fit

The likelihood ratio statistic in this case is

$$\begin{aligned} -2 \log \frac{p(\boldsymbol{y} \mid \boldsymbol{X}; \hat{\boldsymbol{\beta}})}{p_{\text{sat}}(\boldsymbol{y})} &= 2 \sum_{i=1}^{n} \log \text{Po}(y_i; y_i) - \log \text{Po}(y_i; \hat{\mu}_i) \\ &= 2 \sum_{i=1}^{n} y_i \log \frac{y_i}{\hat{\mu}_i} + \hat{\mu}_i - y_i \\ &= \sum_{i=1}^{n} r_{\text{D}}(y_i, \hat{\mu}_i)^2 \end{aligned}$$

where $\hat{\mu}_i = f(\boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}})$ is the predicted mean.

We recognize the likelihood ratio statistic as the sum of squared deviance residuals! (See last lecture.)

# Deviance and Goodness of Fit

Moreover, this statistic is just the deviance (twice the KL divergence) between the two models,

$$
\begin{aligned}
2D_{\mathrm{KL}}\left(p_{\mathrm{sat}}(\boldsymbol{Y}) \,\|\, p(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x}; \hat{\boldsymbol{\beta}})\right) &= 2\mathbb{E}_{p_{\mathrm{sat}}}\left[\log \frac{p_{\mathrm{sat}}(\boldsymbol{y})}{p(\boldsymbol{y} \mid \boldsymbol{X}; \hat{\boldsymbol{\beta}})}\right] \\
&= 2\sum_{i=1}^{n} D_{\mathrm{KL}}\left(\mathrm{Po}(y_i) \,\|\, \mathrm{Po}(\hat{\mu}_i)\right) \\
&\triangleq D(\boldsymbol{y}, \hat{\boldsymbol{\mu}}).
\end{aligned}
$$

The same idea holds for GLMs with other exponential family distributions.

Larger deviance implies a poorer fit.

# Deviance and Goodness of Fit: Baseline Model

Larger deviance relative to what?

The baseline model is typically a GLM with only an intercept term, in which case the MLE is $\mu_i \equiv \frac{1}{n}\sum_{i=1}^{n} y_i = \bar{y}$.

For that baseline model, the deviance is,

$$\begin{aligned} D(\boldsymbol{y}, \bar{y}\mathbf{1}) &= 2\sum_{i=1}^{n} y_i \log\frac{y_i}{\bar{y}} + \bar{y} - y_i \\ &= 2\sum_{i=1}^{n} y_i \log\frac{y_i}{\bar{y}} \\ &= \sum_{i=1}^{n} r_{\mathrm{D}}(y_i, \bar{y})^2. \end{aligned}$$

# Fraction of Deviance Explained

As with linear models where we use the fraction of variance explaiend ($R^2$), in GLMs we can consider the fraction of *deviance* explained.

Note that the deviance is positive for any model that is not saturated.

The **fraction of deviance explained** is

$$1 - \frac{D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})}{D(\boldsymbol{y}; \bar{y}\mathbf{1})}$$

Unless your model is worse than guessing the mean, the fraction of deviance explained is between 0 and 1.

# Model Comparison

The difference in deviance between two models with predicted means $\hat{\mu}_0$ and $\hat{\mu}_1$, the difference in deviance,

$$D(\boldsymbol{y}; \hat{\mu}_0) - D(\boldsymbol{y}; \hat{\mu}_1)$$

has an approximately chi-squared null distribution. We can use this fact to sequentially add or subtract features from a model depending on whether the change in deviance is significant or not.

# Model Checking

Just like in standard linear models, we should inspect the residuals in generalized linear models for evidence of model misspecification. For example, we can plot the residuals as a function of $\hat{\mu}_i$ and they should be approximately normal at all levels of $\hat{\mu}_i$.

# Cross-Validation

Finally, another common approach to model selection and comparison is to use cross-validation. The idea is to approximate out-of-sample predictive accuracy by randomly splitting the training data.

**Leave-one-out cross validation (LOOCV)** withholds one datapoint out at a time, estimates parameters using the other $n - 1$, and then evaluates predictive likelihood on the held out datapoint.

This approximates,

$$\mathbb{E}_{p^\star(x,y)}[\log p(y \mid x, \{x_i, y_i\}_{i=1}^n)] \approx \frac{1}{n} \sum_{i=1}^{n} \log p(y_i \mid x_i, \{x_j, y_j\}_{j \neq i}). \tag{1}$$

where $p^\star(x, y)$ is the true data generating distribution

For small $n$ (or when using a small number of folds), a bias-correction can be used.

# Conclusion

Generalized linear models allow us to build flexible regression models that respect the domain of the response variable.

Logistic regression is a special case of a Bernoulli GLM with the canonical link function.

For categorical data, we could use a categorical distribution with the softmax link function, and for count data, we could use a Poisson GLM with exponential, softplus, or other link functions.

Leveraging the deviance and deviance residuals for exponential family distribiutions, we can derive analogs of familiar terms from linear modeling, like the fraction of variance explained and the residual analysis.