Statistics 203: Introduction to Regression and Analysis of Variance

Generalized Linear Models I

Jonathan Taylor

Today's class

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Poisson regression.
- Residuals for diagnostics.
- Exponential families.
- Fisher scoring.

Poisson regression

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Response: $Y_i \sim \text{Poisson}(\lambda_i)$ independent.
- Popular applications:
  1. anything with count data
  2. contingency tables
- Link functions:
  1. $g(x) = \log(x)$;
  2. $g(x) = 1/x$
- Variance function: $V(\mu) = \mu$.

Canonical link for Poisson

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- In logistic regression, we identified logit as "canonical" link because

$$g'(\mu) = \frac{1}{V(\mu)}.$$

- We have to solve

$$g'(\mu) = \frac{1}{\mu}.$$

- Therefore, in Poisson regression the canonical link is

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

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

Contingency tables

- An $m \times k$ contingency table is a table of counts categorized by two categorical variables A and B

	B(1)	...	B(k)	
A(1)	Y_{11}	...	Y_{1k}	$Y_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A(n)	Y_{n1}	...	Y_{nk}	$Y_{n\cdot}$
	$Y_{\cdot 1}$	...	$Y_{\cdot k}$	$Y_{\cdot\cdot}$

  where $Y_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$.
- Thinking of categorizing a population of birds, say, by two variables: one being occurence (or not) of West Nile virus, the other, which type of bird.
- Question: is there more West Nile virus in one type of bird? Or, if we sample a bird at random are the events $\{$has West Nile virus$\}$ and $\{$is a bird of type $j\}$ independent?
- Equivalent to

$$H_0 : \mathbb{E}(Y_{ij}) = \lambda_{i,A} \times \lambda_{j,B} \iff \log(\mathbb{E}(Y_{ij})) = \eta_i + \gamma_j$$

Test for independence

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Under H_0

$$\widehat{Y}_{ij} = \frac{Y_{i\cdot}Y_{\cdot j}}{Y_{\cdot\cdot}}$$

- Pearson's X^2

$$X^2 = \sum_{i,j} \frac{\left(Y_{ij} - \widehat{Y}_{ij}\right)^2}{\widehat{Y}_{ij}}.$$

- Under H_0, if λ_{ij}'s are not too small, X^2 is approximately $\chi^2_{nk-n-k+1}$.

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

Poisson deviance

- Easy to see that the Poisson deviance is

$$DEV(\mu|Y) = 2\left(Y \log \frac{Y}{\mu} - (Y - \mu)\right).$$

- Because $\phi = 1$ in a Poisson model, under H_0

$$2\sum_{i,j}\left(Y_{ij} \log \frac{Y_{ij}}{\widehat{Y}_{ij}} - (Y_{ij} - \widehat{Y}_{ij})\right) \sim \chi^2_{nk-k+1}$$

Overdispersed Poisson models

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- If the data truly is Poisson, $\phi = 1$ but *not all* count data is Poisson.
- Example that throw off "Poisson"ness: clustering, i.e. counting disease occurence within families and using many families. If disease has a genetic component then the total number of occurences will be more like a mixture of Poissons.
- How to accomodate this in inference (or test for it)?

$$\widehat{\phi} = \frac{1}{n-p}\sum_{i=1}^{n}\frac{(Y_i - \widehat{\mu}_i)^2}{V(\mu_i)} = \frac{X^2}{n-p}.$$

Residuals

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Pearson's X^2 can be thought of as a sum of "weighted" residuals squared.
- Pearson residual

$$r_{i,P} = \frac{Y_i - \widehat{\mu}_i}{V(\widehat{\mu}_i)}$$

- Deviance in a GLM can be expressed as a sum of n terms:

$$DEV(\widehat{\mu}|Y) = \sum_{i=1}^{n} DEV(\widehat{\mu}_i|Y_i)$$

- Leads to deviance residuals

$$r_{i,DEV} = \text{sign}(Y_i - \widehat{\mu}_i) \cdot \sqrt{DEV(\widehat{\mu}_i|Y_i)}$$

these are the default ones in R.

- Once you have residuals, one can do all the standard diagnostic plots.

Multivariate Newton-Raphson

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Univariate root finding: suppose we want to find a root of $f:\mathbb{R}\to\mathbb{R}$. Start at some x_0,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \qquad n \geq 0.$$

- Multivariate root finding: suppose we want to find roots of $f:\mathbb{R}^p \to \mathbb{R}^p$

$$\{x \in \mathbb{R}^p : f(x) = 0 \in \mathbb{R}^p\}.$$

  Start at some x_0

$$x_{n+1} = x_n - Jf(x_n)^{-1} f(x_n), \qquad n \geq 0.$$

  where Jf is the Jacobean of f.

- Convergence to a solution is guaranteed (at least assuming that all solutions points x^* have $\det(Jf(x^*)) \neq 0$.

Finding critical points

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- Multivariate critical point finding: suppose we want to find critical points of $g : \mathbb{R}^p \to \mathbb{R}$. Start at some x_0

$$x_{n+1} = x_n - \nabla^2 g(x_n)^{-1} \nabla g(x_n), \qquad n \geq 0$$

  where $\nabla^2 g(x)$ is the Hessian of g.
- Convergence to a critical point is guaranteed assuming $\det(\nabla^2 g(x^*)) \neq 0$ for all critical points x^*.
- If g is convex: convergence to global minimizer, otherwise can converge to other critical points.

GLM: Fisher scoring

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- In the Poisson model,

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j} DEV(\widehat{\mu}|Y) &= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} DEV(\widehat{\mu}_i|Y_i) \\
&= -\sum_{i=1}^{n} \frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \frac{\partial \widehat{\mu}_i}{\partial \beta_j} \\
&= \sum_{i=1}^{n} \frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \frac{1}{g'(\widehat{\mu}_i)} X_{ij}
\end{aligned}
$$

GLM: Fisher scoring

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- $$\begin{aligned}
&\frac{\partial^2}{\partial\beta_k\partial\beta_j} DEV(\widehat{\mu}|Y) \\
&\quad = -\sum_{i=1}^n \frac{\partial}{\partial\beta_k}\left(\frac{Y_i-\widehat{\mu}_i}{\widehat{\mu}_i}\frac{1}{g'(\widehat{\mu}_i)}X_{ij}\right) \\
&\quad = \sum_{i=1}^n \Bigg(\frac{1}{\widehat{\mu}_i}\frac{1}{g'(\widehat{\mu}_i)^2}X_{ij}X_{ik} + \\
&\qquad\quad + (Y_i-\widehat{\mu}_i)\left(-\frac{1}{V(\widehat{\mu}_i)^2}\cdot\frac{\partial V(\widehat{\mu}_i)}{\partial\widehat{\mu}_i} + \frac{\partial}{\partial\beta_k}\frac{1}{g'(\widehat{\mu}_i)}X_{ij}\right)\Bigg)
\end{aligned}$$

  Therefore Newton-Raphson needs the extra terms not in our algorithm!

- $$\mathbb{E}_{\beta=\widehat{\beta}}\left(\frac{\partial^2}{\partial\beta_k\partial\beta_j} DEV(\widehat{\mu}|Y)\right) = \sum_{i=1}^n \frac{1}{\widehat{\mu}_i}\frac{1}{g'(\widehat{\mu}_i)^2}X_{ij}X_{ik}$$

Fisher scoring with the canonical link

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- If $g'(\mu) = V(\mu)^{-1} = \mu^{-1}$ in the Poisson case, then

$$
\begin{aligned}
\frac{\partial}{\partial \beta_j} DEV(\widehat{\mu}|Y) &= \sum_{i=1}^n \frac{\partial}{\partial \beta_j} DEV(\widehat{\mu}_i|Y_i) \\
&= -\sum_{i=1}^n \frac{Y_i - \widehat{\mu}_i}{\widehat{\mu}_i} \frac{\partial \widehat{\mu}_i}{\partial \beta_j} \\
&= \sum_{i=1}^n (Y_i - \widehat{\mu}_i) X_{ij}
\end{aligned}
$$

-

$$
\begin{aligned}
&\frac{\partial^2}{\partial \beta_k \partial \beta_j} DEV(\widehat{\mu}|Y) \\
&= \sum_{i=1}^n \frac{1}{g'(\widehat{\mu}_i)} X_{ik} X_{ij} \\
&= \sum_{i=1}^n V(\widehat{\mu}_i) X_{ik} X_{ij}
\end{aligned}
$$

Exponential families

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- We have seen logistic and Poisson regression: when else can we talk about GLMs?
- Suppose the density or mass function of Y

$$f(Y|\theta,\phi) \propto \exp((y\theta - b(\theta))/a(\phi) + c(y,\phi))$$

then

$$\ell(\theta,\phi|Y) = (y\theta - b(\theta))/a(\phi) + c(y,\phi)$$

- Using the facts

$$\mathbb{E}_{(\theta,\phi)}\left(\frac{\partial \ell}{\partial \theta}\right) = 0$$

$$\text{Var}_{(\theta,\phi)}\left(\frac{\partial \ell}{\partial \theta}\right) = -\mathbb{E}_{(\theta,\phi)}\left(\frac{\partial \ell^2}{\partial \theta^2}\right)$$

we see

$$\mathbb{E}(Y) = b'(\theta), \qquad \text{Var}(y) = b''(\theta)a(\phi).$$

- Poisson: $b(\theta) = e^{\theta}$, $a(\phi) = 1$.

Canonical link

- Today's class
- Poisson regression
- Canonical link for Poisson
- Contingency tables
- Test for independence
- Poisson deviance
- Overdispersed Poisson models
- Residuals
- Multivariate Newton-Raphson
- Finding critical points
- GLM: Fisher scoring
- GLM: Fisher scoring
- Fisher scoring with the canonical link
- Exponential families
- Example: Poisson

- The variance function is

$$V(\mu) = b'' \circ b'^{-1}(\mu).$$

- The canonical link is $g = b'{-}1$.
- Why?

$$g'(\mu) = \frac{1}{b''(b'^{-1}(\mu))} = \frac{1}{V(\mu)}$$

- Poisson: $g(x) = \log x$....