MLE with mis-specified model

Ciaran Evans

Warmup

Warmup: Poisson

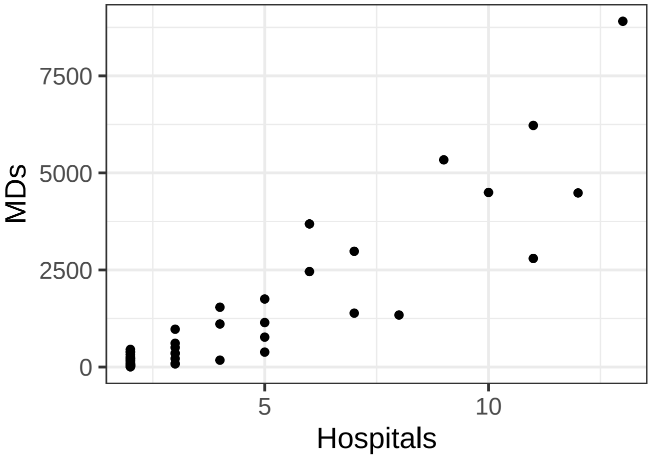
Suppose $Y_1, ..., Y_n \overset{iid}{\sim} Poisson(\lambda)$.

$$\widehat{\lambda} = \overline{Y} \qquad \sqrt{n}(\widehat{\lambda} - \lambda) \xrightarrow{d} N(0, \lambda)$$

Want: variance-stabilizing transformation g such that asymptotic variance of $\sqrt{n}(g(\widehat{\lambda}) - g(\lambda))$ does not depend on λ

Example: non-constant variance

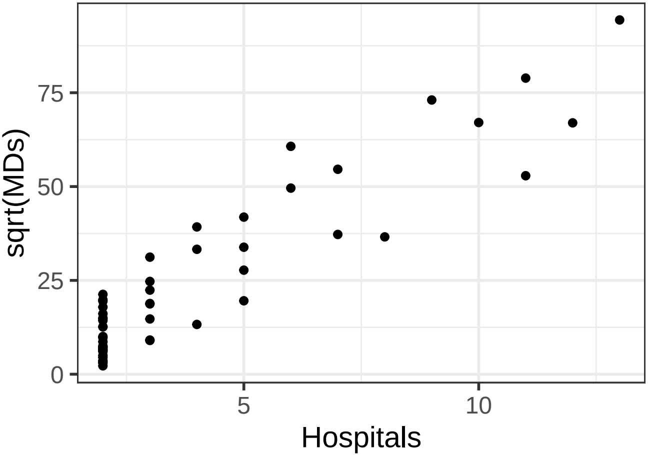
Example: Data on the number of hospitals and number of doctors (MDs) in US counties



Question: How do we adjust for non-constant variance in a linear model?

Example: non-constant variance

Example: Data on the number of hospitals and number of doctors (MDs) in US counties



Variance stablizing transformations

Variance stabilizing transformations are often used when there is a relationship between the *mean* and the *variance* – for example, transforming the response in a linear regression model. Different transformations address different mean-variance relationships. Some examples:

- If $\mathbb{E}[Y] = \mu$ and $Var(Y) = \mu$, then **square root** is a variance stabilizing transformation
- If $\mathbb{E}[Y] = \mu$ and $Var(Y) = \mu^2$, then **log** is a variance stablizing transformation

Why maximum likelihood estimators are nice

Let $\boldsymbol{\theta} \in \mathbb{R}^d$ be a parameter of interest for a distribution with probability function $f(y|\boldsymbol{\theta})$, and $\widehat{\boldsymbol{\theta}}$ the maximum likelihood estimator. Under regularity conditions, the MLE has some very nice properties:

- $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ (consistency)
- $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_1^{-1}(\boldsymbol{\theta}))$ (asymptotic normality)

However: These results assume that we have correctly specified the data distribution $f(y|\boldsymbol{\theta})$! What happens if we were *wrong*?

Revisiting the asymptotic normality proof

Theorem: Let $Y_1, ..., Y_n \overset{iid}{\sim} f(y|\boldsymbol{\theta})$, and let $\widehat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$. Under regularity conditions,

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{d} N(\mathbf{0}, \mathcal{I}_1^{-1}(\boldsymbol{\theta}))$$

Proof sketch:

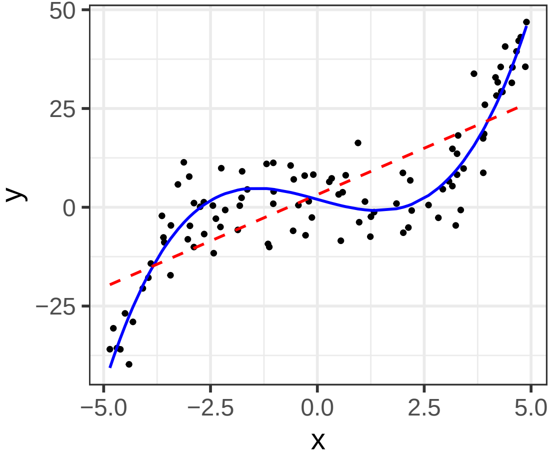
Asymptotics when the distribution is mis-specified?

Setup: Suppose that $Y_1, ..., Y_n \overset{iid}{\sim} g$ for some (unknown) distribution with probability function g. *However*, we instead assume that $Y_i \overset{iid}{\sim} f(y|\boldsymbol{\theta})$ for some specified distribution f with parameter $\boldsymbol{\theta}$.

Example from regression

Truth: (X_i, Y_i) with $Y_i|X_i \sim N(\gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \gamma_3 X_i^3, \sigma^2)$

Assumption: (X_i, Y_i) with $Y_i|X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$



Asymptotics when the distribution is mis-specified?

Setup: Suppose that $Y_1, ..., Y_n \overset{iid}{\sim} g$ for some (unknown) distribution with probability function g. *However*, we instead assume that $Y_i \overset{iid}{\sim} f(y|\boldsymbol{\theta})$ for some specified distribution f with parameter $\boldsymbol{\theta}$.

Asymptotics when the distribution is mis-specified?

Setup: Suppose that $Y_1, \dots, Y_n \overset{iid}{\sim} g$ for some (unknown) distribution with probability function g. *However*, we instead assume that $Y_i \overset{iid}{\sim} f(y|\boldsymbol{\theta})$ for some specified distribution f with parameter $\boldsymbol{\theta}$.

- Let $\boldsymbol{\theta}_0$ solve $\mathbb{E}_g\left[\frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y|\boldsymbol{\theta})\right] = \mathbf{0}$.
- Let $\widehat{\boldsymbol{\theta}}$ solve $\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(Y_i|\boldsymbol{\theta}) = \mathbf{0}$

Key results: Under regularity conditions,

- $\widehat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0$
- $\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{S}(\boldsymbol{\theta}_0))$ where

$$\mathbf{S}(\boldsymbol{\theta}_0) = \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1}$$

$$\mathbf{A} = -\mathbb{E}_g\left[\frac{\partial^2}{\partial \boldsymbol{\theta}_0^2} \log f(Y_i|\boldsymbol{\theta}_0)\right] \qquad \mathbf{B} = Var_g\left(\frac{\partial}{\partial \boldsymbol{\theta}_0} \log f(Y_i|\boldsymbol{\theta}_0)\right)$$

Example: linear regression

Assumed model: $Y_i | \mathbf{x}_i \sim N(\mathbf{x}_i^T \beta, \sigma^2)$