# Maximum Likelihood Estimation

**Simon Wood**, University of Edinburgh, U.K.

# Some models

- *Logistic regression:* $y_i \sim \text{bernouilli}(\mu_i)$ where

$$\log\{\mu_i/(1-\mu_i)\} = \eta_i \equiv \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \dots$$

- *Poisson regression:* $y_i \underset{\text{ind}}{\sim} \text{Poi}(\mu_i)$ where

$$\log(\mu_i) = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \dots$$

- *Linear Mixed Model:*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon} \text{ where } \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_\gamma) \text{ and } \boldsymbol{\epsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$$

  so $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{Z}\psi_\gamma\mathbf{Z}^\mathsf{T} + \mathbf{I}\sigma^2)$.

- What general method can we use to estimate parameters of these model, and others? Least squared no longer the best option.

# Maximum Likelihood Estimation

- Preceding models all specify a p.d.f. $\pi_\theta(\mathbf{y})$ for data vector $\mathbf{y}$.
- $\boldsymbol{\theta}$ is an unknown parameter vector determining the shape of $\pi_\theta$.
- The *Likelihood* of $\boldsymbol{\theta}$ is $\pi_\theta(\mathbf{y})$ considered as a function of $\boldsymbol{\theta}$ with the observed $\mathbf{y}$ plugged in. $L(\boldsymbol{\theta}) = \pi_\theta(\mathbf{y}_{\text{obs.}})$.
- $\boldsymbol{\theta}$ values that make the observed data appear probable are more *likely* than values that make it appear improbable.
- The *log likelihood* is $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.
- The *Maximum Likelihood Estimator* (MLE) is

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmax}}\, l(\boldsymbol{\theta}).$$

- Generally we need numerical optimization to find $\hat{\boldsymbol{\theta}}$.

# MLE properties

- If $n = \dim(\mathbf{y}) \to \infty$ and $l$ is sufficiently regular

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_t, \hat{\boldsymbol{\mathcal{I}}}^{-1})$$

  where $\hat{\boldsymbol{\mathcal{I}}}$ is the Hessian of the negative log likelihood at the MLE ($\hat{\mathcal{I}}_{ij} = -\partial^2 l / \partial\theta_i \partial\theta_j$), and the true parameter value is $\boldsymbol{\theta}_t$.
- Let $\hat{\boldsymbol{\theta}}_0$ be the MLE under $r$ restrictions defining a hypothesis $H_0 : R(\boldsymbol{\theta}) = \mathbf{0}$. If $H_0$ is true, then for regular $l$ and $n \to \infty$

$$2\{l(\hat{\boldsymbol{\theta}}) - l(\hat{\boldsymbol{\theta}}_0)\} \sim \chi^2_r$$

  This is the basis of the *generalized likelihood ratio test* (GLRT) of $H_0$ versus $H_1 : R(\boldsymbol{\theta}) \neq \mathbf{0}$
- Note the generality. If we have a computable likelihood we can use this theory, provided we can maximize the likelihood.

# Programming *log* likelihoods

- The large sample MLE results relate to the log likelihood.
- It is usually much better to optimize the log likelihood than the likelihood, as the likelihood may easily underflow to zero.
- In R the built in densities usually allow you to compute directly on the log probability scale.
- Never compute the likelihood and the take its log!

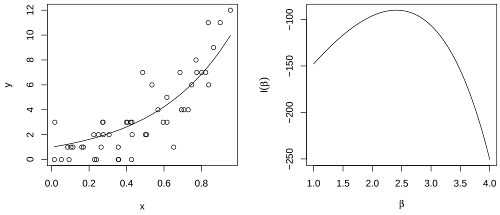
```
> log(prod(dnorm(x,2,2)))  ## 100 obs in x
[1] -219.7226
> sum(dnorm(x,2,2,log=TRUE)) ## stable version
[1] -219.7226
>
> log(prod(dnorm(x,2,2))) ## 400 obs in x - problem!
[1] -Inf
> sum(dnorm(x,2,2,log=TRUE)) ## stable version
[1] -888.4871
```

# Simple one parameter example

- Model: $y_i \sim \text{Poi}\{\exp(\beta x_i)\}$ (independent).
- Poisson p.f. is $\pi(y_i) = \lambda_i^{y_i} \exp(-\lambda_i)/y_i!$ and here $\lambda_i = \exp(\beta x_i)$.
- The log likelihood is therefore

$$l(\beta) = \sum_{i=1}^{n} y_i \beta x_i - \exp(\beta x_i) - \log y_i!$$

- Left is $x_i, y_i$ and fit. Right is $l(\beta)$.



# What distribution of $\hat{\beta}$ means

- Grey are replicate $l(\beta)$ curves for replicate sets of $x_i, y_i$ data.
- Black dots and ticks show MLEs for each. Kernel density estimate the $\hat{\beta}$ distribution. $\beta = 2.5$ was truth here.