STA410 | Programming Portfolio Assignment 3

3.0 - Nonlinearity s| $b=f(x)\not=Ax$

A more interesting problem than $Ax=b$ is the problem $g(x)=b$ for a ***nonlinear function*** g. Such ***nonlinear equations*** are a general version form of a wide set of problems. However, a most frequently encountered instance of the problem is

$$f'(z) = 0 \quad \text{ or more generally } \quad \underbrace{\nabla_z f(z) = \mathbf{0}}_{\text{multivariate form}}$$

because solutions to these equations are ***local minima*** or ***maxima*** of $$f$$ and ***model fitting*** can be framed as a subclass of this general ***optimization problem.***

Two important notes about these ***optimization problems*** can be made.

1. The ***derivative*** $$f'(z)$$ is just some other function, say $$g(z)$$, so outside of its useful interpretation as a derivative, $$f'(z)$$ may be treated just as any other function might be treated.

> E.g., the ***first-order Taylor series approximation*** of $$f'$$ is simply
>
> $$f'(x) \approx f'(x_0) + (x-x_0)f''(x_0)$$
>
> as if $$f'$$ had been replaced with some other function $$g$$.

2. ***Optimization solutions*** $f'(x^*)=0$ are found within regions of curvature of f, while ***roots*** $f(x_0)=0$ need not be.

- The behavior of a function near a ***solution*** to an ***optimization*** problem differs, generally speaking, from it's behavior near a ***root***.
- And the ***numerical precision*** in an ***optimization*** context to in general differs from the vanilla ***root-finding*** context.

> E.g., to the degree that a ***second order Taylor series approximation*** $g_{\tilde x}(x)$ of $f(x)$ is accurate
>
> $$f(x) \approx g_{\tilde x}(x) = f(\tilde x) + (x - \tilde x) f'(\tilde x) + f''(\tilde x)\frac{(x-\tilde x)^2}{2} $$
>
> changes in $g_{\tilde x}(x)$ are likely dominated by
>
> - the linear term $(\underbrace{x-\tilde x}*{\epsilon*{machine}})f'(\tilde x) \longrightarrow$ if $\tilde x$ is near ***root*** x_0
> - the quadratic term $\frac{1}{2}(\underbrace{x-\tilde x}*{\sqrt{\epsilon*{machine}}})^2f''(\tilde x) \longrightarrow$ if $\tilde x$ is near ***optimum*** x^*, since $f'(\tilde x \approx x^*)\approx 0$.
>
> I.e., $f(x) \approx g_{x^*}(x)$ evaluated near an ***optimization problem solution*** x^* only supports about half of the ***numerical accuracy***

> > i.e., the square root of the available precision $\epsilon_{machine}$, since the difference will be squared
>
> compared to evaluating $f(x) \approx g_{x_0}(x)$ near one of its ***roots*** x_0.

\newpage

3.0.1 Maximum Likelihood Estimates (MLEs)

The ***score function*** is the gradient of the ***log likelihood***

$$\nabla_\theta l(\theta') = \left(\frac{\partial l(\theta')}{\partial \theta_1}, \cdots, \frac{\partial l(\theta')}{\partial \theta_p} \right)^T \quad \text{ where } \quad l(\theta) = \log f(x|\theta) \overset{iid}{=} \log \prod_{i=1}^n f(x_i|\theta)$$

Maximum Likelihood Estimates (MLEs) come from the (***nonlinear***) ***score equation*** which sets the ***score function*** equal to $\mathbf{0}$. And for the ***true value*** of the parameter θ^{true}, the ***score function*** has expected value $\mathbf{0}$ (with respect to f_x the distribution of the data). So

$$\underbrace{\nabla_\theta l(\hat \theta) = \mathbf{0}}\text{score equation} \quad \text{ and } \quad E_X !\underbrace{\left[\nabla\theta l(\theta^{\text{true}})\right]}_{\text{score function}} != \mathbf{0}$$

> The expected value of the ***score function*** follows since
>
> \begin{align*} E \left[\nabla_\theta l(\theta)\right]
> = {} & \int \nabla_\theta l(\theta) f(x|\theta) dx \ = {} & \int \left(\frac{\partial l(\theta)}{\partial \theta_1}, \cdots, \frac{\partial l(\theta)}{\partial \theta_p} \right)^T f(x|\theta) dx \ = {} & \int \left(\frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial \theta_1}, \cdots, \frac{1}{f(x|\theta)}\frac{\partial f(x|\theta)}{\partial \theta_p} \right)^T f(x|\theta) dx \ = {} & \int \left(\frac{\partial f(x|\theta)}{\partial \theta_1}, \cdots, \frac{\partial f(x|\theta)}{\partial \theta_p} \right)^T dx \ = {} &\int \nabla_\theta f(x|\theta) dx \ = {} & \nabla_\theta \int f(x|\theta) dx = \nabla_\theta , 1 = \mathbf{0} \ \end{align*}

The ***Fisher information matrix*** $I(\theta^{\text{true}})$, or ***expected Fisher information matrix*** is the expected value of the ***outer product*** of the ***score function*** with itself and is equal to the expected value of the negative of the ***Hessian*** of the log likelihood (all with respect to f_x the distribution of the data), i.e., I $$I(\theta^{\text{true}}) = E_X[\nabla_\theta l(\theta^{\text{true}})(\nabla_\theta l(\theta^{\text{true}})^T] = E_X[-H_{l(\theta)}(\theta^{\text{true}})]$$

The ***observed Fisher information*** is

$$\begin{align*} \hat{I(\theta)} = {} & -H_{l(\theta)}(\hat \theta) = \underbrace{-J(\nabla_\theta l)(\hat \theta) \approx - \sum_{i=1}^n J(\nabla_\theta log f(x_i|\theta))\big|{\hat \theta}}{\text{$J(\nabla_\theta l) (\hat \theta)$ }\textit{Jacobian}\text{ of }\textit{score function} \text{ evaluated at } \hat \theta} \ \text{ or } \quad \hat{I(\theta)} \approx {} & \sum_{i=1}^n \nabla_\theta log f(x_i|\theta)\big|{\hat \theta}\left(\nabla\theta log f(x_i| \theta)\big|_{\hat \theta}\right)^T \end{align*}$$

And the ***asymptotic distribution*** of the MLE is

$$ p(\hat \theta) \overset{n \rightarrow \infty}{\longrightarrow} N!\left(\theta^{\text{true}}, \Sigma = \frac{I(\theta^{\text{true}})^{-1}}{n}\right) \approx N!\left(\theta^{\text{true}}, \Sigma = \frac{\hat{I(\theta)}

{}^{-1}}{n} \approx \frac{I(\theta^{\text{true}})^{-1}}{n}\right)$$

where either the ***observed Fisher information*** or its approximation based on the ***outer product*** of the ***score function*** with itself may be used as plug in estimates for the ***expected Fisher information matrix*** $I(\theta^{\text{true}})$.