Solutions to Problem Set: ML Methods

1. No. We know that $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta} \in \Theta} \mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\theta)],$$

where the expectation is over the joint distribution of $(\mathbf{x}_i, \mathbf{y}_i)$. Therefore, because $\exp(\cdot)$ is an increasing function, $\boldsymbol{\theta}_o$ also maximizes $\exp\{\mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$ over $\boldsymbol{\Theta}$. The problem is that the expectation and the exponential function cannot be interchanged: $\mathrm{E}[f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})] \neq \exp\{\mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$. In fact, Jensen's inequality tells us that

$$\mathrm{E}[f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})] > \exp\{\mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$$

2. (a) Because

$$f(y|\mathbf{x}_i) = (2\pi\sigma_o^2)^{-1/2}\exp[-(y - m(\mathbf{x}_i, \boldsymbol{\beta}_o))^2/(2\sigma_o^2)],$$

it follows that for observation i the log likelihood is

$$\ell_i(\boldsymbol{\beta},\sigma^2) = -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}[y_i - m(\mathbf{x}_i,\boldsymbol{\beta})]^2.$$

Only the last of these terms depends on $\boldsymbol{\beta}$. Further, for any $\sigma^2 > 0$, maximizing $\sum_{i=1}^N \ell_i(\boldsymbol{\beta},\sigma^2)$ with respect to $\boldsymbol{\beta}$ is the same as minimizing

$$\sum_{i=1}^{N}[y_i - m(\mathbf{x}_i,\boldsymbol{\beta})]^2,$$

which means the MLE $\hat{\boldsymbol{\beta}}$ is the NLS estimator.

(b) First,

$$\nabla_{\boldsymbol{\beta}}\ell_i(\boldsymbol{\beta},\sigma^2) = \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i,\boldsymbol{\beta})[y_i - m(\mathbf{x}_i,\boldsymbol{\beta})]/\sigma^2;$$

note that $\nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i,\boldsymbol{\beta})$ is $1 \times P$. Next,

$$\frac{\partial \ell_i(\boldsymbol{\beta},\sigma^2)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}[y_i - m(\mathbf{x}_i,\boldsymbol{\beta})]^2.$$

For notational simplicity, define the residual function $u_i(\boldsymbol{\beta}) \equiv y_i - m(\mathbf{x}_i,\boldsymbol{\beta})$. Then the score is

$$\mathbf{s}_i(\boldsymbol{\theta}) = \begin{pmatrix} \nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})' u_i(\boldsymbol{\beta})/\sigma^2 \\ -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}[u_i(\boldsymbol{\beta})]^2 \end{pmatrix},$$

where $\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) \equiv \nabla_{\boldsymbol{\beta}} m(\mathbf{x}_i,\boldsymbol{\beta})$.

Define the errors as $u_i \equiv u_i(\boldsymbol{\beta}_o)$, so that $\mathrm{E}(u_i|\mathbf{x}_i) = 0$ and $\mathrm{E}(u_i^2|\mathbf{x}_i) = \mathrm{Var}(y_i|\mathbf{x}_i) = \sigma_o^2$. Then, since $\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)$ is a function of $\mathbf{x}_i$, it is easily seen that $\mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i] = \mathbf{0}$. Note that we only use the fact that $\mathrm{E}(y_i|\mathbf{x}_i) = m(\mathbf{x}_i,\boldsymbol{\beta}_o)$ and $\mathrm{Var}(y_i|\mathbf{x}_i) = \sigma_o^2$ in showing this. In other words, only the first two conditional moments of y_i need to be correctly specified; nothing else about the normal distribution is used.

(c) The equation used to obtain $\hat{\sigma}^2$ is

$$\sum_{i=1}^{N}\left(-\frac{1}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}[y_i - m(\mathbf{x}_i,\hat{\boldsymbol{\beta}})]^2\right) = 0,$$

where $\hat{\boldsymbol{\beta}}$ is the nonlinear least squares estimator. Solving gives

$$\hat{\sigma}^2 = N^{-1}\sum_{i=1}^{N}\hat{u}_i^2,$$

where $\hat{u}_i \equiv y_i - m(\mathbf{x}_i,\hat{\boldsymbol{\beta}})$. Thus, the MLE of σ^2 is the sum of squared residuals divided by N. In practice, N is often replaced with $N-P$ as a degrees-of-freedom adjustment, but this makes no difference as $N \to \infty$.

(d) The derivations are a bit tedious but fairly straightforward:

$$
\mathbf{H}_i(\theta) = \begin{pmatrix} -\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})/\sigma^2 + \nabla^2_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) u_i(\boldsymbol{\beta})/\sigma^2 & -\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})' u_i(\boldsymbol{\beta})/\sigma^4 \\ -\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}) u_i(\boldsymbol{\beta})/\sigma^4 & \frac{1}{2\sigma^4} - \frac{1}{\sigma^6}[u_i(\boldsymbol{\beta})]^2 \end{pmatrix},
$$

where $\nabla^2_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta})$ is the $P \times P$ Hessian of $m_i(\boldsymbol{\beta})$.

independence conditional on X.

(d) From part c and $E(u_i|\mathbf{x}_i) = 0$, the off-diagonal blocks are zero. Further,

$$
E[\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)/\sigma_o^2 - \nabla^2_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o) u_i/\sigma_o^2|\mathbf{x}_i] = \nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)/\sigma_o^2
$$

Because , $E(u_i^2|\mathbf{x}_i) = \sigma_o^2$,

$$
E\left(\frac{1}{\sigma_o^6} u_i^2 - \frac{1}{2\sigma_o^4}\middle|\mathbf{x}_i\right) = \frac{1}{\sigma_o^4} - \frac{1}{2\sigma_o^4} = \frac{1}{2\sigma_o^4}.
$$

Therefore,

$$
-E[\mathbf{H}_i(\theta_o)|\mathbf{x}_i] = \begin{pmatrix} \nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)/\sigma_o^2 & \mathbf{0} \\ \mathbf{0} & \frac{1}{2\sigma_o^4} \end{pmatrix} \qquad (1)
$$

where we again use $E(u_i|\mathbf{x}_i) = 0$ and $E(u_i^2|\mathbf{x}_i) = \sigma_o^2$.

(e) To show that $-E[\mathbf{H}_i(\theta_o)|\mathbf{x}_i]$ equals $E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i]$, we need to know that, with u_i defined as above, $E(u_i^3|\mathbf{x}_i) = 0$, which can be used, along with the zero mean and constant (normal distribution not skewed) conditional variance, to show

$$
E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i] = \begin{pmatrix} \nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)/\sigma_o^2 & \mathbf{0} \\ \mathbf{0} & E\left[\left(-\frac{1}{2\sigma_o^2} + \frac{1}{2\sigma_o^4}u_i^2\right)^2\right] \end{pmatrix}.
$$

Further, $E(u_i^4|\mathbf{x}_i) = 3\sigma_o^4$, and so

$$
E\left[\left(-\frac{1}{2\sigma_o^2} + \frac{1}{2\sigma_o^4}u_i^2\right)^2\right] = \frac{1}{4\sigma_o^4} + \frac{3\sigma_o^4}{4\sigma_o^8} - \frac{2\sigma_o^2}{4\sigma_o^6} = \frac{1}{2\sigma_o^4}.
$$

Thus, we have shown $-E[\mathbf{H}_i(\theta_o)|\mathbf{x}_i] = E[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i]$.

(f). From general MLE, we know that Avar$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o)$ is the $P \times P$ upper left hand block of $\{E[\mathbf{A}_i(\boldsymbol{\theta}_o)]\}^{-1}$, where $\mathbf{A}_i(\boldsymbol{\theta}_o)$ is the matrix in (1). Because this matrix is block diagonal, it is easily seen that

$$\text{Avar}\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) = \sigma_o^2\{E[\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)'\nabla_{\boldsymbol{\beta}} m_i(\boldsymbol{\beta}_o)]\}^{-1},$$

and this is consistently estimated by

$$\hat{\sigma}^2\left(N^{-1}\sum_{i=1}^{N}\nabla_{\boldsymbol{\beta}}\hat{m}_i'\nabla_{\boldsymbol{\beta}}\hat{m}_i\right)^{-1}, \qquad (2)$$

which means that $\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}})$ is (2) divided by N, or

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) = \hat{\sigma}^2\left(\sum_{i=1}^{N}\nabla_{\boldsymbol{\beta}}\hat{m}_i'\nabla_{\boldsymbol{\beta}}\hat{m}_i\right)^{-1}.$$

If the model is linear, $\nabla_{\boldsymbol{\beta}}\hat{m}_i = \mathbf{x}_i$, and we obtain exactly the asymptotic variance estimator for the OLS estimator under homoskedasticity.