Solutions to MATH3091 problem sheet 2

18 Feb 2022

1. We have $\boldsymbol{\theta} = (\beta_0, \beta_1, \sigma^2)$, and

$$f_{Y_i}(y_i; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

(the $N(\beta_0 + \beta_1 x_i, \sigma^2)$ p.d.f.), so the likelihood is

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^n f_Y(y_i; \boldsymbol{\theta}) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \beta_0 - \beta_1 x_i)^2\right\} \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2\right\} \end{aligned}$$

The log-likelihood is

$$\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Differentiating with respect to each component of $\boldsymbol{\theta}$, we have

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_0} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i),$$

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_1} = \frac{1}{\sigma^2}\sum_{i=1}^n x_i(y_i - \beta_0 - \beta_1 x_i),$$

and

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

So a stationary point $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\sigma}^2)$ solves

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{1}$$

$$\frac{1}{\hat{\sigma}^2}\sum_{i=1}^n x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0, \tag{2}$$

and

$$-\frac{n}{2\hat{\sigma}^2}+\frac{1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2=0 \tag{3}$$

Rearraging (1)

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}.$$

Rearranging (2) gives

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-n\hat{\beta}_0\bar{x}}{\sum_{i=1}^{n}x_i^2}$$

Substituting $\hat{\beta}_0$, we get

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-n(\bar{y}-\hat{\beta}_1\bar{x})\bar{x}}{\sum_{i=1}^{n}x_i^2},$$

and rearranging for $\hat{\beta}_1$ gives

$$\hat{\beta}_1=\frac{\sum_{i=1}^{n}x_iy_i-n\bar{x}\bar{y}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2}.$$

(b) Rearranging (3), and pulg-in the MLE $\hat{\beta}_0,\hat{\beta}_1$ gives

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2.$$

For (a) and (b), we assume that this stationary point is a maximum, therefore the MLEs are just $(\hat{\beta}_0,\hat{\beta}_1,\hat{\sigma}^2)$.

(c) The MLE of σ is

$$\hat{\sigma}=\sqrt{\hat{\sigma}^2}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)^2}.$$

(d) In this case, let $\boldsymbol{\theta}=(\boldsymbol{\beta}^T,\sigma^2)^T$, we know the likelihood is

$$\begin{aligned}L(\boldsymbol{\theta})&=f_{\boldsymbol{Y}}(\boldsymbol{y};\boldsymbol{\theta})\\&=(2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right],\end{aligned}$$

therefore

$$\ell(\boldsymbol{\theta})=\log L(\boldsymbol{\theta})=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}).$$

Differentiating with respect to each component of $\boldsymbol{\theta}$, we have

$$\frac{\partial\ell(\boldsymbol{\theta})}{\partial\boldsymbol{\beta}}=\frac{1}{\sigma^2}\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}),$$

and

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$$

So a stationary point $\hat{\boldsymbol{\beta}}$ solves

$$\frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma^2} = \boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}),$$

which suggests

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Moreover, So a stationary point $\hat{\sigma}^2$ solves

$$-\frac{n}{2\hat{\sigma}^2} + \frac{1}{2(\hat{\sigma}^2)^2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = 0,$$

which suggests

$$\hat{\sigma}^2 = \frac{1}{n}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})$$

.

(e) Note that as long as $E(Y_i) = \boldsymbol{\beta}^T\boldsymbol{x}_i$, we have

$$E(\boldsymbol{Y}) = E[(Y_1, \ldots, Y_n)^T] = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ is the $n \times p$ matrix of explanatory variables. As a result,

$$E(\hat{\boldsymbol{\beta}}) = E(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}X^TE(\boldsymbol{Y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{\beta} = \boldsymbol{\beta}$$

which suggests $\hat{\boldsymbol{\beta}}$ is an unbiased estimator.

None of the assumptions are required. We only need $E(\epsilon_i) = 0, i = 1\ldots, n$.

(f) Note that $\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = (I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{Y}$, we denote

$$I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = I - P = H,$$

where note that $P^T = P$ and $PP = P$. As a result,

$$\begin{aligned}
(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) =& \boldsymbol{Y}^T(I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)(I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{Y} \\
=& \boldsymbol{Y}^T(I - 2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T + \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{Y} \\
=& \boldsymbol{Y}^T(I - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T)\boldsymbol{Y} \\
=& \boldsymbol{Y}^TH\boldsymbol{Y}.
\end{aligned}$$

Note that $H\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, combined with the fact that $\text{tr}(AB) = \text{tr}(BA)$ we have

$$\begin{aligned}
E(\boldsymbol{Y}^T H \boldsymbol{Y}) &= E\left[(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})^T H (\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\right] \\
&= E(\boldsymbol{\epsilon}^T H \boldsymbol{\epsilon}) \\
&= E(\text{tr}(\boldsymbol{\epsilon}^T H \boldsymbol{\epsilon})) \\
&= E(\text{tr}(H \boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)) \\
&= \text{tr}(H E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)) \\
&= \text{tr}(H\sigma^2 \boldsymbol{I}) \\
&= \sigma^2 \text{tr}(H).
\end{aligned}$$

In addition, we have

$$\begin{aligned}
\text{tr}(H) =& \text{tr}(I_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\
=& \text{tr}(I_n) - \text{tr}(\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T) \\
=& n - \text{tr}((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}) \\
=& n - \text{tr}(I_p) \\
=& n - p
\end{aligned}$$

Hence,

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})\right] = \frac{n-p}{n}\sigma^2,$$

which means the MLE $\hat{\sigma}^2$ is a biased estimator of σ^2.

An unbiased estimator of σ^2 is

$$\tilde{\sigma}^2 = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}).$$

2. (a) We have $p = 2$ and $q = 1$, so

$$F = \frac{(D_0 - D_1)/(p-q)}{D_1/(n-p)} = \frac{(13-12)/(2-1)}{12/(50-2)} = \frac{1}{12/48} = 4.$$

Under H_0, $F \sim F_{p-q,n-p} = F_{1,48}$. For a size $\alpha = 0.05$ test, we reject H_0 if $F > k$, where k is the 95% point of the $F_{1,48}$ distribution, which equals to 4.042652 as calculated by `R`, or 4.04, to two decimal places. So we do not reject H_0.

(b) From the notes, for $j = 1, 2$, we have

$$\max_{\beta,\sigma^2 \in \Theta^{(j)}} L(\boldsymbol{\beta}, \sigma^2) = (2\pi D_j/n)^{-n/2} \exp\left(-n/2\right).$$

So

$$\begin{aligned}
L_{01} &= 2\log\frac{\max_{\beta,\sigma^2\in\Theta^{(1)}} L(\boldsymbol{\beta},\sigma^2)}{\max_{\beta,\sigma^2\in\Theta^{(0)}} L(\boldsymbol{\beta},\sigma^2)} \\
&= 2\log\frac{(2\pi D_1/n)^{-n/2}\exp(-n/2)}{(2\pi D_0/n)^{-n/2}\exp(-n/2)} \\
&= 2\log\left(\frac{D_0}{D_1}\right)^{\frac{n}{2}} \\
&= n\log\frac{D_0}{D_1} \\
&= 50\log\frac{13}{12} \\
&= 4.00 \text{ (2 d.p).}
\end{aligned}$$

Under H_0, $L_{01} \sim \chi^2_1$. For a test of approximate size α, we reject H_0 if $L_{01} > k$, where k is the 95% point of the χ^2_1 distribution, or $\chi^2_{1,0.95}$ which equals to 3.841459, or 3.84, to two decimal places. So we reject H_0.

(c) For the F test, we did not reject H_0, while for the log likelihood ratio test, we did reject H_0. In both cases, the value of the test statistic was quite close to the critical value. The two tests differ because the log likelihood ratio test statistic L_{01} has approximately (not exactly) χ^2_1 distribution under H_0, whereas F has exactly $F_{1,48}$ distribution. Because the approximate distribution for L_{01} is based on an asymptotic result (valid as $n \to \infty$), the two tests will be very similar for large n.

3. (a) Since both F-test and likelihood ratio tests can only applied for testing nested (restricted) models, they are not necessarily suitable for this problem.

If $H_0 : \boldsymbol{C\beta} = 0$ is true, then $\boldsymbol{C\hat{\beta}} \sim N(\mathbf{0}, \sigma^2\boldsymbol{C}(X^TX)^{-1}\boldsymbol{C}^T)$, which further leads to

$$\hat{\boldsymbol{\beta}}^T\boldsymbol{C}^T\left[\sigma^2\boldsymbol{C}(X^TX)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}} \sim \chi^2_m,$$

if matrix $\boldsymbol{C}$ is of rank m.

However, we often do not know σ^2 in practice. This suggests the following Wald test statistic

$$W = \hat{\boldsymbol{\beta}}^T\boldsymbol{C}^T\left[\tilde{\sigma}^2\boldsymbol{C}(X^TX)^{-1}\boldsymbol{C}^T\right]^{-1}\boldsymbol{C}\hat{\boldsymbol{\beta}}$$

(b) In large samples $n \to \infty$, we have $\tilde{\sigma}^2 \to \sigma^2$. As a result, the asymptotic distribution of the Wald test statistic W is χ^2_m.