Statistical Modeling SDS 383D: Excercise 4

Khai Nguyen - kbn397
April 24, 2022

Math Tests :
(A) We have the log likelihood function

$$L(\theta_i) \propto -\frac{1}{2\sigma^2}(\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2)$$

Hence the derivative is

$$\frac{\partial}{\partial\theta_i}L(\theta_i) = \frac{1}{\sigma^2}\sum_{j=1}^{N_i}(y_{ij}-\theta_i)$$

Setting the derivative to 0, we have

$$\hat{\theta}_i = \frac{\sum_{j=1}^{N_i} y_{ij}}{N_i}$$

(B) Explanation: Fewer data samples means higher variance.

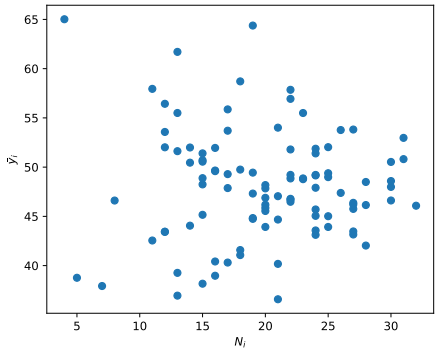

Figure 1:

(C) We have the joint posterior is

$$p(\boldsymbol{\theta},\sigma^2,\tau^2|\boldsymbol{y}) \propto 1\cdot\frac{1}{\sigma^2}\cdot\frac{1}{\tau^2}^{\frac{1}{2}+1}\exp\left(-\frac{1}{2\tau^2}\right)\prod_{i=1}^{P}(\tau^2\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i-\mu)^2\right)$$
$$\prod_{j=1}^{N_i}(\sigma^2)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right)$$

Therefore, we have the conditional posteriors:

$$\begin{aligned}
p(\theta_i|-) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i-\mu)^2\right)\prod_{j=1}^{N_i}\exp\left(-\frac{1}{2\sigma^2}(y_{ij}-\theta_i)^2\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i^2-2\theta_i\mu)\right)\exp\left(-\frac{1}{2\sigma^2}(N_i\theta_i^2-2\theta_i\sum_{j=1}^{N_i}y_{ij})\right)\\
&\propto \exp\left(-\frac{1}{2}\left(\theta_i^2\left(\frac{1}{\sigma^2\tau^2}+\frac{N_i}{\sigma^2}\right)-2\theta_i\left(\frac{\mu}{\sigma^2\tau^2}+\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2}\right)\right)\right)\\
&= \mathcal{N}\left(\left(\frac{1}{\sigma^2\tau^2}+\frac{N_i}{\sigma^2}\right)^{-1}\left(\frac{\mu}{\sigma^2\tau^2}+\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2}\right),\left(\frac{1}{\sigma^2\tau^2}+\frac{N_i}{\sigma^2}\right)^{-1}\right)
\end{aligned}$$

and

$$\begin{aligned}
p(\sigma^2|-) &= (\sigma^2)^{-1-\frac{P}{2}-\frac{\sum_{i=1}^{P}N_i}{2}}\exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\tau^2}+\sum_{i=1}^{P}\frac{\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2}{2}\right)\right)\\
&\quad Inv-Gamma\left(\frac{P+\sum_{i=1}^{P}N_i}{2},\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\tau^2}+\sum_{i=1}^{P}\frac{\sum_{j=1}^{N_i}(y_{ij}-\theta_i)^2}{2}\right)\right)
\end{aligned}$$

and

$$\begin{aligned}
p(\tau^2|-) &= (\tau^2)^{-1-\frac{1}{2}-\frac{P}{2}}\exp\left(-\frac{1}{\tau^2}\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right)\right)\\
&= Inv-gamma\left(\frac{P+1}{2},\frac{\sum_{i=1}^{P}(\theta_i-\mu)^2}{2\sigma^2}+\frac{1}{2}\right)
\end{aligned}$$

and

$$\begin{aligned}
p(\mu|-) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\theta_i-\mu)^2\right)\\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\mu^2-2\mu\theta_i+\theta_i^2)\right)\\
&\propto \exp\left(-\frac{P}{2\sigma^2\tau^2}(\mu^2-2\mu\sum_{i=1}^{P}\theta_i\frac{1}{P}))\right)\\
&= \mathcal{N}\left(\frac{\sum_{i=1}^{P}\theta_i}{P},\frac{\sigma^2\tau^2}{P}\right)
\end{aligned}$$

(D) We have the posterior mean

$$\begin{aligned}
&= \left(\frac{1}{\sigma^2\tau^2}+\frac{N_i}{\sigma^2}\right)^{-1}\left(\frac{\mu}{\sigma^2\tau^2}+\frac{\sum_{j=1}^{N_i}y_{ij}}{\sigma^2}\right)\\
&= \kappa_i\mu+(1-\kappa_i)\bar{y}_i
\end{aligned}$$

where

$$\kappa_i = \left(1 + N_i\tau^2\right)^{-1}$$
$$1 - \kappa_i = 1 - \frac{1}{1 + N_i\tau^2} = \frac{N_i\tau^2}{1 + N_i\tau^2}$$

Explanation: Deceasing

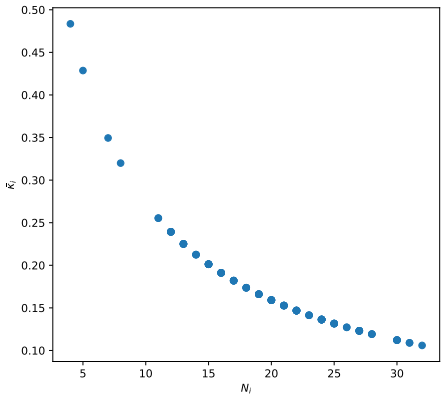

Figure 2: N_i vs $\bar{K}_i$.

(E) We have

$$\begin{aligned} Cov[y_{ij}, y_{ik}] &= E[(\mu + \delta_i + e_{ij} - E[\mu + \delta_i + e_{ij}])(\mu + \delta_i + e_{ik} - E[\mu + \delta_i + e_{ik}]) \\ &= E[\delta_i^2 + \delta_i(e_{ij} + e_{ik}) + e_{ij}e_{i'k}] = \tau^2\sigma^2 \\ Cov[y_{ij}, y_{ik}] &= E[(\mu + \delta_i + e_{ij} - E[\mu + \delta_i + e_{ij}])(\mu + \delta_{i'} + e_{i'k} - E[\mu + \delta_i + e_{ik}]) \\ &= E[\delta_i\delta_{i'} + \delta_i e_{i'k} + \delta_{i'}e_{ij} + e_{ij}e_{i'k}] = 0 \end{aligned}$$

It is reasonable that test scores for students in the same school are expected to be correlated. Also, test scores for students at different schools arr expected to be independent.
(F) I believe that the assumption is reasonably since we has τ for controlling in-school variability while variability between schools is modeled with σ^2.

Blood Pressure :
(A) I got the difference between these two group means is 9.427 with a standard error of 1.004. However, the assumption of the t-test is not reasonable in this scenario since the t-test assumes that observations within groups are independent. From the data, we know that this assumption clearly wrong since we have repeated measurements of the same person.
(B) I got the difference between these two group means is 7.416 with a standard error of 4.511. While this is a better t-test, this testing still cannot obtain standard errors between people and between groups.

(C) Let use a flat prior on β, we have the joint distribution

$$p(\boldsymbol{\theta}, \beta, \sigma^2, \tau^2|\boldsymbol{y}) \propto 1 \cdot \frac{1}{\sigma^2} \cdot \frac{1}{\tau^2}^{\frac{1}{2}+1} \exp\left(-\frac{1}{2\tau^2}\right) \prod_{i=1}^{P} (\tau^2\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i - \mu - \beta x_i)^2\right)$$
$$\prod_{j=1}^{N_i} (\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right)$$

Therefore, we have the conditional posteriors:

$$\begin{aligned}
p(\theta_i|-) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i - \mu - \beta x_i)^2\right) \prod_{j=1}^{N_i} \exp\left(-\frac{1}{2\sigma^2}(y_{ij} - \theta_i)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}(\theta_i^2 - 2\theta_i(\mu + \beta x_i))\right) \exp\left(-\frac{1}{2\sigma^2}(N_i\theta_i^2 - 2\theta_i \sum_{j=1}^{N_i} y_{ij})\right) \\
&\propto \exp\left(-\frac{1}{2}\left(\theta_i^2\left(\frac{1}{\sigma^2\tau^2} + \frac{N_i}{\sigma^2}\right) - 2\theta_i\left(\frac{\mu + \beta x_i}{\sigma^2\tau^2} + \frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2}\right)\right)\right) \\
&= \mathcal{N}\left(\left(\frac{1}{\sigma^2\tau^2} + \frac{N_i}{\sigma^2}\right)^{-1}\left(\frac{\mu + \beta x_i}{\sigma^2\tau^2} + \frac{\sum_{j=1}^{N_i} y_{ij}}{\sigma^2}\right), \left(\frac{1}{\sigma^2\tau^2} + \frac{N_i}{\sigma^2}\right)^{-1}\right)
\end{aligned}$$

and

$$p(\sigma^2|-) = (\sigma^2)^{-1-\frac{P}{2}-\frac{\sum_{i=1}^{P} N_i}{2}} \exp\left(-\frac{1}{\sigma^2}\left(\frac{\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2}{2\tau^2} + \sum_{i=1}^{P}\frac{\sum_{j=1}^{N_i}(y_{ij} - \theta_i)^2}{2}\right)\right)$$
$$Inv - Gamma\left(\frac{P + \sum_{i=1}^{P} N_i}{2}, \left(\frac{\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2}{2\tau^2} + \sum_{i=1}^{P}\frac{\sum_{j=1}^{N_i}(y_{ij} - \theta_i)^2}{2}\right)\right)$$

and

$$\begin{aligned}
p(\tau^2|-) &= (\tau^2)^{-1-\frac{1}{2}-\frac{P}{2}} \exp\left(-\frac{1}{\tau^2}\left(\frac{\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2}{2\sigma^2} + \frac{1}{2}\right)\right) \\
&= Inv - gamma\left(\frac{P+1}{2}, \frac{\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2}{2\sigma^2} + \frac{1}{2}\right)
\end{aligned}$$

and

$$\begin{aligned}
p(\mu|-) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\theta_i - \mu - \beta x_i)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\mu^2 - 2\mu(\theta_i - \beta x_i) + (\theta_i - \beta x_i)^2)\right) \\
&\propto \exp\left(-\frac{P}{2\sigma^2\tau^2}(\mu^2 - 2\mu\sum_{i=1}^{P}(\theta_i - \beta x_i)\frac{1}{P}))\right) \\
&= \mathcal{N}\left(\frac{\sum_{i=1}^{P}\theta_i - \beta x_i}{P}, \frac{\sigma^2\tau^2}{P}\right)
\end{aligned}$$

and

$$
\begin{aligned}
p(\beta|-) &\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\theta_i-\mu-\beta x_i)^2\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}\sum_{i=1}^{P}(\beta^2x_i^2-2(\theta_i-\mu)\beta x_i+(\theta_i-\mu)^)\right) \\
&\propto \exp\left(-\frac{1}{2\sigma^2\tau^2}(\beta^2\sum_{i=1}^{P}x_i^2-2\beta\sum_{i=1}^{P}(\theta_i-\mu)x_i))\right) \\
&= \mathcal{N}\left(\frac{\sum_{i=1}^{P}(\theta_i-\mu)x_i}{\sum_{i=1}^{P}x_i^2}, \frac{\sigma^2\tau^2}{\sum_{i=1}^{P}x_i^2}\right)
\end{aligned}
$$

I run 1000 Gibbs iterations (600 thinned samples). I plot the histogram of β in Figure 3. Compared to (A) and (B), the posterior mean is -8.24 which is quite similar. The standard deviation is 4.73 which is also nearly identical to part (A) and (B).

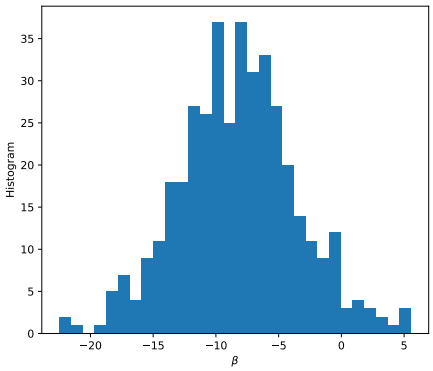

Figure 3: Histogram of β

(D) Based on the data, the assumption to be sensible since blood pressure measurements appear to not be autocorrelated within individuals. In particular, the autocorrelation plots of all subjects are given in Figure 4. From the figures, we observe that the measurements are almost indepedent.

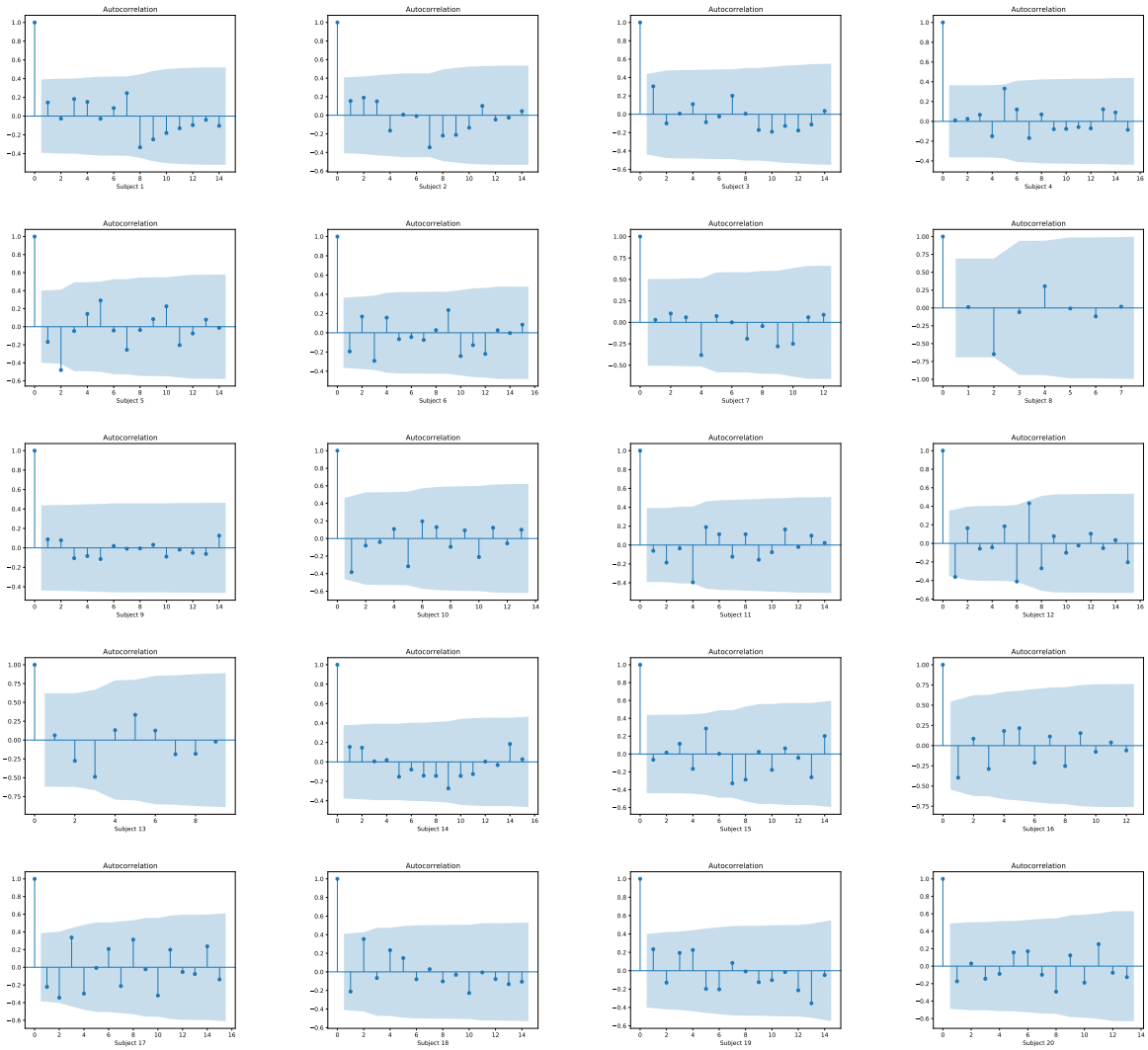

Figure 4: Autocorrelation plots.