Statistical Analysis with Missing Data
Problem 7.9 , 7.16

16212799, 何浩勋

October 7, 2017

1 Problem 7.9

Generate a bivariate normal sample of 20 cases with parameter $\mu_1 = \mu_2 = 0$, $\sigma_{11} = \sigma_{12} = 1, \sigma_{22} = 2$, denoted Y_1, Y_2, and delete values of Y_2 so that $Pr(y_2\ missing|y_1, y_2)$ equals 0.2 if $y_1 < 0$ and 0.8 if $y_1 > 0$.

Label Description
obs : cases that both Y_1 and Y_2 are observed
mis : cases that Y_1 is observed but Y_2 is missing
S.D. : significant difference

Dataset	datasets with S.D.	datasets without S.D.
7.9-dataset	615	385
MCAR-dataset	57	943

Table 1: number of datasets with or without significant difference

(a) Construct a test for whether the data are MCAR and carry out the test on the dataset.
If the missing-data mechanism is MCAR, there will be no significant difference between $Y_1(obs)$ and $Y_1(mis)$. However, the sample size is small, Repeated trial is needed. I generate 1000 datasets(denoted 7.9-datasets) as the problem describes and carry out t-tests between $Y_1(obs)$ and $Y_1(mis)$ on every one of them. If the p-value of the test is lower than 0.05, it is considered that there is significant difference between $Y_1(obs)$ and $Y_1(mis)$ in the dataset. To tell whether the missing

data mechanism is MCAR, I generate another 1000 datasets(denoted MCAR-datasets) that delete 50% values of Y_2 completely at random and carry out t-tests as the same and compare the result of 7.9-datasets and MCAR-datasets.

As Table 1 shows, more than 60% of the 7.9-datasets have significant difference between $Y_1(obs)$ and $Y_1(mis)$. However, most of the MCAR-datasets don't have significant difference between $Y_1(obs)$ and $Y_1(mis)$. Therefore, The missing-data mechanism is no MCAR.

(b) Compute 95% confidence intervals for μ_2 using (i) the data before values were deleted; (ii) the complete cases; (iii) the t-approximation in (2) of Table7.2; Summarize the propeties of these intervals for this missing-data mechanism.

(i) Using the data before values were deleted: μ_2 is estimated by the sample mean and $var(\hat{\mu}_2 - \mu_2)$ is estimated by (7.13). The confidence interval is given by (7.15) That is:

$$\hat{\mu}_2 = \bar{y}_2$$

$$\hat{var}(\hat{\mu}_2 - \mu_2) = \hat{\sigma}_{221}[\frac{1}{r} + \frac{\hat{\rho}^2}{n(1-\hat{\rho}^2)} + \frac{(\bar{y}_1 - \hat{\mu}_1)^2}{rs_{11}}] \quad (1)$$

$$95\%CI : \mu_2 \pm 1.96\sqrt{var(\hat{\mu}_2 - \mu_2)}$$

	lower bound(avg)	upper bound(avg)	length(avg)
(i)	-0.17845176	0.19496054	0.37341230
(ii)	-0.81337044	-0.14332189	0.67004855
(iii)	-0.49194380	0.50547961	0.99742341

Table 2: number of datasets with or without significant difference

(ii) Using the complete cases: μ_2 is estimated by the observed-sample mean and $var(\hat{\mu}_2 - \mu_2)$ is estimated by (7.13). except substituting y in equation 1 with $y(obs)$, the procedure to obtain confidence interval is the same as (i).

(iii) Using t-approximation in (2) of Table 7.2 : μ_2 is estimated by $\bar{y}_2 + \beta_{211}(\hat{\mu}_1 - \bar{y}_1)$ and $var(\hat{\mu}_2 - \mu_2)$ is estimated by (7.13). intervals

are obtain using the complete-case degrees of freedom, the normal percentile 1.96 in equation 1 is replaced by the 97.5th percentile of the t distribution.
the average confidence intervals obtain by (i)(ii)(iii) is list on Tab 2. We can see that the CIs obtain in (i)(iii) are unbiased while those obtain in (ii) are badly biased. Obviously, the bigger values have bigger chance to be deleted. CIs obtain in (ii) is shorter in length than those obtain in (iii),however the real parameter rarely falls in the CIs obtain in (ii).

2 Problem 7.16

(i)Consider the form of the discriminant analysis model for bivariate data with binary X and continuous Y:
(a)X is Bernoulli with $Pr(X=1) = 1 - Pr(X=0) = p$ **and**
(b)Y given X=j is normal with mean μ_j**, variance** σ^2

Derive ML estimates.

$$f(Y|X) = \left\{ \frac{\boldsymbol{I}^{(x)}_{\{x=0\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y-\mu_0)^2}{2\sigma^2}\right] + \frac{\boldsymbol{I}^{(x)}_{\{x=1\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y-\mu_1)^2}{2\sigma^2}\right] \right\}$$

$$f(X) = (1-p)\boldsymbol{I}^{(x)}_{\{x=0\}} + p\boldsymbol{I}^{(x)}_{\{x=1\}}$$

So the joint p.d.f $f(X,Y) = f(Y|X)f(X)$**,that is**

$$f(X,Y) = \left\{ \frac{(1-p)\boldsymbol{I}^{(x)}_{\{x=0\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y-\mu_0)^2}{2\sigma^2}\right] + \frac{p\boldsymbol{I}^{(x)}_{\{x=1\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y-\mu_1)^2}{2\sigma^2}\right] \right\}$$

The likelihood function is

$$\begin{aligned} L &= \prod_{i=1}^{n} \left\{ \frac{(1-p)\boldsymbol{I}^{(x_i)}_{\{x=0\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y_i-\mu_0)^2}{2\sigma^2}\right] + \frac{p\boldsymbol{I}^{(x_i)}_{\{x=1\}}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y_i-\mu_1)^2}{2\sigma^2}\right] \right\} \\ &= (1-p)^{n_0} p^{n_1} (\sqrt{2\pi}\sigma)^{-n} exp\left\{ \sum_{i\in\{i|x_i=0\}} \frac{(y_i-\mu_0)^2}{2\sigma^2} + \sum_{i\in\{i|x_i=1\}} \frac{(y_i-\mu_1)^2}{2\sigma^2} \right\} \end{aligned}$$

where $n_0 = \sum_{i=0}^{n} \boldsymbol{I}^{(x_i)}_{\{x=0\}}$ **,** $n_1 = \sum_{i=0}^{n} \boldsymbol{I}^{(x_i)}_{\{x=1\}}$ **, and** $n = n_0 + n_1$**.**

The log-likelihood function is

$$l = n_0 ln(1-p) + n_1 ln(p) - \frac{n}{2} ln(2\pi) - nln(\sigma) \\ - \sum_{i \in \{i|x_i=0\}} \frac{(y_i - \mu_0)^2}{2\sigma^2} - \sum_{i \in \{i|x_i=1\}} \frac{(y_i - \mu_1)^2}{2\sigma^2} \tag{2}$$

take the derivative of (2) function with respect to $(p, \mu_0, \mu_1, \sigma)$

$$\frac{dl}{dp} = -\frac{n_0}{1-p} + \frac{n_1}{p}$$
$$\frac{dl}{d\mu_0} = -\sum_{i \in \{i|x_i=0\}} \frac{y_i - \mu_0}{\sigma^2} = -\frac{\sum_{i \in \{i|x_i=0\}} y_i - n_0\mu_0}{\sigma^2}$$
$$\frac{dl}{d\mu_1} = -\sum_{i \in \{i|x_i=1\}} \frac{y_i - \mu_1}{\sigma^2} = -\frac{\sum_{i \in \{i|x_i=1\}} y_i - n_1\mu_1}{\sigma^2}$$
$$\frac{dl}{d\sigma} = -\frac{n}{\sigma} + \sum_{i \in \{i|x_i=0\}} \frac{(y_i - \mu_0)^2}{\sigma^3} + \sum_{i \in \{i|x_i=1\}} \frac{(y_i - \mu_1)^2}{\sigma^3}$$

set the derivatives to 0 , we have

$$\hat{p} = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n}$$
$$\hat{\mu}_0 = \frac{\sum_{i \in \{i|x_i=0\}} y_i}{n_0}$$
$$\hat{\mu}_1 = \frac{\sum_{i \in \{i|x_i=1\}} y_i}{n_1}$$
$$\hat{\sigma} = \sqrt{\frac{\sum_{i \in \{i|x_i=0\}} (y_i - \hat{\mu}_0)^2 + \sum_{i \in \{i|x_i=1\}} (y_i - \hat{\mu}_1)^2}{n}}$$

these are the MLE of $(p, \mu_0, \mu_1, \sigma^2)$**.**

Derive marginal mean and variance of Y.

$$\begin{aligned} E(Y) &= \int \int y f(x,y) dx dy \\ &= \int y \int f(x,y) dx dy \\ &= \int y \left[(1-p) f(y|x=0) + p f(y|x=1) \right] dy \\ &= (1-p) E[Y|X=0] + p E[Y|X=1] \\ &= (1-p)\mu_0 + p\mu_1 \end{aligned} \tag{3}$$

$$
\begin{aligned}
Var(Y) &= E(Y^2) - [E(Y)]^2 \\
&= \int\int y^2 f(x,y)dxdy - [E(Y)]^2 \\
&= \int y^2 \left[(1-p)f(y|x=0) + pf(y|x=1)\right] dy - [E(Y)]^2 \\
&= (1-p)E(Y^2|X=0) + pE(Y^2|X=1) - [E(Y)]^2 \\
&= (1-p)(\mu_1^2+\sigma^2) + p(\mu_0^2+\sigma^2) - [(1-p)\mu_0 + p\mu_1]^2 \\
&= \sigma^2 + p(1-p)\mu_0^2 + p(1-p)\mu_1^2 - 2p(1-p)\mu_0\mu_1 \\
&= \sigma^2 + p(1-p)(\mu_0-\mu_1)^2 \qquad (4)
\end{aligned}
$$

Substitute $(p, \mu_0, \mu_1, \sigma)$ with $(\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma})$ in (3)(4) , The marginal mean of Y is $(1-\hat{p})\hat{\mu}_0+\hat{p}\hat{\mu}_1$, and marginal variance of Y is $\hat{\sigma}^2+\hat{p}(1-\hat{p})(\hat{\mu}_0-\hat{\mu}_1)^2$. That is

(ii)Suppose now that X is completely observed, but n-r values of Y are missing, with ignorable mechanism. Use the methods of Chapter 7 to derive the ML estimates of the marginal mean and variance of Y for this monotone pattern.

Arrange the data so that $\{y_1, ..., y_r\}$ is observed and $\{y_{r+1}, ..., y_n\}$, Let $\theta = (p, \mu_1, \mu_2, \sigma)$, and the density of the data (X, Y_{obs}) factors in the following way:

$$
\begin{aligned}
f(X, Y_{obs}|\theta) = &\prod_{i=1}^{r} \left\{ \frac{(1-p)\boldsymbol{I}_{\{x=0\}}^{(x_i)}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y_i-\mu_0)^2}{2\sigma^2}\right] + \frac{p\boldsymbol{I}_{\{x=1\}}^{(x_i)}}{\sqrt{2\pi}\sigma} exp\left[\frac{-(y_i-\mu_1)^2}{2\sigma^2}\right] \right\} \\
&\cdot \prod_{i=r+1}^{n} \left\{ (1-p)\boldsymbol{I}_{\{x=0\}}^{(x_i)} + p\boldsymbol{I}_{\{x=1\}}^{(x_i)} \right\}
\end{aligned}
$$

similarly , the log-likelihood function is

$$
\begin{aligned}
l = &\, n_0 ln(1-p) + n_1 ln(p) - \frac{r}{2} ln(2\pi) - r ln(\sigma) \\
&- \sum_{i\in\{i|x_i=0, i\le r\}} \frac{(y_i-\mu_0)^2}{2\sigma^2} - \sum_{i\in\{i|x_i=1, i\le r\}} \frac{(y_i-\mu_1)^2}{2\sigma^2}
\end{aligned}
$$

where $n_0 = \sum_{i=0}^{n} \boldsymbol{I}_{\{x=0\}}^{(x_i)}$, $n_1 = \sum_{i=0}^{n} \boldsymbol{I}_{\{x=1\}}^{(x_i)}$, and $n = n_0 + n_1$.

set the derivatives to 0 , we have

$$\hat{p} = \frac{n_1}{n_0 + n_1} = \frac{n_1}{n}$$

$$\hat{\mu}_0 = \frac{\sum\limits_{i \in \{i | x_i = 0, i \leq r\}} y_i}{r_0}$$

$$\hat{\mu}_1 = \frac{\sum\limits_{i \in \{i | x_i = 1, i \leq r\}} y_i}{r_1}$$

$$\hat{\sigma} = \sqrt{\frac{\sum\limits_{i \in \{i | x_i = 0, i \leq r\}} (y_i - \hat{\mu}_0)^2 + \sum\limits_{i \in \{i | x_i = 1, i \leq r\}} (y_i - \hat{\mu}_1)^2}{r}}$$

where $r_0 = \sum\limits_{i=0}^{r} \boldsymbol{I}_{\{x=0\}}^{(x_i)}$ **,** $r_1 = \sum\limits_{i=0}^{r} \boldsymbol{I}_{\{x=1\}}^{(x_i)}$ **, and** $r = r_0 + r_1$**.**

Substitute $(p, \mu_0, \mu_1, \sigma)$ **with** $(\hat{p}, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma})$ **in (3)(4) , The marginal mean of Y is** $(1-\hat{p})\hat{\mu}_0 + \hat{p}\hat{\mu}_1$ **, and marginal variance of Y is** $\hat{\sigma}^2 + \hat{p}(1-\hat{p})(\hat{\mu}_0 - \hat{\mu}_1)^2$**.**