University of Edinburgh
School of Mathematics
Incomplete Data Analysis

Assignment–Solutions (sketch)

1. (a) The mean value of the recovery time, provided by a complete case analysis, is 19.273 and the associated standard error is 2.603. The correlation between the recovery time and (log) dose is 0.239, while the correlation between recovery time and blood pressure is −0.020.

```
rm(list=ls())
set.seed(1)
load("databp.Rdata")

ind=which(databp$R==1)
recovtimeccm=mean(databp$recovtime,na.rm=TRUE)
recovtimeccse=sd(databp$recovtime,na.rm=TRUE)/sqrt(length(ind))
recovtimeccm; recovtimeccse

## [1] 19.27273
## [1] 2.603013

cor(databp$recovtime,databp$logdose,use="complete.obs")

## [1] 0.2391256

cor(databp$recovtime,databp$bloodp,use="complete.obs")

## [1] -0.01952862
```

   (b) The mean recovery time using mean imputation is 19.273 and the associated standard error is 2.284. The correlation between the recovery time and (log) dose is 0.215, while the correlation between recovery time and blood pressure is −0.019.

```
rectimemi=ifelse(databp$R==0,mean(databp$recovtime,na.rm=TRUE),databp$recovtime)

n=nrow(databp)
mmi=mean(rectimemi)
semi=sd(rectimemi)/sqrt(n)
mmi; semi

## [1] 19.27273
## [1] 2.284135

cor(rectimemi,databp$logdose); cor(rectimemi,databp$bloodp)

## [1] 0.2150612
## [1] -0.01934126
```

   (c) We will fit a linear regression model to the complete cases, using recovery time as the response and (log) dose and blood pressure as the predictors. The regression equation is

$$\text{RecovTime} = \beta_0 + \beta_1 \text{logDose} + \beta_2 \text{BloodP} + \varepsilon, \qquad \varepsilon \sim \text{N}(0, \sigma^2).$$

The estimated regression coefficients are $\widehat{\beta_0} = 15.216$, $\widehat{\beta_1} = 11.429$, and $\widehat{\beta_2} = -0.277$. The predicted values are

Subject id	log dose	Blood pressure	Imputed recovery time
4	1.54	67	14.263
10	2.32	73	21.516
22	2.62	68	26.329

The estimated mean recovery time is then 19.444 with associated standard error of 2.313. Further, the correlation between the recovery time and (log) dose is 0.280, while the correlation between recovery time and blood pressure is -0.011.

```
fitrectime=lm(recovtime~logdose+bloodp,data=databp)
summary(fitrectime)

##
## Call:
## lm(formula = recovtime ~ logdose + bloodp, data = databp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.768 -10.250  -0.770   3.546  37.394
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  15.2159    19.8203   0.768    0.452
## logdose      11.4290     8.4178   1.358    0.190
## bloodp       -0.2769     0.3411  -0.812    0.427
##
## Residual standard error: 12.25 on 19 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.08879,Adjusted R-squared:  -0.007129
## F-statistic: 0.9257 on 2 and 19 DF,  p-value: 0.4134

predri=predict(fitrectime,newdata=databp)
predri[4]; predri[10]; predri[22]

##        4
## 14.26254
##       10
## 21.51562
##       22
## 26.32896

rectimeri=ifelse(databp$R==1,databp$recovtime,predri)

mean(rectimeri)

## [1] 19.44428

sd(rectimeri)/sqrt(n)

## [1] 2.312845

cor(rectimeri,databp$logdose); cor(rectimeri,databp$bloodp)
```

```
## [1] 0.2801835
## [1] -0.0111364
```

(d) We do a similar analysis to the one in (c) but we now add a random noise to the predictions, i.e.,

$$\widehat{\text{RecovTime}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{logDose} + \widehat{\beta}_2 \text{BloodP} + z, \qquad z \sim \text{N}(0, \widehat{\sigma}^2), \qquad \widehat{\sigma} = 12.25.$$

Note that in this case, depending on the seed, we can obtain negative predicted values for the recovery time, which of course, does not make any sense in practice. Using set.seed(1) we have obtained a mean recovery time of 20.460 with an associated standard error of 2.445. Additionally, the correlation between the recovery time and (log) dose is 0.228, while the correlation between recovery time and blood pressure is -0.018.

```
predsri=predict(fitrectime,newdata=databp)+rnorm(n,0,summary(fitrectime)$sigma)
sbpsri=ifelse(databp$R==1,databp$recovtime,predsri)
sbpsri[4]; sbpsri[10]; sbpsri[22]

## [1] 33.80897
## [1] 17.7738
## [1] 35.91221

mean(sbpsri); sd(sbpsri)/sqrt(n)

## [1] 20.4598
## [1] 2.444571

cor(sbpsri,databp$logdose); cor(sbpsri,databp$bloodp)

## [1] 0.2284537
## [1] -0.01786944
```

(e) Implementing the described predictive mean matching technique, we find that the donor for subject 4 is subject 6, the donor for subject 10 is subject 2, and the donor for subject 22 is subject 17. The resulting mean recovery time, after imputation, is 19.44 (s.e. is 2.464). The correlation between the recovery time and (log) dose is 0.304 and the correlation between recovery time and blood pressure is -0.032.

```
require(mice) #to use the function ic

#predicted values for subjects with missing recovery time
predmis=predict(fitrectime,newdata=ic(databp))

#set the predicted values for the subjects with missing recovery time to
#a very large value  so that they will never be the ones with minimum distance
predri[4]=10000
predri[10]=10000
predri[22]=10000

#finding the donor to subject 4
dis4=numeric(n)
for(i in 1:n){
dis4[i]=(predmis[1]-predri[i])^2
}
```

```
ind4=which(dis4==min(dis4))
ind4

## [1] 6

#finding the donor to subject 10
dis10=numeric(n)
for(i in 1:n){
dis10[i]=(predmis[2]-predri[i])^2
}
ind10=which(dis10==min(dis10))
ind10

## [1] 2

#finding the donor to subject 22
dis22=numeric(n)
for(i in 1:n){
dis22[i]=(predmis[3]-predri[i])^2
}
ind22=which(dis22==min(dis22))
ind22

## [1] 17

recovtimepmm=c(7,10,18,databp$recovtime[6],10,13,21,12,9,databp$recovtime[2],
               20,31,23,22,13,9,39,28,12,60,10,databp$recovtime[17],22,21,14)
mean(recovtimepmm); sd(recovtimepmm)/sqrt(n)

## [1] 19.44
## [1] 2.464467

cor(recovtimepmm,databp$logdose); cor(recovtimepmm,databp$bloodp)

## [1] 0.3037945
## [1] -0.03208685
```

(f) An obvious advantage of predictive mean matching over stochastic regression imputation is that it always produce plausible imputed values, since every imputed value is 'donated' from a subject in the study.