Missing Data Analysis

Kei Sakamoto

Missing data の扱いについて。

```
load("~/計量経済学演習/R data sets for 5e/lawsch85.RData")
lawsch85<-data
```

extract LSAT

```
lsat <- lawsch85$LSAT
head(lsat)
```

```
## [1] 155 160 155 157 162 161
```

Create logical indicator for missing data

```
missLSAT <- is.na(lawsch85$LSAT)
head(missLSAT)
```

```
## [1] FALSE FALSE FALSE FALSE FALSE FALSE
```

LSAT and indicator for Schools No. 120-129(for 2 NA in this range)

```
rbind(lsat,missLSAT)[,120:129]
```

```
##          [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## lsat      156  159  157  167   NA  158  155  157   NA   163
## missLSAT    0    0    0    0    1    0    0    0    1     0
```

missLSAT の型は全体で Numeric なので Logical でなく binary。table にすれば Logical で返ってくる。

Frequencies of indicator

```
table(missLSAT)
```

```
## missLSAT
## FALSE  TRUE
##   150     6
```

Missings for all variables in data frame (counts)

```
colSums(is.na(lawsch85))
```

```
##    rank  salary    cost    LSAT     GPA  libvol faculty     age  clsiz
e
##       0       8       6       6       7       1       4      45
3
```

```
##   north   south    east    west lsalary studfac   top10  r11_25  r26_4
0
##       0       0       0       0       8       6       0       0
0
##  r41_60 llibvol   lcost
##       0       1       6
```

Indicator for complete cases

```
compl <- complete.cases(lawsch85)
table(compl)
```

```
## compl
## FALSE  TRUE
##    66    90
```

全部完璧に報告してるのが 156 校の内なんと 60 校のみ。落ちてる理由が完全に random ならこの後の regression の coef は consistent に保たれるが、なんか理由があって、しかもその理由が dependent variable に説明力を持つものならそのまま regresson したらまずい。でも今回は落ちてる理由はなく完全に random だとする。

```
mean(lsat)
```

```
## [1] NA
```

当然 NA が入ってるので計算できない。

```
mean(lsat,na.rm=TRUE)
```

```
## [1] 158.2933
```

NA のところをとり除けば計算できる。

Regression with missings

```
summary(lm(log(salary)~LSAT+cost+age, data=lawsch85))
```

```
##
## Call:
## lm(formula = log(salary) ~ LSAT + cost + age, data = lawsch85)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.40989 -0.09438  0.00317  0.10436  0.45483
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4.384e+00  6.781e-01   6.465 4.94e-09 ***
## LSAT        3.722e-02  4.501e-03   8.269 1.06e-12 ***
## cost        1.114e-05  4.321e-06   2.577 0.011563 *
## age         1.503e-03  4.354e-04   3.453 0.000843 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1545 on 91 degrees of freedom
##   (61 observations deleted due to missingness)
## Multiple R-squared:  0.6708, Adjusted R-squared:  0.6599
## F-statistic: 61.81 on 3 and 91 DF,  p-value: < 2.2e-16
```

実は default で NA の data は無視して回帰するようになっている。その旨は summary で報告されている。