Learning *R*

Carl James Schwarz

StatMathComp Consulting by Schwarz
cschwarz.stat.sfu.ca @ gmail.com

Missing values

Table of Contents I

Missing Values - NA's, NaN, Infs

R Missing Values

Missing values (NA) can occur

- Data lacking values (are these MCAR, MAR, or IM)?
- Illegal computations ("Carl"/3) or 10/0
- Over/Under flow (e.g. exp(1000))

Default action of *R* is to propagate missing values, i.e. 3 + NA is also NA.
You can change this action in several ways (the na.actions).

R Missing Values

Look at *cereal* dataframe and the *weight* variable.
Some cereals have missing values.
Compare the following results.
Read *help(is.na)* and *help(na.omit)*

```
1  mean(cereal$weight)
2  mean(cereal$weight, na.rm=TRUE)
3  mean(na.omit(cereal$weight))
4
5  length(cereal$weight)  # includes missing values
6  length(na.omit(cereal$weight))
7  is.na(cereal$weight)
8  sum( is.na(cereal$weight)) # count num missing
9
10 dim(cereal)
11 dim(na.omit(cereal)) # drop row with missing data
12
13 complete.cases(cereal)
14 dim(cereal[complete.cases(cereal),])
```

R Missing Values

Look at cereal dataframe and the *weight* variable.
Compute the % weight of fat, i.e. grams of fat/serving size.
Compute the protein:fat ratio.

```
1  # Using NA in operations leads to NA's
2  cereal$prop.fat <- cereal$fat /cereal$weight
3  cereal$prop.fat
4  is.na(cereal$protein.fat) # Inf is a value
5
6
7  # NA is different from Inf;   protein/fat ratio
8  cereal$protein.fat <- cereal$protein / cereal$fat
9  cereal$protein.fat
10 iis.na(cereal$protein.fat) # Inf is a value
11 is.infinite(cereal$protein.fat)
```

R Missing Values - Caution with plotting functions

Missing values typically don't show up on plots!

```
1 ggplot(data=cereal, aes(x=weight, y=calories))+
2   ggtitle("calories vs. Weight per serving")+
3   xlab("Weight per serving")+ylab("calories")+
4   geom_jitter()+
5   geom_smooth(method="lm",se=FALSE)
```

```
Warning messages:
1: Removed 2 rows containing missing values (stat_smooth).
2: Removed 2 rows containing missing values (geom_point).
```

R Missing Values - Caution with modeling functions

Look at cereal dataframe and the weight variable.
Regress calories against serving size.

```
1 fit.cal.serving <- lm(calories ~ weight,
2                       data=cereal)
3 summary(fit.cal.serving) # note missingness
```

```
Residual standard error: 12.89 on 73 degrees of freedom
  (2 observations deleted due to missingness)
```

R Missing Values - Caution with modeling functions

Look at cereal dataframe and the weight variable.
Regress calories against serving size.

```
1 fitted(fit.cal.serving) # only length 75 despite having ind
2 length(fitted(fit.cal.serving))
```

So the following fails:

```
1 cereal$fitted <- fitted(fit.cal.serving)
```

```
Error in `$<-.data.frame`(`*tmp*`, "fitted", value = c(100.6
  replacement has 75 rows, data has 77
```

R Missing Values - Caution with modeling functions

Look at cereal dataframe and the weight variable.
Regress calories against serving size.

It is possible to have *lm()* deal *nicely* with NA, but this approach is not implemented consistently in *R*.

```
1 fit.cal.serving2 <- lm(calories ~ weight,
2           na.action=na.exclude,
3           data=cereal)
4 summary(fit.cal.serving2)
5 fitted(fit.cal.serving2) # now padded with NA
6 length(fitted(fit.cal.serving2))
7 cereal$fitted <- fitted(fit.cal.serving2)
```

R Missing Values - Caution with modeling functions

Look at cereal dataframe and the weight variable.
Regress calories against serving size.

I prefer to use the predict with new data that propagates missing values properly.

Compare:

```
predict(fit.cal.serving)
predict(fit.cal.serving, newdata=cereal)

predict(fit.cal.serving2)
```

R Missing Values - Summary

- ALWAYS check for NA's in data and ask if MCAR, MAR, or IM
- Sometime special codes are used, e.g. -1 means missing

```
1 mydf [ is.na(mydf$var), ] # list rows where var is miss
2 mydf [ mydf == -1 ] <- NA # common code
```

- NA and Inf's propagate through computation but *R* is NOT consistent.
  - *mean(var)* is NA if var contains any missing
  - *mean(var, na.rm=TRUE)* drops the NAs
  - *lm(Y∼ X, data=blah)* drops missing rows automatically
- Different results from methods with *na.action=na.exclude* vs. *na.action = na.omit*.