Code ▾

ANLY 699: Assignment 2

Subhash Pemmaraju

06/07/2020

Hide

```
missing_sum%>%kable("html", digits=1,
 col.names = c("% w Parks access",
               "Yrs of Pot. life lost Rate",
               "% w fair/poor health",
               "Avg days phys. unhappy",
               "Avg days ment. unhappy",
               "% obese",
               "% Phys inactive",
               "% smokers"))%>%kable_styling(bootstrap_options="striped",full_width=FALSE)
```

	% w Parks access	Yrs of Pot. life lost Rate	% w fair/poor health	Avg days phys. unhappy	Avg days ment. unhappy	% obese	% Phys inactive	% smokers
Count	3142.0	3142.0	3142.0	3142.0	3142.0	3142.0	3142.0	3142.0
Mean	25.9	8572.2	17.9	4.0	4.2	32.9	27.4	17.5
SD	23.8	2570.9	4.7	0.7	0.6	5.5	5.7	3.6
Missing_Num	0.0	295.0	2.0	2.0	2.0	2.0	2.0	2.0
Missing_Perc	0.0	9.4	0.1	0.1	0.1	0.1	0.1	0.1

Hide

```
complete_data <- merged_data[complete.cases(merged_data[,c(6,9:14,16)]),]
```

Summary of missing variables

There is very little data missing in the dataset. Among all of the variables, Years of potential life lost is the only variable which has 295 missing records, which is 9.4% of the total sample. All the other variables have at most 2 missing records which is under 0.5% of the total sample.

No. of missing overall = 295

% missing overall = 9.3889243

Hide

```
missing_data <- subset(health_ind_cty, is.na(yrs_plr) == T )
#missing_data
```

Summary of missing data in Years of potential life lost

There is no apparent pattern to the missing data in years of potential life lost. It is missing in counties across multiple states.Furthermore, MCAR test below shows that normality can be rejected and consequently, that MCAR cannot be rejected. Therefore, the pattern of missing data is MCAR.

Hide

```
library(BaylorEdPsych)
library(MissMech)

TestMCARNormality(merged_data[,c(6,9:14,16)])
```

```
## Call:
## TestMCARNormality(data = merged_data[, c(6, 9:14, 16)])
## 
## Number of Patterns:  2 
## 
## Total number of cases used in the analysis:  3140 
## 
##  Pattern(s) used:
##          Value   yrs_plr   perc_fair_poor_health   avg_phy_unh_days
## group.1       1         1                       1                  1
## group.2       1        NA                       1                  1
##          avg_mental_unh_days   perc_obese   perc_phy_inact   perc_smokers
## group.1                    1            1                1              1
## group.2                    1            1                1              1
##          Number of cases
## group.1             2847
## group.2              293
## 
## 
##     Test of normality and Homoscedasticity:
##   -------------------------------------------
## 
## Hawkins Test:
## 
##     P-value for the Hawkins test of normality and homoscedasticity:  2.162954e-61 
## 
##     Either the test of multivariate normality or homoscedasticity (or both) is rejected.
##     Provided that normality can be assumed, the hypothesis of MCAR is 
##     rejected at 0.05 significance level. 
## 
## Non-Parametric Test:
## 
##     P-value for the non-parametric test of homoscedasticity:  0.08326334 
## 
##     Reject Normality at 0.05 significance level.
##     There is not sufficient evidence to reject MCAR at 0.05 significance level.
```

Summary

Because the sample size is larege enough (3142) that excluding 295 missing cases will still leave a large enough sample, we can consider using list-wise deletions and run the analysis only on complete cases. Alternatively, for the missing values, we can also use a mean substitution technique to replace all the missing values for years of potential life lost with the mean for the available data. This technique is superior in that there is potentially some value to the data in the other variables for the purpose of this analysis and it might be worth it to retain them.