IST438-W3-Applications

3/13/23

Supervised Learning: Logistic Regression Models

In this application, we will interest the regression problem under the following sections:

- Training model
- Measuring model performance
- Checking over and underfitting problem

Packages

We need to install `{naniar}` and `{DALEX}` package to use functions to handle missing data and `titanic` data set in applications. Please use the two-step codes below: (1) install, (2) load the package.

```
#install.packages("naniar")
#install.packages("DALEX")
library(naniar)
library(DALEX)
```

Missing data summaries

We can summarizes the missing values in vector or data frame format. To summarize the missing values in a data set, `{naniar}` provides very useful functions as below:

- `n_miss()` returns number of missing values in data set
- `n_complete()` returns number of completed (aka not missing) values in data set

- `miss_var_summary()` returns number and percentage of missing values in data set for each variable
- `miss_case_summary()` returns number and percentage of missing values in data set for each observation

```
n_miss(titanic)
```

```
[1] 129
```

```
n_complete(titanic)
```

```
[1] 19734
```

```
miss_var_summary(titanic)
```

```
# A tibble: 9 x 3
  variable n_miss pct_miss
  <chr>     <int>    <dbl>
1 country      81   3.67
2 fare         26   1.18
3 sibsp        10   0.453
4 parch        10   0.453
5 age           2   0.0906
6 gender        0   0
7 class         0   0
8 embarked      0   0
9 survived      0   0
```

```
miss_case_summary(titanic)
```

```
# A tibble: 2,207 x 3
    case n_miss pct_miss
   <int>  <int>    <dbl>
 1   577      4     44.4
 2   145      3     33.3
 3   151      3     33.3
 4   238      3     33.3
 5   517      3     33.3
 6   616      3     33.3
 7   681      3     33.3
 8  1095      3     33.3
 9  1190      3     33.3
10  1305      3     33.3
# ... with 2,197 more rows
```

- `miss_var_table()` returns a summary table consists number and percentage of missing values over variables.

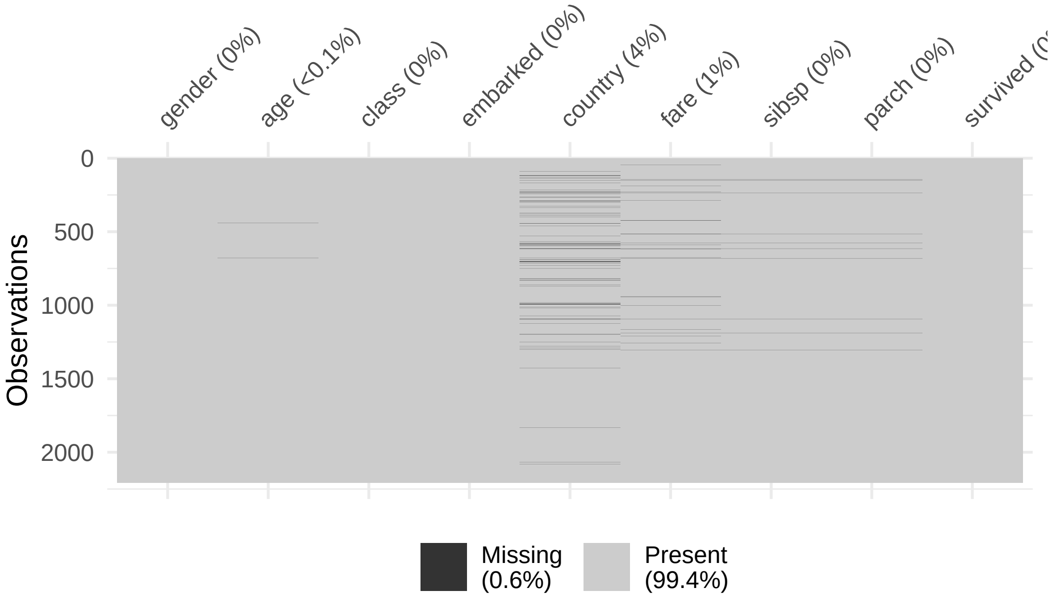
```
miss_var_table(titanic)
```

```
# A tibble: 5 x 3
  n_miss_in_var n_vars pct_vars
          <int>  <int>    <dbl>
1             0      4     44.4
2             2      1     11.1
3            10      2     22.2
4            26      1     11.1
5            81      1     11.1
```

The table shows that there are four variables do not have and missing values, and the other variables have different number of missing values.

We can visualize the missing values to see the big picture!

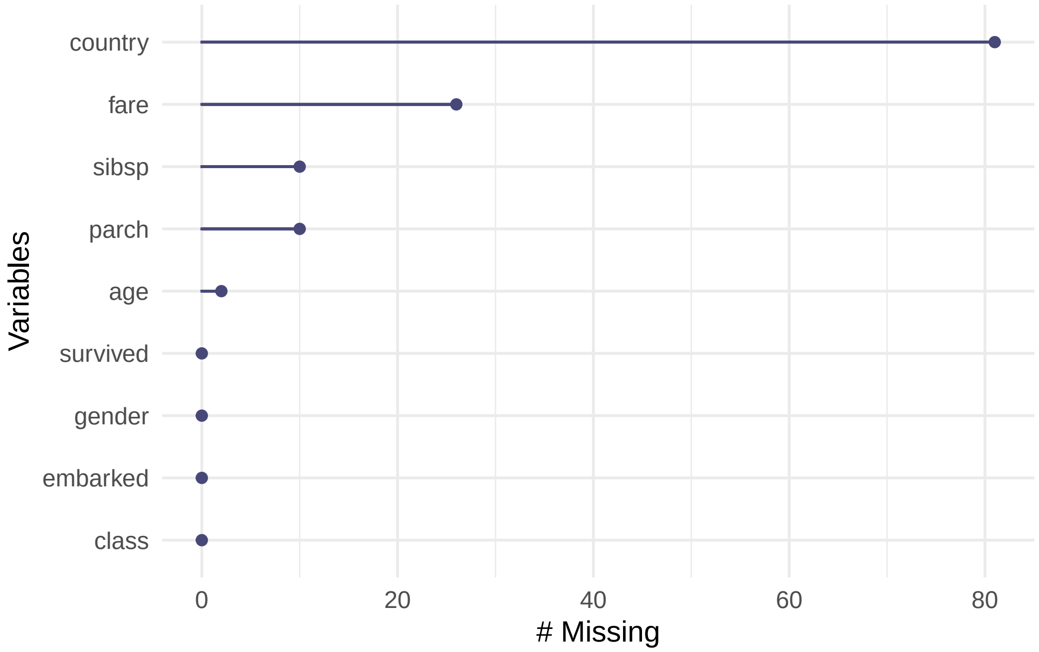
```
vis_miss(titanic)
```



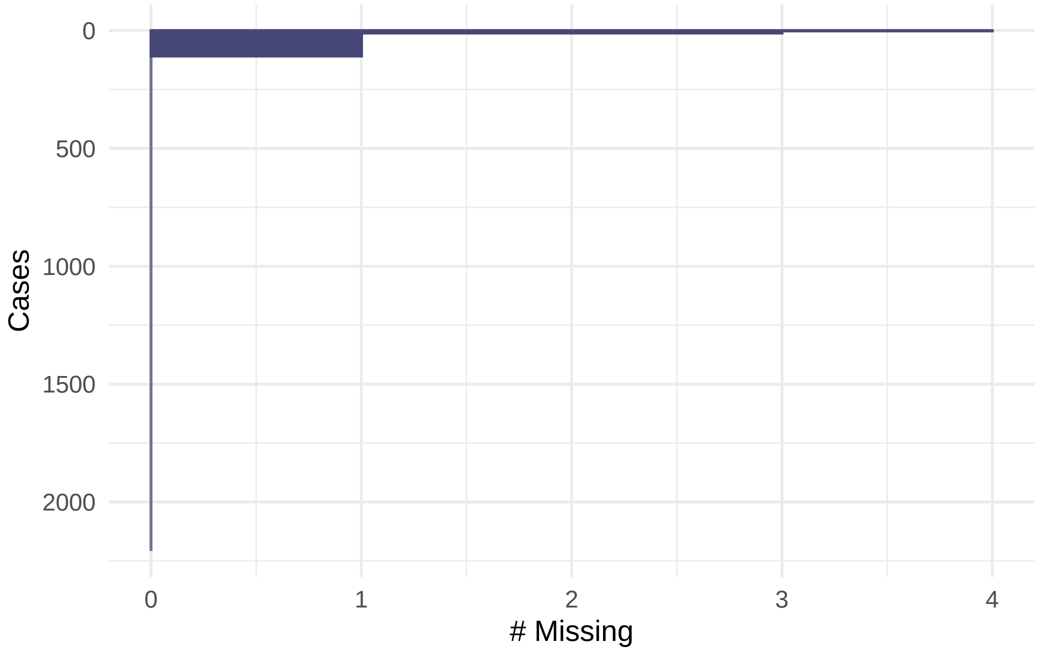
The graph above shows that the 0.6% of the observations is missing. Most of these missing values is in the `country` feature (aka variable). Also, the variables `fare`, `sibsp`, `parch`, and `age` have some missing values.

We can also visualize the missing values by variable and observation level.

```
gg_miss_var(titanic)
```



```
gg_miss_case(titanic)
```


Cases
# Missing