

p8130 HW4 Regression

Eleanor Zhang

11/11/2018

Problem 2 Heart disease

We are interested in if there is an association between **total cost** in dollars diagnosed with heart disease and the **number of ER visits**. Other factors will be adjusted later on.

a) short description of data and look at the data

```
heart_disease <- read_csv("../data/HeartDisease.csv")
```

```
## Parsed with column specification:
## cols(
##   id = col_integer(),
##   totalcost = col_double(),
##   age = col_integer(),
##   gender = col_integer(),
##   interventions = col_integer(),
##   drugs = col_integer(),
##   ERvisits = col_integer(),
##   complications = col_integer(),
##   comorbidities = col_integer(),
##   duration = col_integer()
## )
```

```
head(heart_disease)
```

```
## # A tibble: 6 x 10
##       id totalcost  age gender interventions drugs ERvisits complications
##   <int>    <dbl> <int> <int>         <int> <int>   <int>         <int>
## 1     1    179.   63     0             2     1     4             0
## 2     2    319   59     0             2     0     6             0
## 3     3   9311.  62     0            17     0     2             0
## 4     4    281.  60     1             9     0     7             0
## 5     5  18727.  55     0             5     2     7             0
## 6     6    453.  66     0             1     0     3             0
## # ... with 2 more variables: comorbidities <int>, duration <int>
```

```
str(heart_disease)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':   788 obs. of  10 variables:
## $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
## $ totalcost : num  179 319 9311 281 18727 ...
## $ age      : int  63 59 62 60 55 66 64 45 68 64 ...
## $ gender   : int  0 0 0 1 0 0 1 1 0 1 ...
## $ interventions: int  2 2 17 9 5 1 2 3 6 3 ...
## $ drugs    : int  1 0 0 0 2 0 0 0 2 0 ...
## $ ERvisits  : int  4 6 2 7 7 3 3 5 5 2 ...
## $ complications: int  0 0 0 0 0 0 0 0 0 0 ...
```

```
## $ comorbidities: int 3 0 5 2 0 4 1 1 4 0 ...
## $ duration      : int 300 120 353 332 18 296 247 82 334 85 ...
## - attr(*, "spec")=List of 2
## ..$ cols      :List of 10
## .. ..$ id      : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ totalcost : list()
## .. .. ..- attr(*, "class")= chr "collector_double" "collector"
## .. ..$ age      : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ gender   : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ interventions: list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ drugs    : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ ERvisits  : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ complications: list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ comorbidities: list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## .. ..$ duration   : list()
## .. .. ..- attr(*, "class")= chr "collector_integer" "collector"
## ..$ default: list()
## .. ..- attr(*, "class")= chr "collector_guess" "collector"
## ..- attr(*, "class")= chr "col_spec"
```

```
names(heart_disease)
```

```
## [1] "id"          "totalcost"   "age"         "gender"
## [5] "interventions" "drugs"       "ERvisits"    "complications"
## [9] "comorbidities" "duration"
```

In this dataset, there are 788 observations of patients with 10 variables:

- **id**: patient id
- **totalcost**: total cost (\$) of patients who are diagnosed with heart disease
- **age**: age of patients
- **interventions**: number of interventions (integers)
- **drugs**: ? number of drugs.
- **ERvisits**: number of ER visits
- **complications**: number of complications
- **comorbidities**: number of co-presence of other diseases in addition to heart disease
- **duration**: duration of heart disease (in days)

Based on our investigation interest, the main outcome is **total cost** of patients with heart disease and the main predictor is **ERvisits** (number of ER visits). Other important covariates also need to be considered because they could potentially have differential effect on the association relationship between our main predictor and main outcome, including age, interventions, drugs used, complications, and duration of disease. We will first take a look at these variables:

i) First we took a look at the main outcome and main predictor

number summaries for variables:

```
variable_set1 <- dplyr::select(heart_disease, totalcost, ERvisits, everything(), -c(id, gender, complications))
variable_set2 <- dplyr::select(heart_disease, gender, complications)
```

```
knitr::kable(summary(variable_set1))
```

totalcost	ERvisits	age	interventions	drugs	comorbidities	duration
Min. : 0.0	Min. : 0.000	Min. :24.00	Min. : 0.000	Min. :0.0000	Min. : 0.000	Min. : 0.00
1st Qu.: 161.1	1st Qu.: 2.000	1st Qu.:55.00	1st Qu.: 1.000	1st Qu.:0.0000	1st Qu.: 0.000	1st Qu.: 41.75
Median : 507.2	Median : 3.000	Median :60.00	Median : 3.000	Median :0.0000	Median : 1.000	Median :165.50
Mean : 2800.0	Mean : 3.425	Mean :58.72	Mean : 4.707	Mean :0.4467	Mean : 3.767	Mean :164.03
3rd Qu.: 1905.5	3rd Qu.: 5.000	3rd Qu.:64.00	3rd Qu.: 6.000	3rd Qu.:0.0000	3rd Qu.: 5.000	3rd Qu.:281.00
Max. :52664.9	Max. :20.000	Max. :70.00	Max. :47.000	Max. :9.0000	Max. :60.000	Max. :372.00

```
table(variable_set2)
```

```
##      complications
## gender    0    1    3
##      0 576  32    0
##      1 169  10    1
```

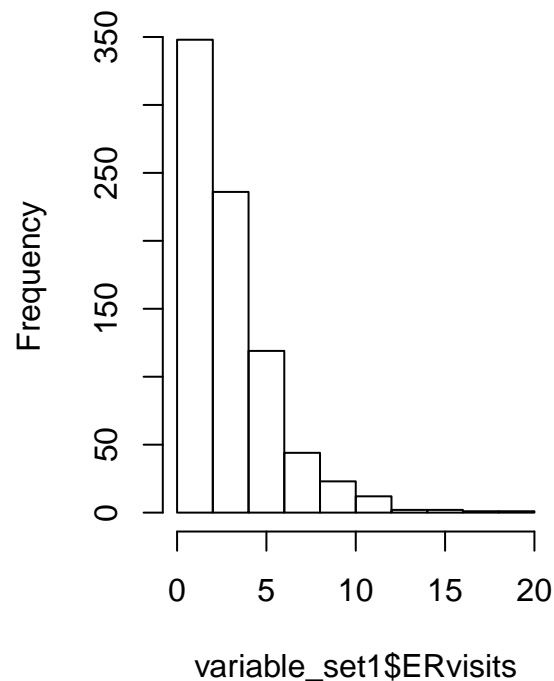
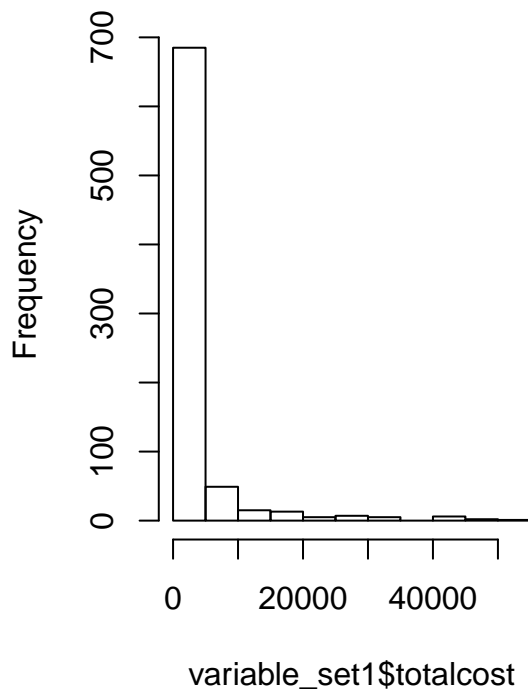
```
#margin.table(table(variable_set2))
```

```
#prop.table(table(variable_set2))
```

Visualize the distribution of these variables

```
par(mfrow = c(1,2))
hist(variable_set1$totalcost)
hist(variable_set1$ERvisits)
```

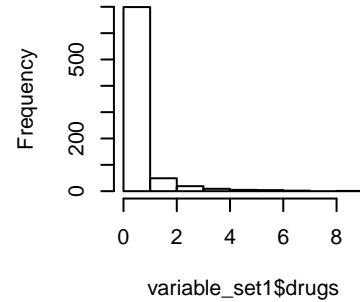
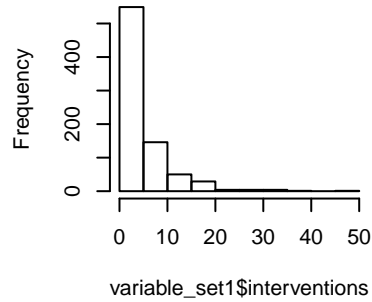
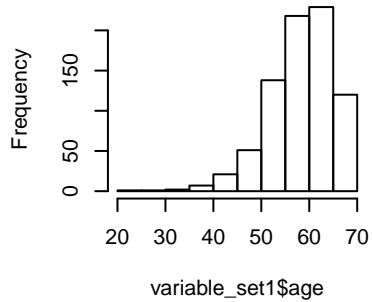
Histogram of variable_set1\$totalcost Histogram of variable_set1\$ERvisits



```
par(mfrow = c(2,3))
hist(variable_set1$age)
hist(variable_set1$interventions)
```

```
hist(variable_set1$drugs)
hist(variable_set1$comorbidities)
hist(variable_set1$duration)
```

Histogram of variable_set1\$age **Histogram of variable_set1\$interventions** **Histogram of variable_set1\$drugs**



Histogram of variable_set1\$comorbidities **Histogram of variable_set1\$duration**

