December 5, 2021

Assignment – 1

AST 405: Lifetime data analysis

The Western Collaborative Group Study (WCGS) (Rosenman et al., 1966) was designed to test the hypothesis that the so-called Type A behavior pattern (TABP) - "characterized particularly by excessive drive, aggressiveness, and ambition, frequently in association with a relatively greater preoccupation with competitive activity, vocational deadlines, and similar pressures" – is a cause of CHD.

A total of 3524 men aged 39–59 and employed in the San Francisco Bay or Los Angeles areas were enrolled in 1960 and 1961. In addition to determinations of behavior pattern, the initial examination included medical and parental history, socioeconomic factors, exercise, diet, smoking, alcohol consumption, diet, serum lipid and lipoprotein studies, blood coagulation studies, and cardiovascular examination. Men continuing in the study were re-examined annually and follow-up for CHD incidence was terminated in 1969.

Download `wcgs.xls` file from the google classroom and create an R object (data frame) `wcgs`, which has the following variables. You can also download the `wcgs.Rdata` file to get the R object `wcgs` (e.g. use the R code `>load("wcgs.Rdata")` to get `wcgs` object in the R environment that you are using.)

```
## Rows: 3,154
## Columns: 22
## $ age      <dbl> 50, 51, 59, 51, 44, 47, 40, 41, 50, 43, 59, 54, 48, 39, 49, 5~
## $ arcus    <dbl> 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1~
## $ behpat   <chr> "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "A1", "~
## $ bmi      <dbl> 31.32101, 25.32858, 28.69388, 22.14871, 22.31303, 27.11768, 2~
## $ chd69    <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No", "~
## $ chol     <dbl> 249, 194, 258, 173, 214, 206, 190, 212, 130, 233, 181, 214, 2~
## $ dbp      <dbl> 90, 74, 94, 80, 80, 76, 78, 84, 70, 80, 86, 76, 78, 74, 80, 7~
## $ dibpat   <chr> "Type A", "Type A", "Type A", "Type A", "Type A", "Type A", "~
## $ height   <dbl> 67, 73, 70, 69, 71, 64, 70, 70, 71, 68, 72, 67, 71, 70, 73, 7~
## $ id       <dbl> 2343, 3656, 3526, 22057, 12927, 16029, 3894, 11389, 12681, 10~
## $ lnsbp    <dbl> 4.882802, 4.787492, 5.062595, 4.836282, 4.836282, 4.753590, 4~
## $ lnwght   <dbl> 5.298317, 5.257495, 5.298317, 5.010635, 5.075174, 5.062595, 5~
## $ ncigs    <dbl> 25, 25, 0, 0, 0, 80, 0, 25, 0, 25, 10, 0, 20, 0, 4, 0, 0, 20,~
## $ sbp      <dbl> 132, 120, 158, 126, 126, 116, 122, 130, 112, 120, 130, 118, 1~
## $ smoke    <chr> "Yes", "Yes", "No", "No", "No", "Yes", "No", "Yes", "No", "Ye~
## $ t1       <dbl> -1.6333529, -4.0633659, 0.6397287, 1.1217681, 2.4250107, -0.7~
## $ time169  <dbl> 1367, 2991, 2960, 3069, 3081, 2114, 2929, 3010, 3104, 2861, 2~
## $ typchd69 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ uni      <dbl> 0.48607379, 0.18595429, 0.72779906, 0.62446356, 0.37897763, 0~
## $ weight   <dbl> 200, 192, 200, 150, 160, 158, 162, 160, 195, 187, 206, 152, 1~
## $ wghtcat  <chr> "170-200", "170-200", "170-200", "140-170", "140-170", "140-1~
## $ agec     <chr> "46-50", "51-55", "56-60", "51-55", "41-45", "46-50", "35-40"~
```

Run the following R codes to create your own data set `mydat`.

```
> library(tidyverse)
> sid <- 203 # replace 203 by your class roll number (numeric part only)
> set.seed(sid)
> mydat <- sample_n(wcgs, size = n(), replace = T)
```

Use R object `mydat` to answer the following questions.

1. For the variables `age`, `arcus`, `behpat`, `bmi`, and `chd69`, obtain appropriate (numeric) descriptive statistics. Note `age` and `bmi` are continuous variables, and others are categorical variable.
2. Examine whether each of the variables `age`, `bmi`, `arcus`, and `behpat` has significant effect on `chol` (a continuous variable). You need to mention appropriate null and alternative hypothesis, test statistic, decision criterion, etc.
3. Examine whether `behpat` and `chd69` are significantly associated. Both `behpat` and `chd69` are categorical variables. You need to mention the appropriate null and alternative hypothesis, test statistic, decision criterion, etc.
4. Consider a regression model `chol` on `age` and interpret the results.
5. Consider a regression model `chol` on `dibpat` and interpret the results.
6. Consider a regression model `chol` on `behpat` and interpret the results.
7. Consider a regression model `chol` on `age` and `dibpat`, and interpret the results. Compare the results with the model considered in 4 and 5.
8. Create a variable `age40` by subtracting 40 from the variable `age`. consider a regression model `chol` on `age40` and interpret the results. What is the main difference of this model compared to the model 4.
9. Consider a regression model to compare the effect of `dibpat` on `chol` between different levels of `smoke`, and interpret the results.

Prepare a hand-written solutions (report) of the above questions and upload the scan copies of the solution to google classroom. Your solution must include a cover-page that clearly mention your name, class roll number, etc. You need to add R codes used for the analysis at the end of your solution.

References

Rosenman, R. H., Friedman, M., Straus, R., Wurm, M., Jenkins, C. D., and Messinger, H. B. (1966). Coronary heart disease in the western collaborative group study: A follow-up experience of two years. *JAMA*, 195(2):86–92.