December 12, 2021

Solution of 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.)

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.

	Overall
n	3154
age (mean (SD))	46.30 (5.52)
arcus = 1 (%)	934 (29.6)
behpat (%)	
A1	264 (8.4)
A2	1327 (42.1)
B3	1229 (39.0)
B4	334 (10.6)
bmi (mean (SD))	24.49 (2.57)
chd69 = Yes (%)	253 (8.0)

```
## computation of mean, SD, median
mydat %>%
  summarise(mage = mean(age), sage = sd(age),
            mdage = median(age), mbmi = mean(bmi),
            sbmi = sd(bmi), mdbmi = median(bmi))
###
### another way
mydat %>%
  summarise(across(c(age, bmi),
                   .fns = list(mean = mean, sd = sd, median = median),
                   .names = "{.col}_{.fn}"))
###
### computation of frequency and corresponding proportions
wcgs %>%
  count(behpat) %>%
  mutate(`%` = 100 * n / sum(n))
```

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.

term	estimate	statistic	p.value
age	0.070	3.937	<0.001
bmi	0.079	4.412	<0.001
arcus	-11.905	-7.156	<0.001

term	df	sumsq	meansq	statistic	p.value
behpat	3	12502.22	4167.407	2.262	0.079
Residuals	3136	5777036.91	1842.167	NA	NA

```r
## Correlation test
cor.test(mydat$chol, mydat$chol)
## to get nicer output, cor.test object
broom::tidy(cor.test(mydat$chol, mydat$chol))
##
## two-sample t.test
t.test(chol ~ arcus, data = mydat)
broom::tidy(t.test(chol ~ arcus, data = mydat))
#
## Anova
aov(chol ~ behpat, data = mydat)
broom::tidy(aov(chol ~ behpat, data = mydat))
```

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.

	CHD	
behpat	No	Yes
A1	230	34
A2	1176	151
B3	1174	55
B4	321	13

statistic	p.value
57.385	<0.001

```
## Contingency table
mydat %>%
  count(chd69, behpat) %>%
  pivot_wider(names_from = chd69, values_from = n)
#
## Chi-square test
chisq.test(mydat$chd69, mydat$behpat)
broom::tidy(chisq.test(mydat$chd69, mydat$behpat))
```

4. Consider a regression model `chol` on `age` and interpret the results.

term	estimate	std.error	statistic	p.value
(Intercept)	200.630	6.466	31.026	<0.001
age	0.546	0.139	3.937	<0.001

$$\widehat{\text{chol}} = 200.63 + 0.546(\text{age})$$

- For 1-year increase of age, average cholesterol level increases for about 0.546 unit.

```r
# regression model for chol on age
lm(chol ~ age, data = mydat)
broom::tidy(lm(chol ~ age, data = mydat))
```

5. Consider a regression model `chol` on `dibpat` and interpret the results.

term	estimate	std.error	statistic	p.value
(Intercept)	227.745	1.078	211.220	<0.001
dibpatType B	-3.704	1.532	-2.419	0.016

$$\widehat{\text{chol}} = 227.745 - 3.704(\text{dibpat}_{\text{Type B}})$$

- Mean cholesterol level of Type A subjects is 227.745 unit.
- The difference of mean cholesterol between Type A and Type B subjects is 3.704 unit.
- Mean cholesterol level of Type A subjects is 3.704 unit higher than that of Type B subjects.

```r
# regression model for chol on dibpat
lm(chol ~ dibpat, data = mydat) %>%
  broom::tidy()
```

6. Consider a regression model `chol` on `behpat` and interpret the results.

term	estimate	std.error	statistic	p.value
(Intercept)	227.790	2.652	85.905	<0.001
behpatA2	-0.054	2.903	-0.019	0.985
behpatB3	-3.202	2.921	-1.096	0.273
behpatB4	-5.778	3.549	-1.628	0.104

$$\widehat{\text{chol}} = 227.79 - 0.054(\text{behpat}_{\text{A2}}) - 3.202(\text{behpat}_{\text{B3}}) - 5.778(\text{behpat}_{\text{B4}})$$

- Mean cholesterol level of Type A1 subjects is 227.790 unit.
- Mean cholesterol level of Type A2 subjects is 0.054 unit lower compared to that of Type A1 subjects.
- Similarly, mean cholesterol levels of Type B3 and B4 subjects are 3.202 and 5.778 unit lower compared to that of Type A1 subjects, respectively.

```r
# regression model for chol on dibpat
lm(chol ~ behpat, data = mydat) %>%
  broom::tidy()
```

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.

term	estimate	std.error	statistic	p.value
(Intercept)	203.350	6.591	30.853	<0.001
age	0.522	0.139	3.752	<0.001
dibpatType B	-3.229	1.534	-2.105	0.035

$$\widehat{\text{chol}} = 227.79 - 0.054(\text{behpat}_{\text{A2}}) - 3.202(\text{behpat}_{\text{B3}}) - 5.778(\text{behpat}_{\text{B4}})$$

- Mean cholesterol level of subjects with age 0 and behavioral pattern A is 202.555 unit.
- For 1-year increase of age, mean cholesterol level increases about 0.523 unit, on an average provided behavioral pattern remains fixed.
- On an average, cholesterol level of Type B subjects is 0.908 unit lower compared to that of Type A subjects provided age remains fixed.

```r
# regression model for chol on dibpat
lm(chol ~ age + dibpat, data = mydat) %>%
  broom::tidy()
```

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.

term	estimate	std.error	statistic	p.value
(Intercept)	222.472	1.161	191.702	<0.001
age40	0.546	0.139	3.937	<0.001

$$\widehat{\text{chol}} = 222.472 + 0.546(\text{age40})$$

- On an average, cholesterol level of a 40-year old subject is 222.472 unit
- For 1-year increase of age, mean cholesterol increases about 0.546 unit

```
lm(chol ~ age40,
   data = mydat %>%
     mutate(age40 = age - 40)) %>%
  broom::tidy()
```

9. Consider a regression model to compare the effect of `dibpat` on `chol` between different levels of `smoke`, and interpret the results.

smoke	term	estimate	std.error	statistic	p.value
No	(Intercept)	225.607	1.451	155.483	<0.001
No	dibpatType B	-7.145	2.015	-3.546	<0.001
Yes	(Intercept)	229.916	1.587	144.887	<0.001
Yes	dibpatType B	1.018	2.316	0.440	0.660

term	estimate	std.error	statistic	p.value
(Intercept)	225.607	1.511	149.336	<0.001
dibpatType B	-7.145	2.098	-3.406	<0.001
smokeYes	4.310	2.145	2.009	0.045
dibpatType B:smokeYes	8.163	3.055	2.672	0.008

$$\widehat{\text{chol}} = 225.607 - 7.145(\text{dibpat}_{\text{Type B}}) + 4.31(\text{smoke}_{\text{Yes}}) + 8.163(\text{dibpat}_{\text{Type B}} \times \text{smoke}_{\text{Yes}})$$

- On an average, cholesterol level of a non-smoker Type A subject is 225.607 unit
- Among non-smokers, mean cholesterol level of Type B subjects is -7.145 unit lower compared to Type A subject
- Among Type A subjects, mean cholesterol level of smokers is 4.310 unit higher compared to non-smokers
- Difference of mean cholesterol level between smokers and non-smokers is 8.163 unit higher in Type B subjects compared to that of Type A subjects (Difference-in-differences)

```r
## model fit to two separate data
mydat %>%
  group_by(smoke) %>%
  do(broom::tidy(lm(chol ~ dibpat, data = .)))
###
## model fit with interaction
lm(chol ~ dibpat * smoke, data = mydat) %>%
  broom::tidy()
```

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.