

---

# Assignment On Lifetime Data Analysis AST 405

---

*Submitted By:*

Shafayet Khan Shafee

FH-033-011

4th Year

*Submitted To:*

Dr. Mahbub Latif

Professor

ISRT, DU

January 22, 2022



# Contents

<b>Answers to Questions</b>	<b>2</b>
Question 01 . . . . .	2
Question 02 . . . . .	3
Question 03 . . . . .	5
Question 04 . . . . .	6
Question 05 . . . . .	6
Question 06 . . . . .	7
Question 07 . . . . .	7
Question 08 . . . . .	8
Question 09 . . . . .	9
<b>R-code</b>	<b>10</b>

# Answers to Questions

## Question 01

**Answer:** For the continuous variables `age` and `bmi`, mean and standard deviation (SD) is obtained. And for categorical variable `arcus`, `behpat`, and `chd69`, frequency and proportion (in percentage) is obtained as descriptive statistics, which are shown in the following table:

Table 1: Descriptive Statistics

	level	Overall
n		3154
age (mean (SD))		46.35 (5.56)
bmi (mean (SD))		24.48 (2.55)
arcus (%)	0	2219 (70.4)
	1	934 (29.6)
behpat (%)	A1	275 ( 8.7)
	A2	1290 (40.9)
	B3	1236 (39.2)
	B4	353 (11.2)
chd69 (%)	No	2900 (91.9)
	Yes	254 ( 8.1)

## Question 02

**Answer:**

**Effect of age and bmi:** To check whether each of **age** and **bmi** has significant effect on **chol**, we need to do correlation test (pearson).

For each correlation test of **age** and **bmi** with **chol**, the hypotheses are,

$$H_o : \rho = 0$$

$$H_a : \rho \neq 0$$

where  $\rho$  is the population correlation coefficient.

The appropriate test statistic for testing the hypothesis is,

$$t_o = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

where  $r$  is the sample correlation coefficient and  $t_o$  follows  $t_{n-2}$  distribution under  $H_o$ .

Here we will reject  $H_o$  if the associated **p.value** is less than  $\alpha = 0.05$ , where  $\alpha$  is the level of significance.

Now from Table: 2, we can see that both of the variables **age** and **bmi** is significantly associated with **chol** variable.

Table 2: Correlation Test (age, bmi) and T-test (arcus) with chol

term	estimate	statistic	p.value
age	0.081	4.550	<.001
bmi	0.060	3.385	<.001
arcus	-12.672	-7.604	<.001

**Effect of arcus:** Now to check the effect of **arcus** (a categorical variable with level 0 and 1) on **chol**, we can do two sample t-test, where one sample is people with **arcus** 0 and other sample is people with **arcus** 1. Therefore,  $\mu_o$  and  $\mu_1$  are the population mean of arcus 0 and arcus 1 group respectively.

Then our hypotheses are:

$$H_o : \mu_o = \mu_1$$

$$H_a : \mu_o \neq \mu_1$$

and appropriate test statistic (assuming unequal variance),

$$t_o = \frac{\bar{x}_o - \bar{x}_1}{\sqrt{\frac{s_o^2}{n_o} + \frac{s_1^2}{n_1}}}$$

where  $\bar{x}$  and  $s^2$  denotes sample mean and sample variance with subscript 0 and 1 for **arcus** group 0 and 1 respectively and  $n_o$  and  $n_1$  are the corresponding sample sizes.

Here we will reject  $H_o$  if the associated **p.value** is less then  $\alpha = 0.05$ , where  $\alpha$  is the level of significance.

Then from Table: 2, since **p.value** is less than 0.001 we can conclude that mean **chol** differs significantly for 0 and 1 group of **arcus**, that is, **arcus** has significant effect on **chol**.

**Effect of behpat:** To check the effect of **behpat** on **chol**, we can do oneway ANOVA. In this case the hypotheses are:

$$H_o : \mu_1 = \mu_2$$

$$H_a : \mu_i \neq \mu_j \quad \text{for at least one } i \neq j$$

and test statistic is  $F_o = \frac{MS_{reg}}{MS_E}$ . We will reject  $H_o$  if associated **p.value** is less than 0.05.

Table 3: One Way Analysis of Variance for chol on behpat

term	df	sumsq	meansq	statistic	p.value
behpat	3	30741.67	10247.223	5.475	<.001
Residuals	3135	5867632.50	1871.653	NA	NA

Then from Table: 3, we conclude that mean **chol** differs significantly over the levels of **behpat**.

### Question 03

**Answer:** To examine the association between `behpat` and `chd69`, the hypotheses are:

$H_o$  : There's no association between `behpat` and `chd69`.

$H_a$  : There's association between `behpat` and `chd69`.

Here, the test statistics is:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where,  $O_{ij}$  is observed frequency and  $E_{ij}$  is the expected frequency and  $r, c$  is the row and column number of contingency table 4

Here, we would reject  $H_o$  if associated `p.value` corresponding to test statistics is less the 0.05.

	CHD	
	No	Yes
behpat		
A1	246	29
A2	1158	132
B3	1164	72
B4	332	21

Table 4: Contingency table

statistic	p.value
20.978	<.001

Table 5: chi-square test

Now from the Table 5, since `p.value` is less than 0.001, we can conclude that there's a significant association between `behpat` and `chd69`.

## Question 04

Answer:

Table 6: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	196.971	6.483	30.381	<.001
age	0.632	0.139	4.550	<.001

The fitted regression line is:

$$\widehat{\text{chol}} = 196.97 + 0.63(\text{age}) \quad (1)$$

Then, the cholesterol level is expected to increase by 0.63 unit for 1-year increase of age.

## Question 05

Answer:

Table 7: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	228.352	1.097	208.102	<.001
dibpatType B	-4.147	1.546	-2.682	0.007

The fitted regression line is:

$$\widehat{\text{chol}} = 228.35 - 4.15(\text{dibpat}_{\text{Type B}}) \quad (2)$$

So, we can say that, Mean cholesterol level of **Type A dibpat** subjects is 228.35. Also, mean cholesterol level of **Type A dibpat** subjects is 4.15 unit higher than **Type B dibpat** subjects.

## Question 06

Answer:

Table 8: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	235.073	2.618	89.778	<.001
behpatA2	-8.150	2.883	-2.827	0.005
behpatB3	-10.226	2.894	-3.533	<.001
behpatB4	-13.122	3.491	-3.759	<.001

The fitted regression line is:

$$\widehat{\text{chol}} = 235.07 - 8.15(\text{behpat}_{A2}) - 10.23(\text{behpat}_{B3}) - 13.12(\text{behpat}_{B4}) \quad (3)$$

So, we can say, since **Type A1** Behavior pattern subjects are reference group,

- Mean cholesterol level of **Type A1** subjects is 235.07 unit.
- Mean cholesterol level of **Type A2** subjects is 8.15 unit lower than that of **Type A1** subjects.
- Mean cholesterol level of **Type B3** subjects is 10.23 unit lower than that of **Type A1** subjects.
- Mean cholesterol level of **Type B4** subjects is 13.12 unit lower than that of **Type A1** subjects.

## Question 07

Answer:

Table 9: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	200.191	6.637	30.163	<.001
age	0.600	0.140	4.302	<.001
dibpatType B	-3.469	1.550	-2.238	0.025



The fitted regression line is:

$$\widehat{\text{chol}} = 200.19 + 0.6(\text{age}) - 3.47(\text{dibpat}_{\text{Type B}}) \quad (4)$$

Since **Type A1** subjects are reference group,

- Mean cholesterol level of subjects with age 0 and **Type A** pattern is 200.19 unit.
- Mean cholesterol level increases about 0.6 for 1-year increase of age, holding dibpat fixed.
- Mean cholesterol level of **Type B2** subjects is 3.47 unit lower than that of **Type A** dibpat subjects, holding the subject's age fixed.

Here, Both the regression coefficients corresponding to age and dibpat (Eq 4) have changed from the case of simple linear regression in Eq 1 and Eq 2. And also the value of  $R_{adj}^2$  increased.

## Question 08

**Answer:**

Table 10: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	222.252	1.171	189.755	<.001
age40	0.632	0.139	4.550	<.001

The new fitted model after subtracting 40 from the variable **age** is:

$$\widehat{\text{chol}} = 222.25 + 0.63(\text{age40}) \quad (5)$$

- Since the explanatory variable is age minus 40, we can say, the mean cholesterol level of a 40 year old subject is 222.25.
- Mean cholesterol level is expected to increase by 0.63 for 1 year increase of age.

The main difference of this model (Eq 1) compared to model 5 is that, we can interpret the intercept term for this model logically.

## Question 09

Answer:

Table 11: Estimate of Model Parameters

term	estimate	std.error	statistic	p.value
(Intercept)	223.702	1.597	140.043	<.001
dibpatType B	-3.644	2.175	-1.675	0.094
smokeYes	8.728	2.189	3.988	<.001
dibpatType B:smokeYes	0.265	3.085	0.086	0.932

In this case, the fitted model is:

$$\widehat{\text{chol}} = 223.7 - 3.64(\text{dibpat}_{\text{Type B}}) + 8.73(\text{smoke}_{\text{Yes}}) + 0.27(\text{dibpat}_{\text{Type B}} \times \text{smoke}_{\text{Yes}}) \quad (6)$$

Here the reference groups are **Non-smoker** and **dibpat Type A** subjects. So,

- The mean cholesterol level of Type A non smoker subject is 223.7 unit.
- Among the non-smokers, mean cholesterol level of dibpat **Type B** subjects is 3.64 unit lower compared to **Type A** subject.
- Among the **Type A** subjects, mean cholesterol level of smokers is 8.73 unit higher than that of non-smokers.
- Difference of mean cholesterol level between smokers and non-smokers is 0.27 unit higher in **Type B** dibpat subjects compared to that of **Type A** dibpat subjects.

## R-code

```
knitr::opts_chunk$set(  
  echo = FALSE,  
  message = FALSE,  
  warning = FALSE  
)  
  
## ----- package setup -----  
  
library(dplyr)  
library(purrr)  
library(knitr)  
library(broom)  
library(tableone)  
library(kableExtra)  
library(equationomatic)  
  
## ----- data setup -----  
  
load(here::here("data", "wcgs.Rdata"))  
sid <- 011  
set.seed(sid)  
mydat <- sample_n(wcgs, size = n(), replace = TRUE)  
  
## ----- utility functions -----  
  
kab_tab <- function(tab, ...) {  
  knitr::kable(tab,  
    format = "latex",  
    booktabs = TRUE,  
    digits = 3,  
    ...)  
}
```

```

p_format <- function(pval) {
  ifelse(pval < .001, "<.001", as.character(round(pval, 3)))
}

mod_tab <- function(mod, ...) {
  mod %>%
    tidy() %>%
    mutate(p.value = p_format(p.value)) %>%
    kable_tab(align = "lrrrr",
              caption = "Estimate of Model Parameters", ...) %>%
    kable_styling(latex_options = "HOLD_position")
}

reg_eq <- function(mod, ref, ...) {
  extract_eq(mod,
             use_coefs = TRUE,
             intercept = "beta",
             wrap = TRUE,
             label = paste0("eq", ref),
             ...)
}

params <- function(mod, param, dec = 2) {
  round(mod$coefficients[[param]], dec)
}

## ----- Code for Question-01 -----

tab <- CreateTableOne(
  data = mydat,
  vars = c("age", "bmi", "arcus", "behpat", "chd69"),
  factorVars = "arcus",
  addOverall = TRUE
)

```

```
tab_p <- print(tab, showAllLevels = TRUE, printToggle = FALSE)
```

```
kab_tab(tab_p, caption = "Descriptive Statistics") %>%
```

```
  kable_styling(latex_options = "HOLD_position")
```

```
## ----- Code for Question-02 -----
```

```
arcus <- t.test(chol ~ arcus, data = mydat) %>%
```

```
  tidy() %>%
```

```
  mutate(term = "arcus") %>%
```

```
  select(term, estimate, statistic, p.value)
```

```
mydat %>%
```

```
  select(age, bmi) %>%
```

```
  map(~ cor.test(x = .x, y = mydat$chol)) %>%
```

```
  map_dfr(broom::tidy, .id = "term") %>%
```

```
  select(term:p.value) %>%
```

```
  bind_rows(arcus) %>%
```

```
  mutate(p.value = p_format(p.value)) %>%
```

```
  kab_tab(align = "lrrr",
```

```
    caption = "Correlation Test (age, bmi) and T-test (arcus) with chol") %>%
```

```
  kable_styling(latex_options = "HOLD_position")
```

```
anova(lm(chol ~ behpat, data = mydat)) %>%
```

```
  tidy() %>%
```

```
  mutate(p.value = p_format(p.value)) %>%
```

```
  kab_tab(align = "lrrrrr",
```

```
    caption = "One Way Analysis of Variance for chol on behpat") %>%
```

```
  kable_styling(latex_options = "HOLD_position")
```

```
## ----- Code for Question-03 -----
```

```
df_cont <- mydat %>% janitor::tabyl(behpat, chd69)
```

```
tab_cont <- df_cont %>%  
  kab_tab() %>%  
  add_header_above(header = c(" " = 1, "CHD" = 2))
```

```
tab_chi <- df_cont %>%  
  janitor::chisq.test() %>%  
  tidy() %>%  
  select(statistic, p.value) %>%  
  mutate(p.value = p_format(p.value)) %>%  
  kab_tab()
```

```
tab_side <- c(  
  "\\begin{table}[H]  
    \\begin{minipage}{.5\\linewidth}  
      \\centering",  
  tab_cont,  
  "\\caption{Contingency table}  
  \\label{Table-04}  
  \\end{minipage}%  
  \\begin{minipage}{.5\\linewidth}  
    \\centering",  
  tab_chi,  
  "\\caption{chi-square test}  
  \\label{Table-05}  
  \\end{minipage}  
  \\end{table}"  
)
```

```
## ----- Code for Question-04 -----
```

```
m1 <- lm(chol ~ age, data = mydat)  
m1 %>% mod_tab()
```

```
reg_eq(m1, 1)
```

```
## ----- Code for Question-05 -----
```

```
m2 <- lm(chol ~ dibpat, data = mydat)
m2 %>% mod_tab()

reg_eq(m2, 2)
```

```
## ----- Code for Question-05 -----
```

```
m3 <- lm(chol ~ behpat, data = mydat)
m3 %>% mod_tab()

reg_eq(m3, 3)
```

```
## ----- Code for Question-08 -----
```

```
m4 <- lm(chol ~ age + dibpat, data = mydat)
m4 %>% mod_tab()

reg_eq(m4, 4)
```

```
## ----- Code for Question-08 -----
```

```
mydat %>%
  mutate(age40 = age - 40) %>%
  lm(chol ~ age40, data = .) -> m5
```

```
m5 %>% mod_tab()
```

```
reg_eq(m5, 5)
```

```
## ----- Code for Question-08 -----
```

```
m6 <- lm(chol ~ dibpat * smoke, data = mydat)
m6 %>% mod_tab()

reg_eq(m6, 6)
```