

Exercises day 2 – Linear regression

The dataset provided to you is from an imaginary prospective pregnancy cohort in India. The pregnant women in this cohort were enrolled in the first trimester of their pregnancy and followed up till delivery. At delivery, the birth weight of the baby and gestational age at delivery were documented.

"**pregnancies.dta**" contains the following variables:

id: This variable represents the unique identifier for each participant in the dataset.

gestational_age: The gestational age of the pregnancy, measured from conception to delivery.

birth_weight: The weight of the baby at delivery.

low_birth_weight: A binary variable indicating whether the baby's weight was less than 2500 grams (low birth weight).

bmi: The pre-pregnancy body mass index (BMI) of the mother at enrollment.

bmi_group: Categorical variable representing the pre-pregnancy BMI categories at enrollment. It includes the following categories:

"normal" (corresponding to a BMI group of 1)

"underweight" (corresponding to a BMI group of 2)

"overweight" (corresponding to a BMI group of 3)

"obese" (corresponding to a BMI group of 4)

preterm_birth: A binary variable indicating whether the baby was born before 37 weeks of gestational age (preterm birth).

passive_smoking: A binary variable indicating whether the mother was exposed to passive smoking during pregnancy.

biomass_fuel: A binary variable indicating whether the household used non-LPG fuel sources for cooking (biomass fuel use).

parity: A categorical variable indicating the parity of the mother (number of previous pregnancies). It includes the following categories:

0 (nulliparous, meaning no previous pregnancies)

1 (parous, meaning at least one previous pregnancy)

part_education: The number of years of education completed by the mother.

family_inc: The total family income in Indian Rupees (INR).

1) Load and describe the data.

```
# Load packages
library(tidyverse)
library(ggthemes)
```

```
library(mgcv)

# Set working directory
setwd("C:/...")

# Load the .csv file
pregnancies <- read_csv("pregnancies.csv")
# Describe data
glimpse(pregnancies)
str(pregnancies)
summary(pregnancies)
head(pregnancies)
tail(pregnancies)
typeof(pregnancies)
class(pregnancies)
View(pregnancies)
```

2) Visualize the relationship between gestational age and birth weight using a scatter plot.

```
#Scatterplot gestational age and birth weight
ggplot(
  data = pregnancies,
  mapping = aes(x = birth_weight , y = gestational_age)) +
  geom_point(color = "steelblue", alpha = 0.5) +
  theme_minimal() +
  labs(x = "Birth weight (grams)", y = "Gestational Age (weeks)")
```

3) Perform a linear regression analysis to investigate the association between gestational age and birth weight.

```
> # Fit linear regression model
> model1 <- lm(birth_weight ~ gestational_age, data = pregnancies)
> # Display results
> summary(model1)
```

- a. Are these two characteristics associated with each other?
- b. What seems to be the influence of gestational age at delivery on birth weight?
- c. How do you interpret R^2 in the test report?
- d. What does regression analysis inform beyond a correlation analysis for the relationship between two continuous variables?
- e. What is residual and how is it different from random error?

- 4) Conduct another linear regression analysis to examine the relationship between birth weight and preterm birth

- a. How do you interpret the results?
- b. Can you relate this with t test?

```
# t-test
t.test(birth_weight ~ preterm_birth,
       data = pregnancies,
       var.equal = TRUE)
```

- 5) Visualize the relationship between gestational age and pre-pregnancy BMI using a scatter plot. What are your thoughts on the scatterplot that you see? How are these two measures related to each other?

```
#Scatterplot gestational age and BMI
ggplot(
  data = pregnancies,
  mapping = aes(x = bmi, y = gestational_age)) +
  geom_point(color = "steelblue", alpha = 0.5) +
  theme_minimal() +
  labs(x = "Pre-pregnancy BMI", y = "Gestational Age (weeks)")
```

- 6) Perform a linear regression analysis to assess the association between gestational age and pre-pregnancy BMI.

- a. Interpret the results.
- b. Can you interpret the association between these two variables as you did for the previous example (gestational age and birthweight)? If not, why?
- c. What are the solutions to overcome this problem?

- 7) Treat pre-pregnancy BMI as a categorical variable and conduct a linear regression analysis.

```
# Convert bmi_group to a factor

pregnancies <- pregnancies |>
  mutate(
    bmi_group = factor(
      bmi_group,
      levels = c(1, 2, 3, 4),
      labels = c(
        "Normal",
        "Underweight",
        "Overweight",
```

```

    "obese"
  )
)

```

a. How do you interpret the results?

- 8) Conduct non-linear analysis between gestational age and pre-pregnancy BMI using GAM and interpret the result.

```

model4 <- gam(gestational_age ~ s(bmi), family = gaussian(),
  data = pregnancies)
plot(model4)

```

#or

```

ggplot(
  data = pregnancies,
  mapping = aes(x=bmi,y=gestational_age))+
  geom_point() +
  geom_smooth(method = "gam") +
  scale_x_continuous(limits = c(5,40), n.breaks = 10) +
  labs(x = "Pre-pregnancy BMI", y = "Gestational Age
    (weeks)")

```