

Exercises day 3 - Interactions

For this exercise we have synthetically generated a population of adults aged 50 years and above, with explicitly defined confounding effects and interactions. We then carried out a cross-sectional survey by randomly selecting around 1,000 people from this population. For each person we recorded whether they have ever had a heart attack and 18 lifestyle and health exposures that could potentially be associated with it. The data are synthetic and created for exercises only; the numbers are not real disease rates.

Variables in SurveyData.csv

Variable	Meaning
heart_attack	Outcome: 1 = has ever had a heart attack, 0 = no
smoker	1 = current smoker, 0 = non-smoker
high_bp	1 = has high blood pressure (hypertension), 0 = no
family_history	1 = a parent or sibling has had a heart attack, 0 = no
drinks_coffee, vegetarian, left_handed, owns_fridge, night_shift, drinks_tea, uses_internet, married, street_food, plays_sport, owns_motorbike, wears_glasses, has_pet, spicy_food, social_media	15 further lifestyle and household variables, each coded 1 = yes, 0 = no, used in exercise 13

- 1) Import the data, which is comma-separated values (csv) formatted, and check that it has been correctly imported.

```
# Load packages
library(tidyverse)

# Set working directory
setwd("__directory path__")

# Load the .csv file
SurveyData <- read.csv("SurveyData.csv")

# Check the import
glimpse(SurveyData)
summary(SurveyData)
```

You should see 19 yes/no (0/1) variables and 1,065 observations. The outcome is heart_attack (1 = yes, 0 = no).

From past studies we know that smoking and high blood pressure may be associated with heart attack. In the data these are the variables smoker, high_bp and heart_attack. We will examine these associations on an additive scale (risk differences).

- 2) Perform separate univariate binomial regression analyses with an identity link for smoker and for high_bp.

```
model_smoker <- glm(
  heart_attack ~ smoker,
  family = binomial(link = "identity"),
  data = SurveyData
)
summary(model_smoker)

model_bp <- glm(
  heart_attack ~ high_bp,
  family = binomial(link = "identity"),
  data = SurveyData
)
summary(model_bp)
```

- a. Are smoking and high blood pressure associated with heart attack? What do the intercept and the coefficient mean with an identity link?
- 3) Smoking is our main exposure. Add high blood pressure to the model to check whether it confounds the association between smoking and heart attack.

```
model_both <- glm(
  heart_attack ~ smoker + high_bp,
  family = binomial(link = "identity"),
  data = SurveyData
)
summary(model_both)
```

- a. Does it look like high blood pressure confounds the association of the smoking with heart attack?
- 4) Estimate the interaction coefficient using the observed risks in the four groups.

Interaction coefficient = (Risk of smoking in high BP) – (Risk of smoking in normal BP)

$$\beta_{\text{interaction}} = (R_{\text{Smoker, HighBP}} - R_{\text{NonSmoker, HighBP}}) - (R_{\text{Smoker, NormalBP}} - R_{\text{NonSmoker, NormalBP}})$$

Run below R commands

```
# Risk in the four exposure groups
table_risk <- SurveyData |>
  mutate(
    smoker = ifelse(smoker == 1, "Smoker", "Non-smoker"),
    high_bp = ifelse(high_bp == 1, "High BP", "Normal BP")
  ) |>
  summarise(
    n = n(),
    risk = mean(heart_attack),
    .by = c(smoker, high_bp)
```

)

```
R_Smoker_HighBP <- table_risk |>
  filter(smoker == "Smoker", high_bp == "High BP") |>
  pull(risk)
```

```
R_NonSmoker_HighBP <- table_risk |>
  filter(smoker == "Non-smoker", high_bp == "High BP") |>
  pull(risk)
```

```
R_Smoker_NormalBP <- table_risk |>
  filter(smoker == "Smoker", high_bp == "Normal BP") |>
  pull(risk)
```

```
R_NonSmoker_NormalBP <- table_risk |>
  filter(smoker == "Non-smoker", high_bp == "Normal BP") |>
  pull(risk)
```

```
beta_interaction <- (R_Smoker_HighBP - R_NonSmoker_HighBP) -
  (R_Smoker_NormalBP - R_NonSmoker_NormalBP)
```

```
table_risk
beta_interaction
```

- 5) Estimate the interaction coefficient using the generalized linear regression model with an interaction term on an additive scale (identity link).

```
model_int <- glm(
  heart_attack ~ smoker * high_bp,
  family = binomial(link = "identity"),
  data = SurveyData
)
summary(model_int)
```

- a. What is the difference between the interaction coefficients obtained from the two approaches?
- b. Is there an interaction between smoking and blood pressure?
- c. What additional information have you found from the GLM model compared to the manual approach?
- d. The coefficient for smoker fell from 0.182 in exercise 2 to 0.097 here. Is smoking now less harmful?