Linear regression (running + interpreting)

CRISP R Mini-Course

Day 6

Review from last time

What does the code below do? Explain to your neighbor!

```
library(dplyr)

df <- read.csv("data.csv")

df_new <- df %>%
   mutate(sex = factor(sex, labels = c("Female", "Male"))) %>%
   filter(bmi < 30)

# disease is a binary variable 0/1 indicating presence of disease
tx_disease_table <- table(df_new %>% select(tx, disease))

prop.test(tx_disease_table)
```

data processing

"table" object

two sample z-test

	disease	
tx	0	1
0		
1		

Follow-ups from last time

- Last time, we learned how to perform a t-test comparing the mean of two groups
- Here is alternative way of doing it that does not require creating subsetting datasets

Original way:

```
# method 1
df <- read.csv("data.csv")

df_tx1 <- df %>% filter(tx == 1)
df_tx0 <- df %>% filter(tx == 0)

t.test(df_tx1$age, df_tx0$age)
```

Alternative (better) way:

```
# method 2
df <- read.csv("data.csv")

# the first argument is the formula
# looks like: variable ~ group
# second argument is the dataframe
t.test(age ~ tx, data = df)
```

Today’s agenda

- Linear regression – conceptual tutorial
- Running and interpreting linear regression in R

Linear regression w/ continuous variable (1)

- We are trying to determine if a blood pressure is associated with age. We obtain data from an observational study with both variables.
- We answer this question using linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Age
- We do this in R and obtain the following output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.95507    1.13096   81.31   <2e-16 ***
age          0.46052    0.02166   21.26   <2e-16 ***
```

- How can we interpret this output?

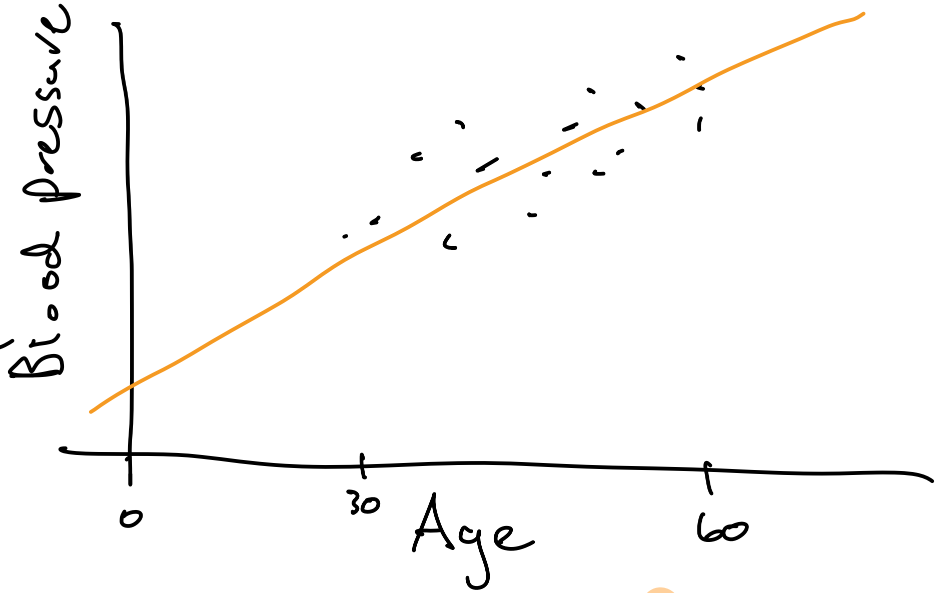
Linear regression w/ continuous variable (2)

- Linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Age

used to calculate CI

p-value

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 91.95507    1.13096   81.31   <2e-16 ***
age          0.46052    0.02166   21.26   <2e-16 ***
```



$$\text{Blood Pressure} = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1}_{\text{slope}} \cdot \text{Age}$$

β_0 estimate: 91.955

predicted BP when age = 0

β_1 estimate: 0.461

mean difference in BP for every 1-year difference in age

p-value: $H_0: \beta_1 = 0$

"significant association" b/w age and BP

Linear regression w/ binary variable (1)

- We are trying to determine if a blood pressure is associated with treatment status. We obtain data from an observational study with both variables.
- *Q: What is one way we can analyze our data to answer this question?*

↳ Could use a t.test

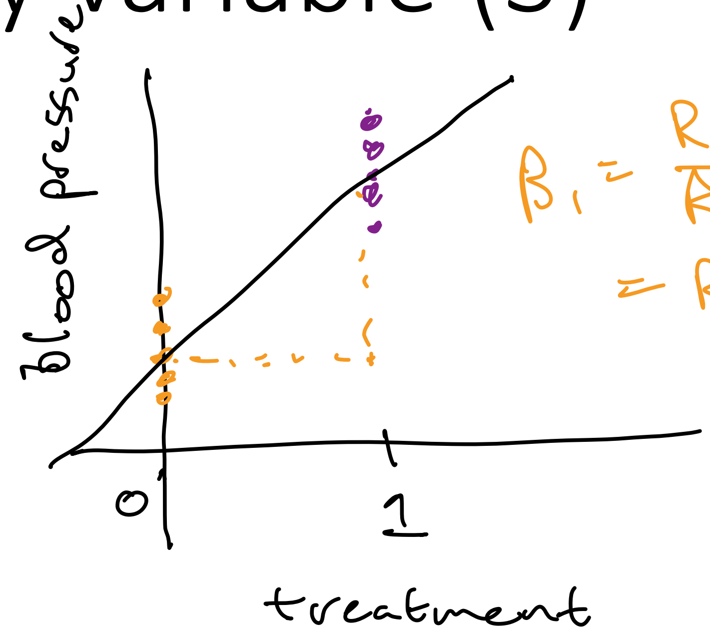
Linear regression w/ binary variable (2)

- We are trying to determine if a blood pressure is associated with treatment status. We obtain data from an observational study with both variables.
- We can also analyze this using linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Treatment status
- We do this in R and obtain the following output:

```
Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      110.4340     0.6407  172.37   <2e-16 ***
treatmentTreated   6.7961     0.7758    8.76   <2e-16 ***
```

Linear regression w/ binary variable (3)

- Linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Treatment status (0/1)



$\beta_1 = \frac{\text{Rise}}{\text{Run}} = \text{Rise}$

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     110.4340     0.6407  172.37   <2e-16 ***
treatmentTreated  6.7961     0.7758    8.76   <2e-16 ***
```

β_0 : 110.43

Mean BP for tx = 0 group

$\text{BloodPressure} = \beta_0 + \beta_1 \text{ treatment}$

β_1 : 6.796

Difference in the mean BP for Tx = 1 and Tx = 0

p-value : $H_0: \beta_1 = 0$

"Significant difference in means"

Linear regression w/ adjustment variable (1)

- We are trying to determine if a ~~disease severity score~~ blood pressure is associated with treatment status adjusting for age. We obtain data from an observational study with all variables.
- We answer this question using linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Treatment status
  - **Adjustment covariate**: Age
- We do this in R and obtain the following output:

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     91.86909    1.13694  80.804   <2e-16 ***
treatmentTreated -0.58538    0.77626  -0.754    0.451
age              0.47019    0.02518  18.675   <2e-16 ***
```

Linear regression w/ adjustment variable (2)

$$BloodPressure = \beta_0 + \beta_1 Treatment + \beta_2 Age$$

- Linear regression:
  - **Outcome**: Blood pressure
  - **Exposure**: Treatment status
  - **Adjustment covariate**: Age

β_0 : 91.869
predicted BP for those w/ tx = 0 and age = 0

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	91.86909	1.13694	80.804	<2e-16	***
treatmentTreated	-0.58538	0.77626	-0.754	0.451	
age	0.47019	0.02518	18.675	<2e-16	***

β_1 : −0.585
Difference in the mean BP for Tx = 1 and Tx = 0 keeping age constant

p-value: after adjusting for age, we do not find a significant association b/w BP and tx

β_2 : 0.470
mean difference in BP for every 1-year difference in age keeping tx status constant

Can construct CI's from estimates and SE

Ex) CI: $(0.5, 5.9)$ expect a p-value < 0.05

Ex) CI: $(-5, 5)$ expect a p-value > 0.05

Ex) CI $(-0.1, 6)$ expect a p-value close to 0.05

Guided tutorial

Today, we will learn the basics of dataset processing.

1. Go to bit.ly/crisp2025.
2. Download Rmd file for today into your CRISP R notes folder.
3. We will go through the tutorial (until the exercises) together! Try to follow along, and type and run the code as I do it.