# Logistic regression (running + interpreting)

CRISP R Mini-Course

Day 7

# Review from last time

outcome

exposure We wish to determine the association between **disease severity score and treatment status**, adjusted by age and sex. We run the following code in R, and obtain the output shown.

*Explain what we are doing and how to interpret the results.*

linear regression

```
# df is a dataframe with age, tx (0/1), sex (0/1), disease_severity
mod <- lm(disease_severity ~ tx + age + sex, data = df)

summary(mod)
confint(mod)
```

formula

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72527    0.87731  -0.827  0.40881
tx (0/1)     1.43562    0.44416   3.232  0.00131 **
age          0.04127    0.02283   1.808  0.07124 .
sex          0.16322    0.44421   0.367  0.71345
```

```
                   2.5 %     97.5 %
(Intercept) -2.448965779 0.99843098
tx           0.562955300 2.30828457
age         -0.003582779 0.08612001
sex         -0.709540823 1.03598060
```

Coefficient on tx : 1.435 (95% CI : 0.564 2.30)

Expected difference in disease Severity Score comparing treated vs untreated participants (of the same age + sex)

On average treated participants have a score of 1.435 pts higher than untreated participants of the same age and sex.

---

tx is a binary variable

if numeric, the lowest number will be considered the ref

---

p-value : 0.00131

↳ difference is significantly different than 0

↳ find signficant association

# This week's schedule and next steps

- This week's schedule:
  - **Today**: Logistic regression
  - **Wednesday**: Plotting using `ggplot2` + tips for next steps
- Additional tutorials available on website
  - R stuff: lists, loops, functions
  - Risk differences, risk ratios, odds ratios
  - Adding interaction terms in linear regression (i.e. effect modification)
  - Survival analysis

# Today’s agenda

- Logistic regression – conceptual tutorial
- Running and interpreting logistic regression in R

# **Review**: Risk difference, risk ratio, odds ratio

binary outcome

exposure

We run a randomized control trial testing the efficacy of a vaccine in preventing a disease. We obtain the following results:

- *In the vaccine arm*: 20/100 developed the disease
- *In the placebo arm*: 50/100 developed the disease

different

logistic regression

Calculate the following:

Risk of disease in the placebo arm:

$\frac{50}{100} = 0.5$

Risk of disease in the vaccine arm:

$\frac{20}{100} = 0.2$

Risk difference (vaccine vs placebo):

$0.2 - 0.5 = -0.3$

Risk ratio (vaccine vs placebo):

$\frac{0.2}{0.5} = 0.4$

Odds ratio (vaccine vs placebo):

$\frac{0.25}{1} = 0.25$

Odds of disease in the placebo arm:

$\frac{\text{Cases}}{\text{Non-cases}} = \frac{50}{50} = 1$

Odds of disease in the vaccine arm:

$\frac{20}{80} = 0.25$

# Logistic regression overview

- Linear regression – continuous outcomes (e.g. disease severity score)
  - Model:

$$\underbrace{SeverityScore}_{\text{outcome}} = \beta_0 + \beta_1 * \underbrace{TxDose}_{\text{exposure}}$$

  - Obtain estimates for coefficients $\beta_0, \beta_1$
  - $\beta_0$: expected score for those TxDose = 0, $\beta_1$: expected difference in score for groups w/ a 1-unit difference in dose
- Logistic regression – binary outcomes (e.g. disease incidence)
  - Model:

$$\log(OddsDisease) = \beta_0 + \beta_1 * TxDose$$

  - Obtain estimates for coefficients $\beta_0, \beta_1$
  - How can we interpret $\beta_0, \beta_1$?

# Logistic regression w/ binary variable (1)

outcome

- We are trying to determine if a hypertension is associated with exposure treatment status. We obtain data from an observational study with both variables.
- We can analyze this using logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Treatment status
- We do this in R and obtain the following output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2329     0.1503 -14.856   <2e-16 ***
tx            0.3099     0.2020   1.534    0.125
```

# Logistic regression w/ binary variable (2)

- Logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Treatment status

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2329     0.1503 -14.856   <2e-16 ***
tx            0.3099     0.2020   1.534    0.125
```

**Exponentiated coefficients:**

```
(Intercept)          tx
   0.107221    1.363275
```

$\log(\text{odds Htn}) = \beta_0 + \beta_1 \cdot \text{treatment}$

↳ $\text{Odds Htn} = \exp(\beta_0 + \beta_1 \cdot \text{treatment})$

$\beta_0 \Rightarrow \exp(\beta_0) = 0.107$

↳ expected odds of htn when untreated ($tx = 0$)

$\beta_1 \Rightarrow \exp(\beta_1) = 1.363$

↳ odds ratio comparing treated vs untreated people.

↳ treated people have an estimated 36% higher odds of hypertension

p-value: is the odds ratio significantly different than 1?
no, not significantly different

$$\text{Odds Ratio } (tx=1 \text{ vs } tx=0) = \frac{\text{Odds for } tx=1}{\text{Odds for } tx=0}$$

$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)}$$

$$= \exp(\beta_1)$$

# Logistic regression w/ continuous variable (1)

- We are trying to determine if a odds of hypertension is associated with age. We obtain data from an observational study with both variables.
- We answer this question using logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Age
- We do this in R and obtain the following output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.63319    0.59852  -9.412  < 2e-16 ***
age          0.06767    0.01073   6.305 2.87e-10 ***
```

- How can we interpret this output?

# Logistic regression w/ continuous variable (2)

- Logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Age

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.63319    0.59852  -9.412  < 2e-16 ***
age          0.06767    0.01073   6.305 2.87e-10 ***
```

**Exponentiated coefficients:**

```
(Intercept)         age
0.00357716  1.07001190
```

$\log(\text{odds Htn}) = \beta_0 + \beta_1 \cdot \text{Age}$

↳ $\text{odds Htn} = \exp(\beta_0 + \beta_1 \cdot \text{Age})$

$\beta_0$: ⇒ $\exp(\beta_0) = 0.0036$

Predicted odds of HTN when age = 0

$\beta_1$ ⇒ $\exp(\beta_1) = 1.07$

↳ Odds ratio comparing the odds of HTN between two groups age 1-year apart

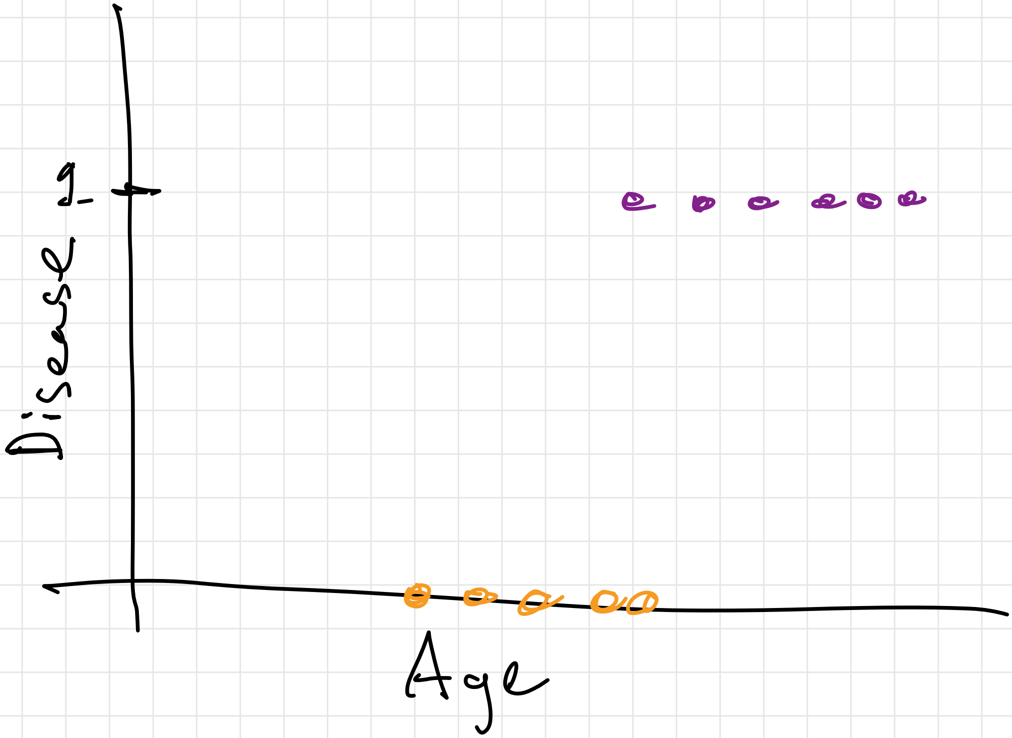
↳ on average the 1-year older group has 7% higher odds of having HTN

10-year

$\exp(\beta_1 \cdot 10)$

$\exp(0.0677 \cdot 10)$

p-value ⇒ find a significant association b/w HTN and age



1
Disease
Age

# Logistic regression w/ adjustment variable (1)

- We are trying to determine if a odds of hypertension is associated with treatment status adjusting for age. We obtain data from an observational study with all variables.
- We answer this question using logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Treatment status
  - **Adjustment covariate**: Age
- We do this in R and obtain the following output:

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.74480    0.60642  -9.473  < 2e-16 ***
tx           0.26707    0.20679   1.292    0.197    
age          0.06714    0.01074   6.252 4.06e-10 ***
```

# Logistic regression w/ adjustment variable (2)

- Logistic regression:
  - **Outcome**: Hypertension (binary)
  - **Exposure**: Treatment status
  - **Adjustment covariate**: Age

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -5.74480    0.60642  -9.473  < 2e-16 ***
tx           0.26707    0.20679   1.292    0.197    
age          0.06714    0.01074   6.252 4.06e-10 ***
```

**Exponentiated coefficients:**

```
(Intercept)          tx         age 
0.003199387 1.306137258 1.069444178 
```

# 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.