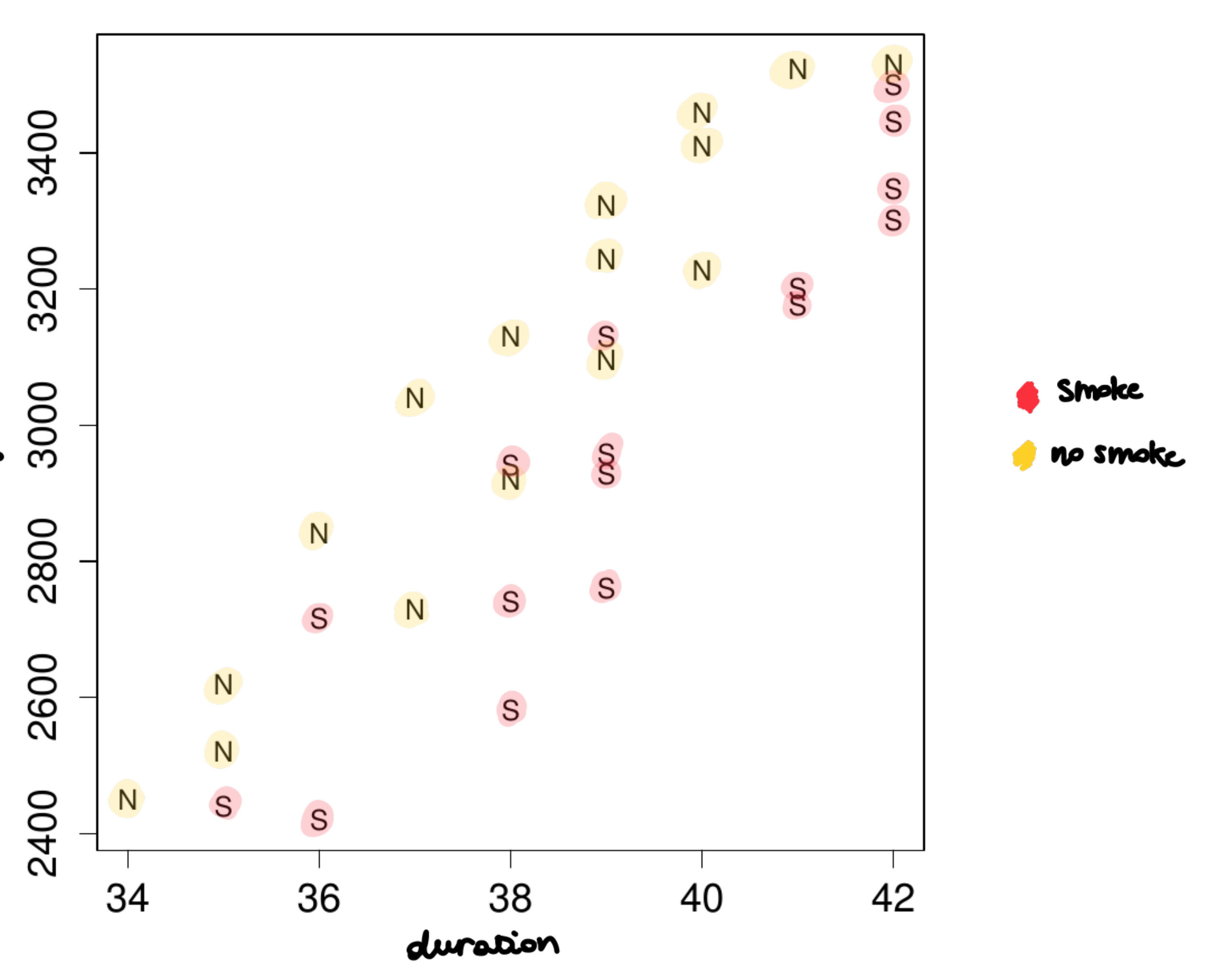
ANALYSIS OF COVARIANCE (ANCOVA)

• Assume that now we have two groups and a continuous covariate x

Example: y_i = weight of a baby at birth

x_i = duration of the pregnancy

group = smoke / no smoke (of the mother)



The interest is understanding WHETHER SMOKING AFFECTS THE WEIGHT AT BIRTH, WHILE CONTROLLING FOR THE PREGNANCY DURATION.

The weight is clearly influenced by the duration: premature babies have a lower weight compared to babies born later. This is true in general, regardless of the smoking factor.

Hence, it does not make sense to compare the weight of a child whose mother smokes with the weight of a child whose mother does not smoke, if the duration of the pregnancy is different. In that case it would not be clear if an observed difference in the weight is due to smoke or to the duration.

The individual effect of smoke is obtained only if we consider pregnancies with a similar duration.

We have two groups and a covariate.

Consider first modelling the two groups separately

SMOKE GROUP "S" : $Y_i^S = \beta_1^S + \beta_2^S x_i + \varepsilon_i \qquad i = 1, \dots, n_S$

NO-SMOKE GROUP "N" : $Y_i^N = \beta_1^N + \beta_2^N x_i + \varepsilon_i \qquad i = 1, \dots, n_N$

We ask whether the weight depends on the smoking habit, given the duration x

If we fix a duration x_0

$$\mu_0^S = \mathbb{E}[Y^S] = \beta_1^S + \beta_2^S x_0$$

$$\mu_0^N = \mathbb{E}[Y^N] = \beta_1^N + \beta_2^N x_0$$

Given the specific duration x_0, is there an effect of smoking?

This question corresponds to a test

$$\begin{cases} H_0: \mu_0^S = \mu_0^N \\ H_1: \mu_0^S \neq \mu_0^N \end{cases}$$

However, we are not interested in the effect of smoking only on babies born at the duration x_0. We want to study the effect of smoking, given the duration, for all durations.

We can do it using a unique linear model for the two groups.

We define, for $i = 1, \dots, n_S + n_N$

• x_i = duration i

• s_i : indicator of smoke $\quad s_i = \begin{cases} 1 & \text{if woman } i \text{ smokes} \\ 0 & \text{if woman } i \text{ does not smoke} \end{cases}$

• $x_i \cdot s_i$: interaction $\quad x_i \cdot s_i = \begin{cases} x_i & \text{if } s_i = 1 \\ 0 & \text{if } s_i = 0 \end{cases}$

and consider the following model:

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 s_i + \beta_4 x_i \cdot s_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

In matrix form we get:

$$X = [\, \underbrace{\underline{1}}\ \underbrace{\underline{x}}_{\text{duration}}\ \underbrace{\underline{s}}_{\substack{\text{dummy}\\\text{smoke}}}\ \underbrace{\underline{x} \cdot \underline{s}}_{\text{interaction}} \,]$$

$$\begin{bmatrix} 1 & x_1 & 1 & x_1 \\ 1 & x_2 & 1 & x_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_S} & 1 & x_{n_S} \\ 1 & x_{n_S+1} & 0 & 0 \\ 1 & x_{n_S+2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_S+n_N} & 0 & 0 \end{bmatrix} \begin{array}{l} \left.\vphantom{\begin{matrix}1\\1\\\vdots\\1\end{matrix}}\right\} \text{smoke group } i = 1, \dots, n_S \\ \left.\vphantom{\begin{matrix}1\\1\\\vdots\\1\end{matrix}}\right\} \text{no-smoke group } i = n_S+1, \dots, n_S+n_N \end{array}$$

Let's look at the mean of Y_i for different combinations of x_i and s_i

• if individual i smokes:

$$\mu_i = \beta_1 + \beta_2 x_i + \beta_3 \cdot 1 + \beta_4 \cdot s_i \cdot 1$$

$$= \underbrace{(\beta_1 + \beta_3)}_{\beta_1^S} + \underbrace{(\beta_2 + \beta_4)}_{\beta_2^S} x_i$$

• if individual i doesn't smoke

$$\mu_i = \beta_1 + \beta_2 x_i + \beta_3 \cdot 0 + \beta_4 \cdot s_i \cdot 0$$

$$= \underbrace{\beta_1}_{\beta_1^N} + \underbrace{\beta_2}_{\beta_2^N} x_i$$

INTERPRETATION OF THE PARAMETERS

• β_1 is the intercept in the "no-smoke" group

• $\beta_1 + \beta_3$ is the intercept in the "smoke" group

• β_2 is the effect on $\mathbb{E}[Y_i]$ of increasing the duration x_i by 1 unit in the "no-smoke" group

• $\beta_2 + \beta_4$ is the effect on $\mathbb{E}[Y_i]$ of increasing the duration x_i by 1 unit in the "smoke" group

We are interested in whether smoking has an effect on the weight, while controlling for the pregnancy duration.

If there is no effect, the two groups will have the same estimated regression line.

i.e., equal intercept and slope : $\beta_1^S = \beta_1^N$ and $\beta_2^S = \beta_2^N$

With the model for both groups it means:

$$\beta_1^N = \beta_1^S \Rightarrow \beta_1 = \beta_1 + \beta_3 \Rightarrow \beta_3 = 0$$

$$\beta_2^N = \beta_2^S \Rightarrow \beta_2 = \beta_2 + \beta_4 \Rightarrow \beta_4 = 0$$

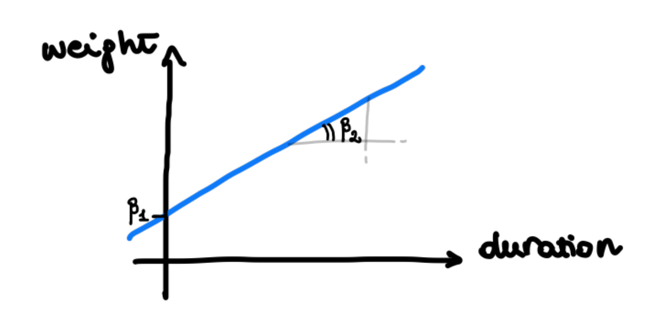
Hence, to test the absence of an effect of smoking on the weig, we test

$$\begin{cases} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \end{cases}$$

Test on whether smoking affects the weight at birth, controlling for the duration

• Under H_0: no effect of smoking

we have a single regression line for both groups

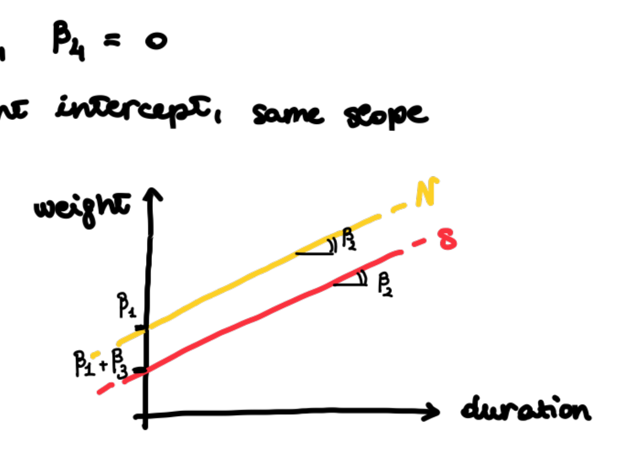


• Under H_1,

if I reject H_0, I can have different scenarios

1. $\beta_3 \neq 0$, $\beta_4 = 0$

different intercept, same slope



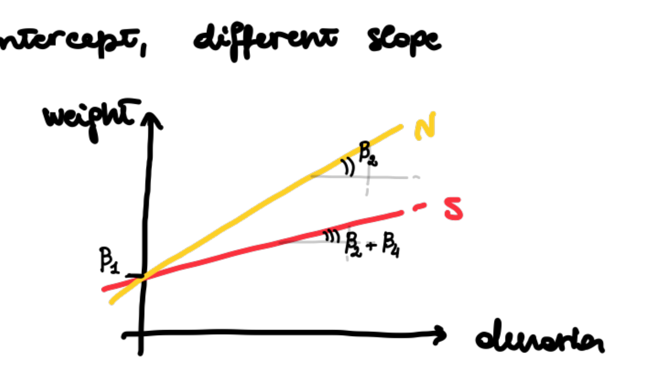
In this example, $\beta_3 < 0$.

The effect of smoking is constant, regardless of the duration.

→ SMOKING REDUCES THE EXPECTED WEIGHT BY β_3, FOR ALL DURATIONS.

2. $\beta_3 = 0$, $\beta_4 \neq 0$

same intercept, different slope

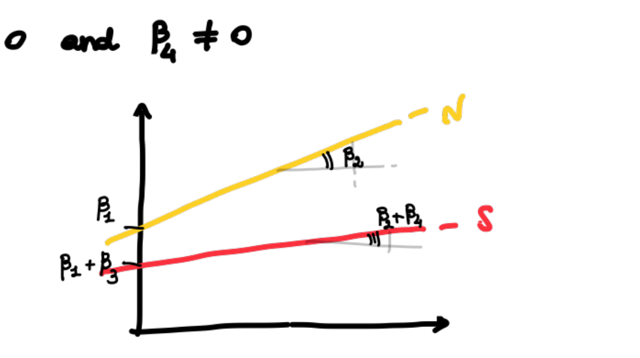


Here, $\beta_4 < 0$.

At a duration $x = 0$ (not meaningful here) smoking has no effect.

The effect increases for increasing duration.

3. $\beta_3 \neq 0$ and $\beta_4 \neq 0$



In the example, $\beta_3 < 0$ and $\beta_4 < 0$

At a duration $x = 0$ the two groups have different means.

Moreover, the effect of smoking changes for different durations.