- 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 if smoking affects the weight of the newborn, <u>while controlling for the pregnancy duration</u>. Indeed the weight is clearly influenced by the duration: premature babies have lower weight compared to babies born later (in general). So 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 effect of smoke is obtained only if we consider babies born after similar duration of the pregnancy ("for a given $x = x_0$").

Again, we are comparing 2 groups. However, now we also have a covariate x: we can specify a separate linear model for each group

smoke group: "S" $\quad Y_i^S = \beta_1^S + \beta_2^S x_i + \varepsilon_i \quad i = 1, \dots, n_S$

no-smoke group: "N" $\quad Y_i^N = \beta_1^N + \beta_2^N x_i + \varepsilon_i \quad i = 1, \dots, n_N$

the weight depends on the smoking habit, given the duration x

if we fix a duration x_0

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

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

given a specific duration, is there an effect of "smoke"? $H_0: \mu_0^S = \mu_0^N$

We can test this type of hyp. by writing the model in a single LM.

$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2) \qquad i = 1, \dots, n_S + n_N$

with x_{i2} = duration

x_{i3} = indicator of "smoke" (dummy) $= \begin{cases} 0 & \text{no smoke} \\ 1 & \text{smoke} \end{cases}$

$x_{i4} = x_{i2} \cdot x_{i3}$ = duration . smoke "interaction"

$= \begin{cases} x_{i2} = \text{duration} & \text{if smoke} = 1 \\ 0 & \text{if smoke} = 0 \end{cases}$

$$X = [\ \underline{1} \quad \underset{\uparrow}{\overset{\text{duration}}{\underline{x}_2}} \quad \underset{\text{dummy smoke}}{\underline{x}_3} \quad \underline{x}_2 \cdot \underline{x}_3\ ] = \left[\begin{array}{cccc} 1 & x_{12} & 1 & x_{12} \\ 1 & x_{22} & 1 & x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n_S 2} & 1 & x_{n_S 2} \\ 1 & x_{n_S+1,2} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & 0 & 0 \end{array}\right] \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \\ \end{array}\right\} \text{smoke group } i = 1, \dots, n_S \\ \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{no-smoke group } i = n_S+1, \dots, n_S+n_N \end{array}$$

Let's look at the mean of Y_i for different combinations of x_{i2}, x_{i3}, x_{i4}

- if individual i smokes: $\mu_i = \beta_1 + \beta_2 x_{i2} + \beta_3 \cdot 1 + \beta_4 \cdot (x_{i2} \cdot 1)$

$= \underbrace{(\beta_1 + \beta_3)}_{\beta_1^S} + \underbrace{(\beta_2 + \beta_4)}_{\beta_2^S} x_{i2}$

- if individual i doesn't smoke $\quad \mu_i = \underbrace{\beta_1}_{\beta_1^N} + \underbrace{\beta_2}_{\beta_2^N} x_{i2}$

$\Rightarrow$ 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 of x_{i2} on Y_i in the "no smoke" group

$\beta_2 + \beta_4$ is the effect of x_{i2} on Y_i 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.

ie: $\beta_1^S = \beta_1^N$ and $\beta_2^S = \beta_2^N$

With the new parameters $(\beta_1, \beta_2, \beta_3, \beta_4)$ it means:

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

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

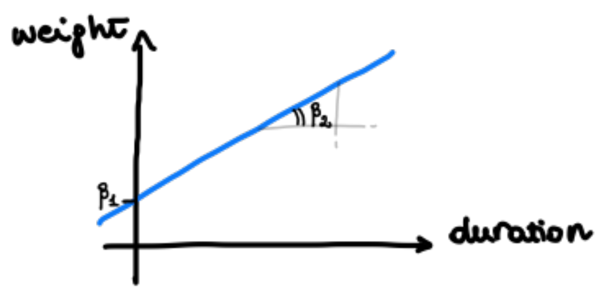
Hence we can write:

$$\left.\begin{array}{l} H_0: \beta_3 = \beta_4 = 0 \\ H_1: \overline{H_0} \quad (\text{at least one is} \neq 0) \end{array}\right\} \text{test on whether smoking affects the weight at birth, controlling for the duration}$$

↳ test about a subset of $\underline{\beta}$

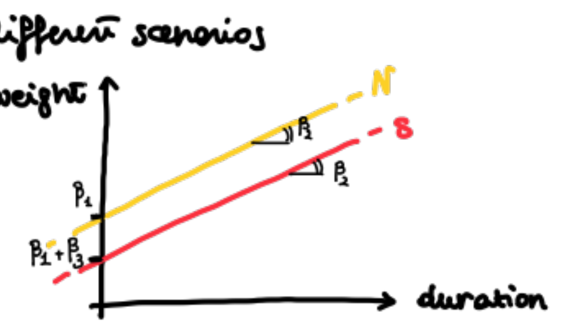
possible cases:

H_0: no effect: one regression line for both groups



if I reject H_0, I can have different scenarios

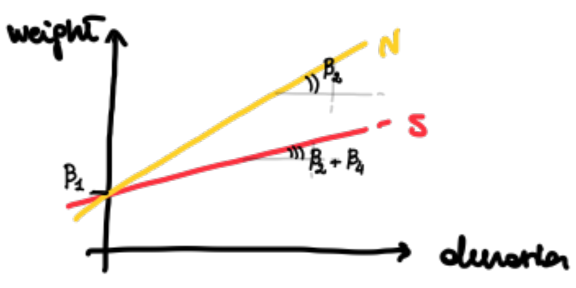
1) $\beta_3 \neq 0$, $\beta_4 = 0$
different intercept, same slope



$\beta_3 < 0$ here

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

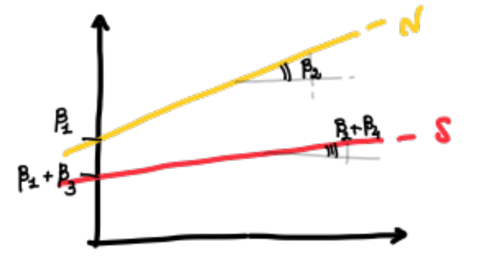
2) $\beta_3 = 0$, $\beta_4 \neq 0$
same intercept
different slope



$\beta_4 < 0$ here

At duration = 0 (not meaningful here...) smoking has no effect. The effect increases for increasing duration.

3) $\beta_3 \neq 0$ $\beta_4 \neq 0$
different slope and intercept



At duration = 0 the two groups have different means. Moreover, there is an effect also on the slope.

with the data, how do I do the test?

Fit the restricted model (H_0) $\quad Y_i = \beta_1 + \beta_2 x_{i2} + \varepsilon_i \quad \rightsquigarrow$ I obtain $\tilde{\beta}_1, \tilde{\beta}_2 \Rightarrow \tilde{y}_i = \tilde{\beta}_1 + \tilde{\beta}_2 x_{i2} \quad p_0 = 2$ covariates

compute the estimated variance $\tilde{\sigma}^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \cdot \frac{1}{n}$

Fit the unconstrained model (H_1) $\quad Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} \quad \rightsquigarrow$ estimates $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4$

$\Rightarrow \hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3} + \hat{\beta}_4 x_{i4} \qquad p = 4$ covariates

estimated variance $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$F = \frac{\dfrac{\tilde{\sigma}^2(Y) - \hat{\sigma}^2(Y)}{4 - 2}}{\dfrac{\hat{\sigma}^2(Y)}{n - 4}} \overset{H_0}{\sim} F_{2, n-4}$$

$$f^{obs} = \frac{(\tilde{\sigma}^2 - \hat{\sigma}^2)/2}{\hat{\sigma}^2 / n-4}$$

$$\alpha^{obs} = \mathbb{P}_{H_0}(F \geq f^{obs})$$