Multiple Linear regression

STATS – Modelação Estatística

PhD Programme in Health Data Science

Cristina Costa Santos & Andreia Teixeira

Multiple Linear Regression

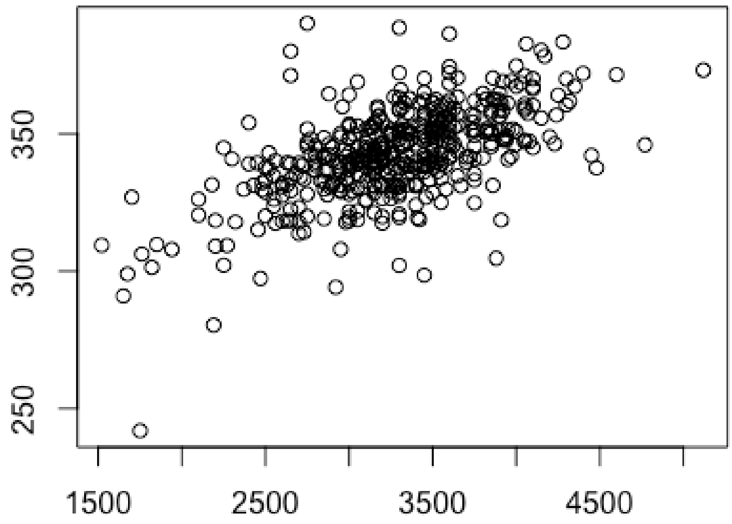
Multiple linear regression is an extension of the simple linear regression model to several covariates

$$\boxed{y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)}$$

$$\boxed{\mu_{y_i|x_{1i},x_{2i},\ldots,x_{ki}} = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}}$$

Multiple Linear Regression

Is the newborn weight related to head circumference?




```
> regressao <- lm(alcohol$ofc~alcohol$birthwt, data=alcohol)
> summary(regressao)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt, data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.016  -7.760   1.330   7.688  57.259 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.805e+02  4.015e+00   69.88   <2e-16 ***
alcohol$birthwt 1.905e-02  1.209e-03   15.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.54 on 452 degrees of freedom
Multiple R-squared:  0.3546,	Adjusted R-squared:  0.3532 
F-statistic: 248.4 on 1 and 452 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

Is the newborn weight related to head circumference?

```
> regressao <- lm(alcohol$ofc~alcohol$birthwt, data=alcohol)
> summary(regressao)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt, data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.016  -7.760   1.330   7.688  57.259 

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.805e+02  4.015e+00   69.88   <2e-16 ***
alcohol$birthwt 1.905e-02  1.209e-03   15.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.54 on 452 degrees of freedom
Multiple R-squared:  0.3546,    Adjusted R-squared:  0.3532 
F-statistic: 248.4 on 1 and 452 DF,  p-value: < 2.2e-16
```

Multiple Linear Regression

Is the newborn weight related to head circumference?

```
> regressao <- lm(alcohol$ofc~alcohol$birthwt, data=alcohol)
> summary(regressao)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt, data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.016  -7.760   1.330   7.688  57.259 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.805e+02  4.015e+00   69.88   <2e-16 ***
alcohol$birthwt 1.905e-02  1.209e-03   15.76   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.54 on 452 degrees of freedom
Multiple R-squared:  0.3546,	Adjusted R-squared:  0.3532 
F-statistic: 248.4 on 1 and 452 DF,  p-value: < 2.2e-16
```

Head circumference increases by an average of 0.019 mm for each gram of increase in baby weight.

Multiple Linear Regression

Is the newborn weight related to head circumference?

Head circumference increases by an average of 0.019 mm for each gram of increase in baby weight.

And if we adjust for gestational age, newborn weight mantain the same relation with head circumference?

Multiple Linear Regression

And if we adjust for gestational age, newborn weight mantain the same relation with head circumference?

```
> regressao2 <- lm(alcohol$ofc~alcohol$birthwt + alcohol$gestlmp, data=alcohol)
> summary(regressao2)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt + alcohol$gestlmp,
    data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.079  -8.251   1.346   7.632  54.835 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.484e+02  1.486e+01  16.716   <2e-16 ***
alcohol$birthwt 1.767e-02  1.367e-03  12.922   <2e-16 ***
alcohol$gestlmp 9.374e-01  4.169e-01   2.248    0.025 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.53 on 447 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.3591,    Adjusted R-squared:  0.3563 
F-statistic: 125.3 on 2 and 447 DF,  p-value: < 2.2e-16
```

The effect of both variables becomes significant for a level of significance of 0.05.

Multiple Linear Regression

```
> regressao2 <- lm(alcohol$ofc~alcohol$birthwt + alcohol$gestlmp, data=alcohol)
> summary(regressao2)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt + alcohol$gestlmp,
    data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-72.079  -8.251   1.346   7.632  54.835 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     2.484e+02  1.486e+01  16.716   <2e-16 ***
alcohol$birthwt 1.767e-02  1.367e-03  12.922   <2e-16 ***
alcohol$gestlmp 9.374e-01  4.169e-01   2.248    0.025 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.53 on 447 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared:  0.3591,	Adjusted R-squared:  0.3563 
F-statistic: 125.3 on 2 and 447 DF,  p-value: < 2.2e-16
```

For babies of the same gestational age (fixing the gestational age) the head circumference increases on average 0.018 mm for each gram of baby weight gain.

Multiple Linear Regression

As in the simple linear regression model we can ask how much variation of y (here head circumference) can be explained by the model (now with 2 variables).

The answer is 36%.

$$\frac{44928 + 925}{44928 + 925 + 81821} = \frac{45853}{127674} = 0,359$$

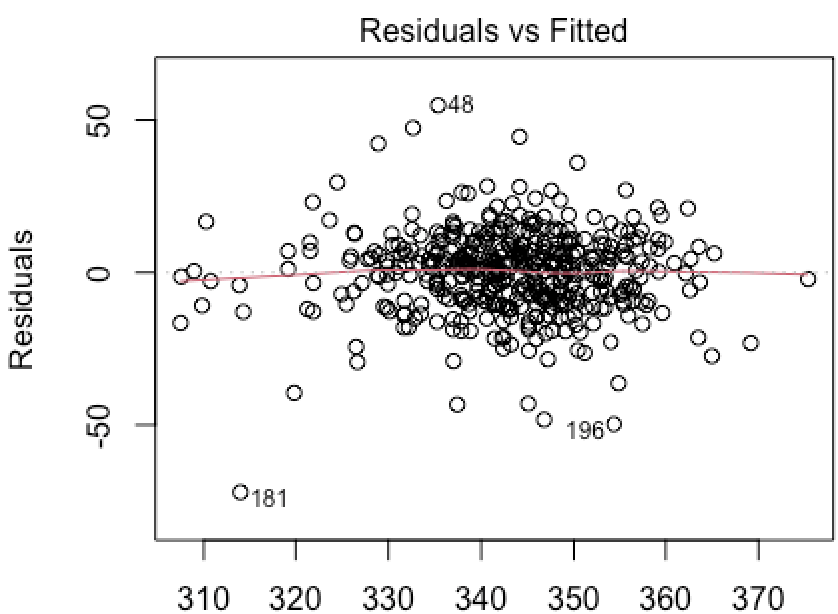
```
> anova(regressao2)
Analysis of Variance Table

Response: alcohol$ofc
                 Df Sum Sq Mean Sq  F value   Pr(>F)
alcohol$birthwt   1  44928   44928 245.4500 < 2e-16 ***
alcohol$gestlmp   1    925     925   5.0556 0.02503 *
Residuals       447  81821     183
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

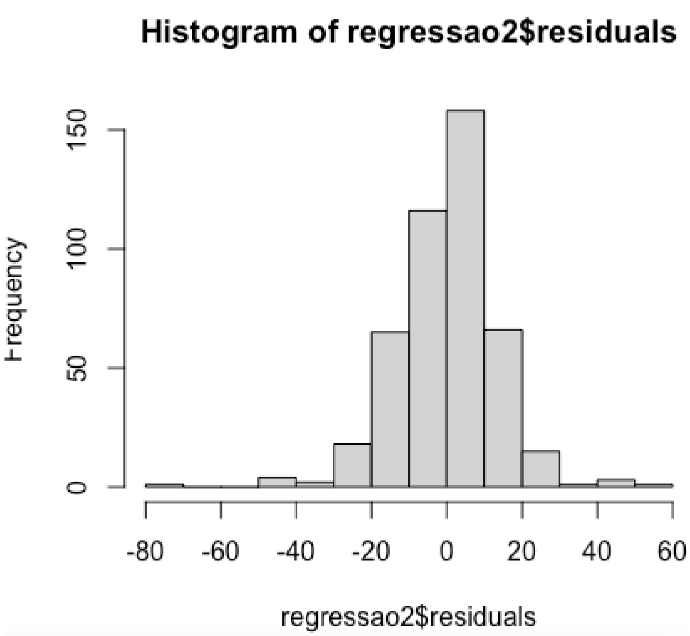
Multiple Linear Regression

To check the assumptions of the model we can analyze the residuals as we did in the regression simple linear.

```
> plot(regressao2, 1)
```




```
> hist(regressao2$residuals)
```



Multiple Linear Regression

And if we consider the same example but with a categorical covariate:

Sex (0=female; 1=male)

$$\boxed{HC_i = \beta_0 + \beta_1 * bweight_i + \beta_2 * sex_i + \epsilon_i}$$

For girls:

$$\boxed{HC_i = \beta_0 + \beta_1 * bweight_i + \epsilon_i}$$

For boys:

$$\boxed{HC_i = (\beta_0 + \beta_2) + \beta_1 * bweight_i + \epsilon_i}$$

Multiple Linear Regression

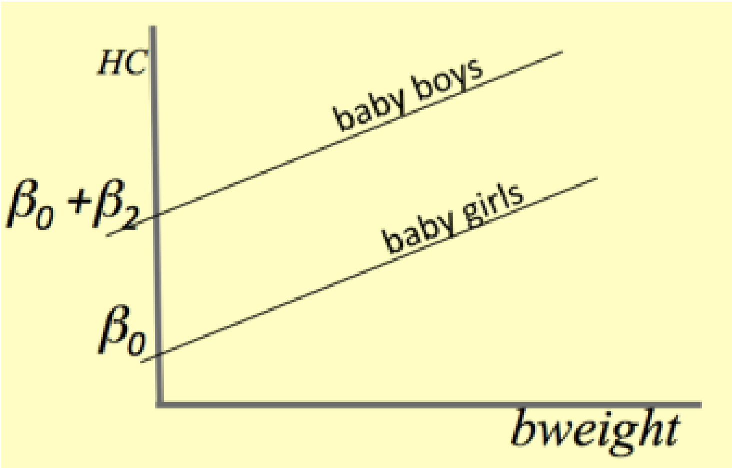
For girls:

$$HC_i = \beta_0 + \beta_1 * bweight_i + \epsilon_i$$

For boys:

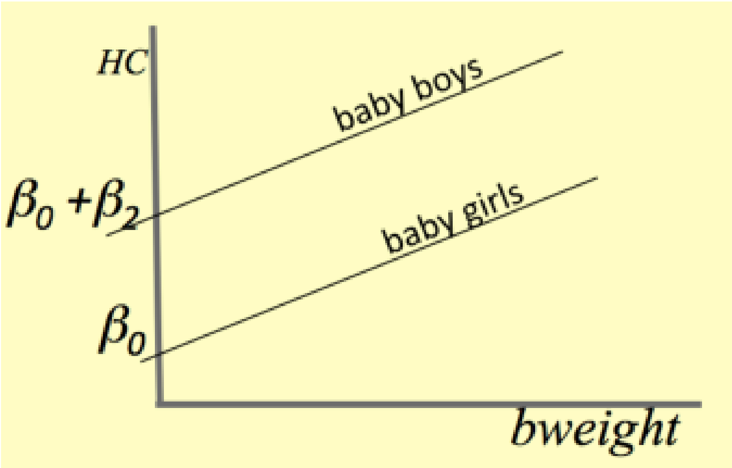
$$HC_i = (\beta_0 + \beta_2) + \beta_1 * bweight_i + \epsilon_i$$

Each equation corresponds to a line with the same slope but with different intercepts



Multiple Linear Regression

Girls have, on average, a constant head circumference difference to boys, but the effect of baby weight on head circumference is the same for both sexes.

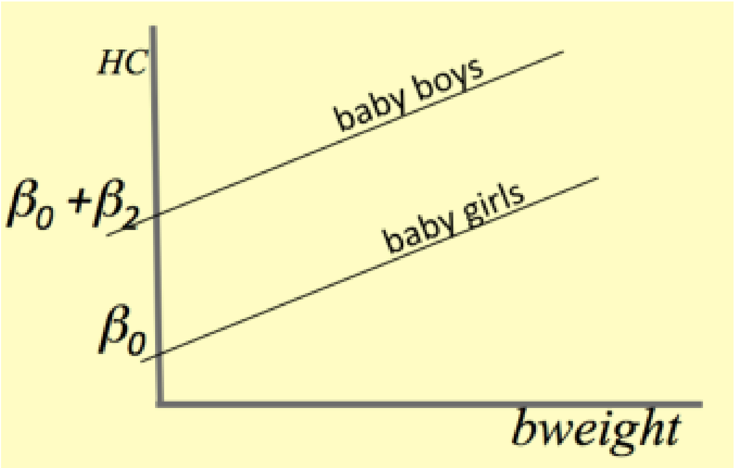


Multiple Linear Regression

Girls have, on average, a constant head circumference difference to boys, but the effect of baby weight on head circumference is the same for both sexes.

BUT

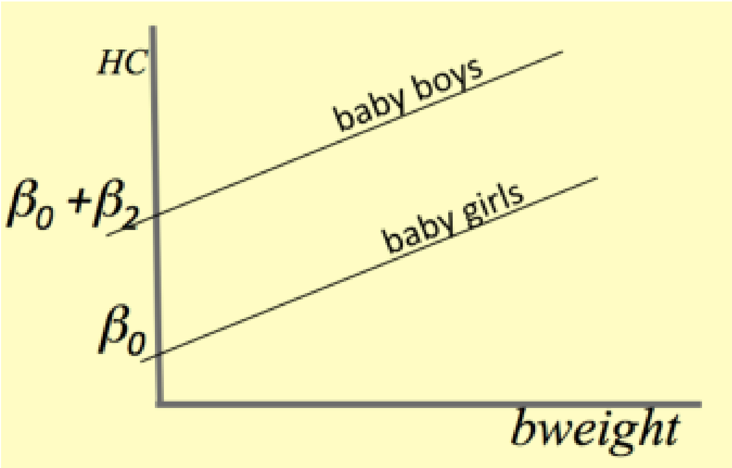
we forced it when
we chose this model



Multiple Linear Regression

Girls have, on average, a constant head circumference difference to boys, but the effect of baby weight on head circumference is the same for both sexes.

But, in fact, the effect of the baby's weight on the head circumference may be different in each sex, but the model we choose doesn't allows this difference.



Multiple linear regression

```
> regressao3 <- lm(alcohol$ofc~alcohol$birthwt + alcohol$sex, data=alcohol)
> summary(regressao3)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt + alcohol$sex, data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max 
-73.213  -7.823   0.807   8.069  56.204 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     279.86750    4.02765  69.487   <2e-16 ***
alcohol$birthwt   0.01891    0.00121  15.635   <2e-16 ***
alcohol$sex       2.11979    1.27486   1.663   0.0971 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.51 on 451 degrees of freedom
Multiple R-squared:  0.3586,	Adjusted R-squared:  0.3557 
F-statistic: 126.1 on 2 and 451 DF,  p-value: < 2.2e-16
```

Gender is not significant, so we have no evidence of differences between sex in the head circumference, even adjusting for the baby's weight.

Multiple linear regression

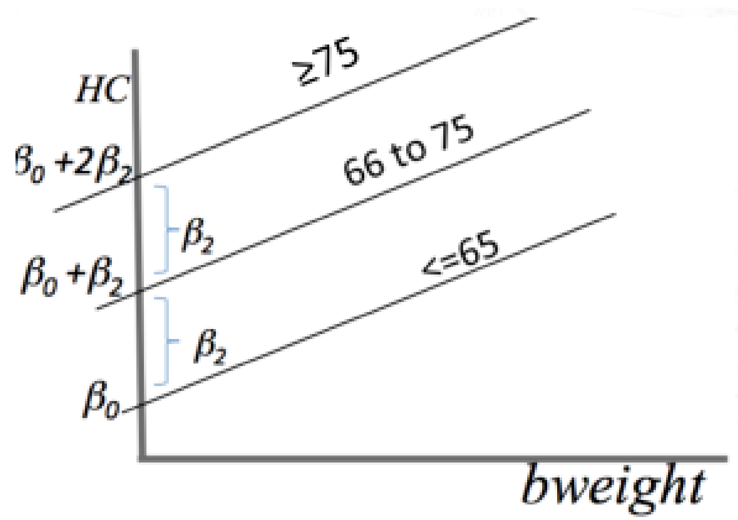
If the categorical variable has more than two categories there are two possible approaches.

Mothers' weight can be a categorical variable:

0: ≤ 65 kg
1: between 65 and 75 kg
2: ≥ 75 kg

$$HC_i = \beta_0 + \beta_1 * bweight_i + \beta_2 * mweight_i + \epsilon_i$$

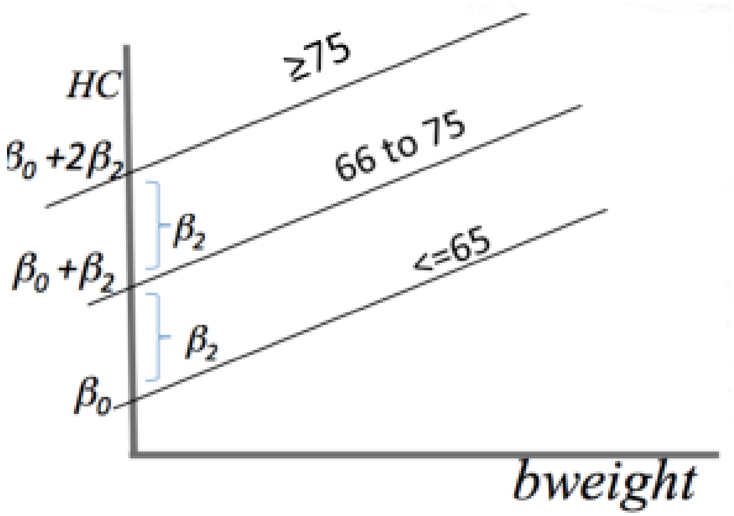
$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 0**
$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 1**
$HC_i = (\beta_0 + 2\beta_2) + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 2**



Multiple linear regression

In addition to the lines being parallel (forced by the model) it is also imposed that the difference between the 3rd and 2nd groups is the same as the difference between the 1st and 2nd groups.

$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 0**
$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 1**
$HC_i = (\beta_0 + 2\beta_2) + \beta_1 bweight_i + \varepsilon_i$	**For mweight = 2**



Multiple linear regression

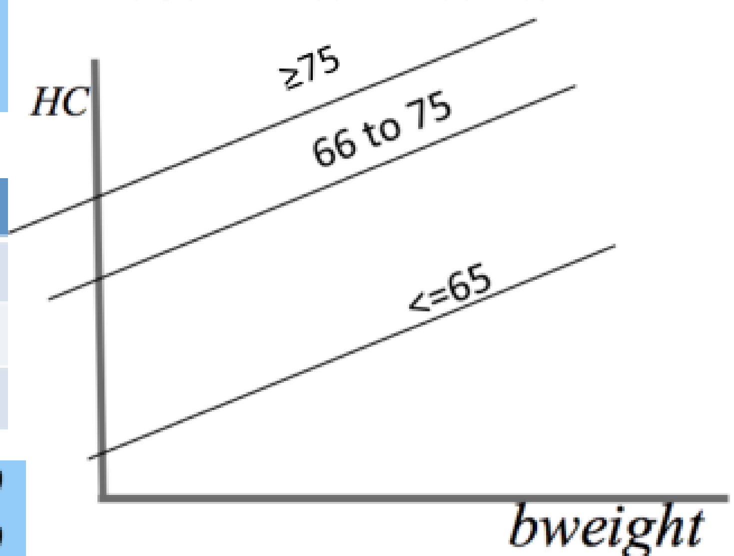
We can allow different differences between the groups.

For that we have to create indicator variables (dummy).

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 I_{1i} + \beta_3 I_{2i} + \varepsilon_i$$

$$I_{1i} = \begin{cases} 1 \text{ if mweight}_i = 1 \\ 0 \;\text{ otherwise} \end{cases} \qquad I_{2i} = \begin{cases} 1 \text{ if mweight}_i = 2 \\ 0 \;\text{ otherwise} \end{cases}$$

Mweight	I_1	I_2
0	0	0
1	1	0
2	0	1



$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$ **For mweight = 0, i.e,** $I_1=0,\ I_2=0$

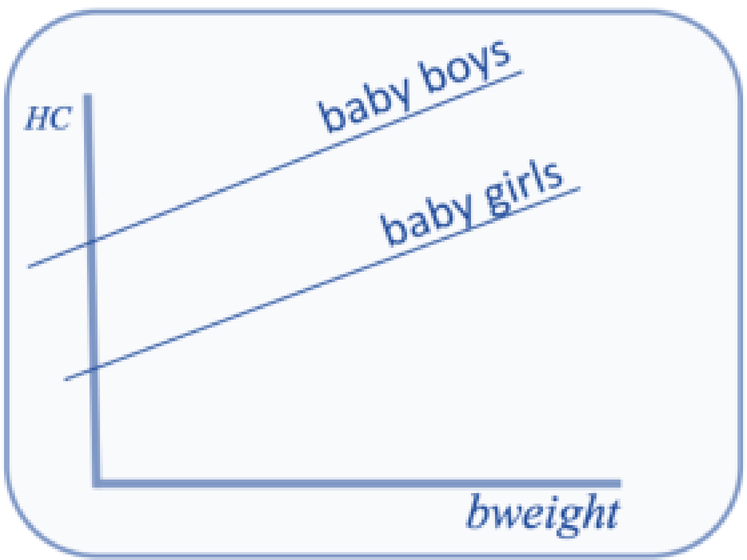
$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$ **For mweight = 1, i.e,** $I_1=1,\ I_2=0$

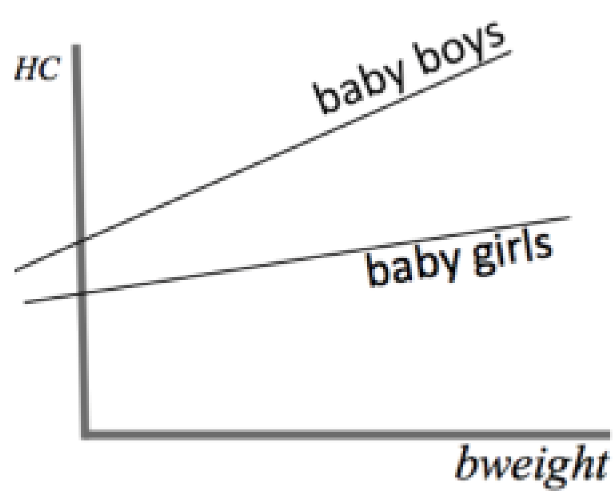
$HC_i = (\beta_0 + \beta_3) + \beta_1 bweight_i + \varepsilon_i$ **For mweight = 2, i.e,** $I_1=0,\ I_2=1$

Interactions

$$HC_i = \beta_0 + \beta_1 * bweight_i + \beta_2 * sex_i + \epsilon_i$$

assumes parallel lines in both sexes



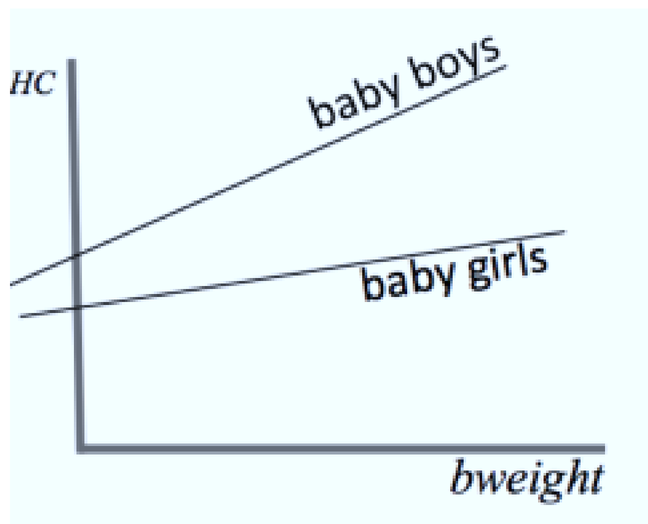


Interactions

The lines being parallel means that the effect of the weight in head circumference is the same in both sexes.

The increase of 1 gr of weight leads to the same increases in head circumference in boys and in the girls.

If the effect of newborn weight is different for each sex we say that there is a **interaction** between weight and gender.



Interactions

That is, we will have two straight lines with diferente intercepts and slopes.

$$\boxed{HC_i = \beta_0 + \beta_1 * bweight_i + \beta_2 * sex_i + \beta_3 * bweight_i * sex_i + \epsilon_i}$$

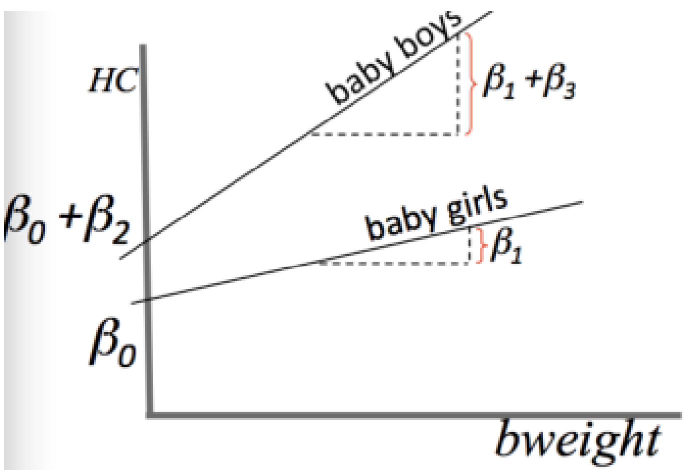
$$\boxed{HC_i = \beta_0 + \beta_1 * bweight_i + \epsilon_i}$$

$$\boxed{HC_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) * bweight_i + \epsilon_i}$$

Interactions

$$HC_i = \beta_0 + \beta_1 * bweight_i + \epsilon_i$$

$$HC_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) * bweight_i + \epsilon_i$$



Interactions

```
> regressao3 <- lm(alcohol$ofc~alcohol$birthwt + alcohol$sex + alcohol$sex*alcohol$birthwt, data=alcohol)
> summary(regressao3)

Call:
lm(formula = alcohol$ofc ~ alcohol$birthwt + alcohol$sex + alcohol$sex *
    alcohol$birthwt, data = alcohol)

Residuals:
    Min      1Q  Median      3Q     Max
-71.704  -7.749   1.017   8.138  56.747

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                284.327309   6.175274  46.043   <2e-16 ***
alcohol$birthwt              0.017535   0.001884   9.305   <2e-16 ***
alcohol$sex                 -5.537319   8.136821  -0.681    0.497
alcohol$birthwt:alcohol$sex  0.002342   0.002458   0.953    0.341
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.51 on 450 degrees of freedom
Multiple R-squared:  0.3599,	Adjusted R-squared:  0.3556
F-statistic: 84.33 on 3 and 450 DF,  p-value: < 2.2e-16
```

The effect of weight on head circumference is 0.018 for girls (sex=0) and 0.018+0.002 for the boys (sex=1). However, this interaction it is not significant(p=0.341).

When we have an interaction in the model, the main effects test no longer has big importance.

If an interaction or covariable effect is non-significant, should we remove it from the model?

It depends...

How do we decide what variables to include or exclude from the model?

Common sense might be the right way to go