Multiple Linear Regression

Basics of Health Intelligent Data Analysis

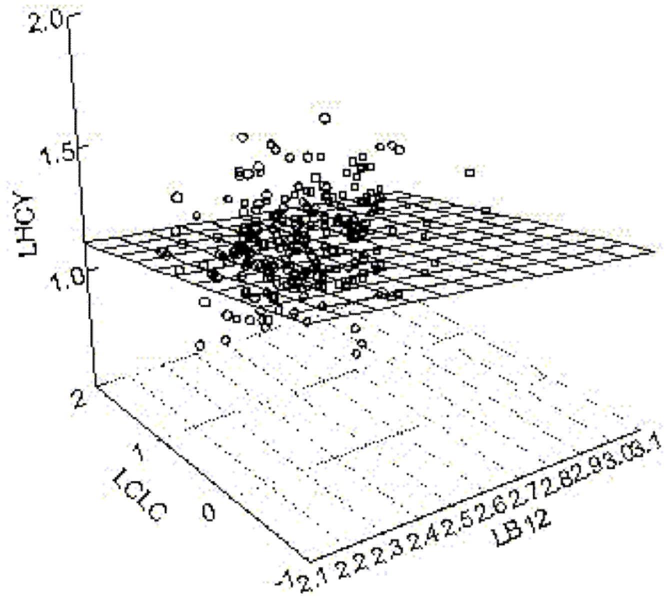
PhD Programme in Health Data Science

Cláudia Camila Dias

Pedro Pereira Rodrigues

Multiple Linear Regression

- The multiple linear regression is the natural extension of the simple linear model for several covariates



$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i$$

$$\varepsilon_i \sim N(0, \sigma^2)$$

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

- Geometrically this corresponds to fit a hyperplane to the data

Multiple Linear Regression

- In the example, it is not surprising that the head circumference may also be related with birthweight
- First lets look at the simple linear regression with birthweight as a covariate

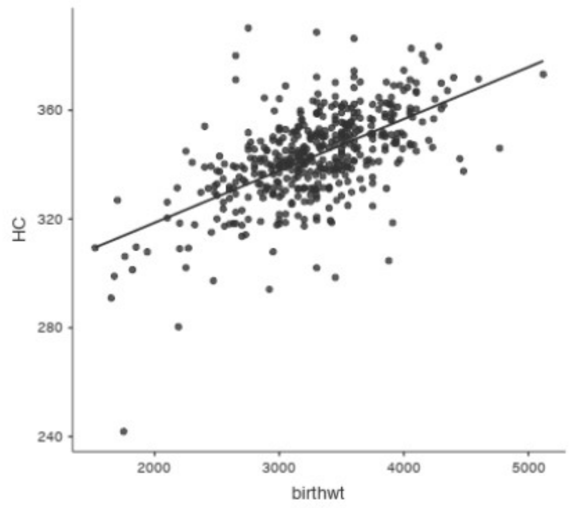
Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
birthwt	45515	1	45515	248	<.001
Residuals	82825	452	183		

Note. Type 3 sum of squares

Model Coefficients - HC

			95% Confidence Interval			
Predictor	Estimate	SE	Lower	Upper	t	p
Intercept	280.5438	4.01487	272.6537	288.4339	69.9	<.001
birthwt	0.0191	0.00121	0.0167	0.0214	15.8	<.001

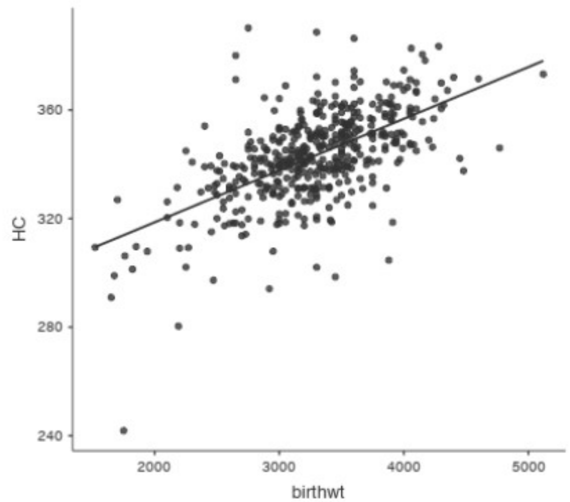


- Birthweigth is (linearly) associated with the head circumference

Multiple Linear Regression

- In the example, it is not surprising that the head circumference may also be related with birthweight
- First lets look at the simple linear regression with birthweight as a covariate

```
> lm(ofc ~ gestlmp,data=alcohol)

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

Coefficients:
(Intercept)      gestlmp
    210.042        3.392

>

> anova(lm(ofc ~ gestlmp,data=alcohol))
Analysis of Variance Table

Response: ofc
           Df Sum Sq Mean Sq F value    Pr(>F)
gestlmp     1  15291 15290.8  60.955 4.171e-14 ***
Residuals 448 112383   250.9
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



- Birthweigth is (linearly) associated with the head circumference

Multiple Linear Regression

- What about taking gestational age also into accout? i.e, if we adust for gestational age, is birthweight still associated with the head circumference?
- Fitting a model with both covariates we get:

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	248.3505	14.85690	16.72	<.001
gestimp	0.9374	0.41692	2.25	0.025
birthwt	0.0177	0.00137	12.92	<.001

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$

The effect of the two covariates remains significant at the 0.05 level

Multiple Linear Regression

- What about taking gestational age also into accout? i.e, if we adust for gestational age, is birthweight still associated with the head circumference?
- Fitting a model with both covariates we get:

```
> summary(lm(ofc ~ gestlmp + birthwt,data=alcohol))

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

Residuals:
<Labelled double>
    Min      1Q  Median      3Q     Max 
-72.079  -8.251   1.346   7.632  54.835 

Labels:
 value   label
   999 Missing

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.484e+02  1.486e+01  16.716   <2e-16 ***
gestlmp     9.374e-01  4.169e-01   2.248    0.025 *  
birthwt     1.767e-02  1.367e-03  12.922   <2e-16 ***
---
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
```

$\hat{\beta}_0$ $\hat{\beta}_1$ $\hat{\beta}_2$

The effect of the two covariates remains significant at the 0.05 level

Multiple Linear Regression

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	248.3505	14.85690	16.72	< .001
gestimp	0.9374	0.41692	2.25	0.025
birthwt	0.0177	0.00137	12.92	< .001

```
> summary(lm(ofc ~ gestlmp + birthwt,data=alcohol))

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

Residuals:
<Labelled double>
    Min      1Q  Median      3Q     Max
-72.079  -8.251   1.346   7.632  54.835

Labels:
 value   label
   999 Missing

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.484e+02  1.486e+01  16.716   <2e-16 ***
gestlmp     9.374e-01  4.169e-01   2.248    0.025 *
birthwt     1.767e-02  1.367e-03  12.922   <2e-16 ***
---
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
```

Now, the regression coefficients, β_1 and β_2, have a different interpretation

- **For babies of the same gestational age** (fixing the gestational age), the head circumference increases on average **0.018 mm for an increase of one gram** in the birthweigh
- **For babies of the the same birthweight,** the head circumference increases on average **0.937 mm per week** of gestational age

Multiple Linear Regression

Note for example the change on the coefficient for gestational age:

Simple linear regression

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	210.04	17.043	12.32	<.001
gestimp	$\boxed{3.39}$	0.434	7.81	<.001

Multiple linear regression

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	248.3505	14.85690	16.72	<.001
gestimp	$\boxed{0.9374}$	0.41692	2.25	0.025
birthwt	0.0177	0.00137	12.92	<.001

Model Fit

As before, we can ask how much variation of y can be explained by the model (with the two covariates)

Simple linear regression

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestimp	15291	1	15291	61.0	<.001
Residuals	112383	448	251		

Note. Type 3 sum of squares

Multiple linear regression

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestimp	925	1	925	5.06	0.025
birthwt	30563	1	30563	166.97	<.001
Residuals	81821	447	183		

Note. Type 3 sum of squares

- Note that the total in both tables is the same.
- However, the model with more covariates explains more variation

Model Fit

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
gestimp	925	1	925	5.06	0.025
birthwt	30563	1	30563	166.97	<.001
Residuals	81821	447	183		

Note. Type 3 sum of squares

- The test above shows that the amount of variation explained by the model is significantly higher than zero
- This is equivalent to test H_0: $\beta_1=\beta_2=0$
- A good practice is to look first at this test and only then look at the individual tests H_0: $\beta_k=0$

Model Fit

Model Fit Measures

Model	R	R^2
1	0.599	0.359

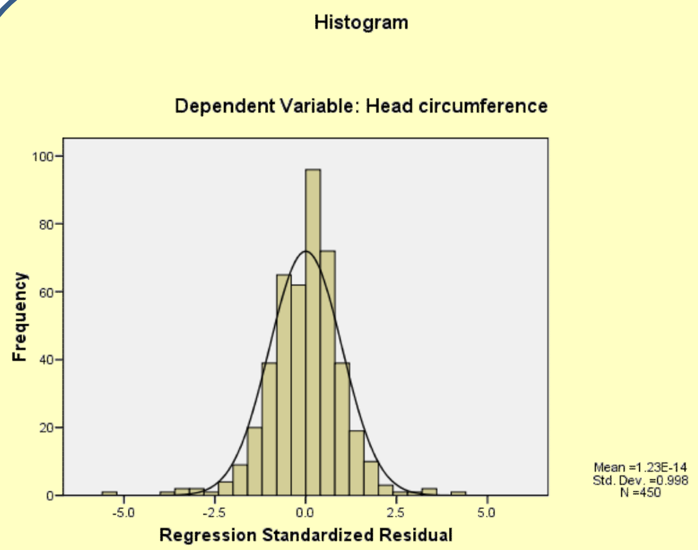
- However, the inclusion of an additional variable will allways increase the *r^2* so we should be carefull interpreting this statistics

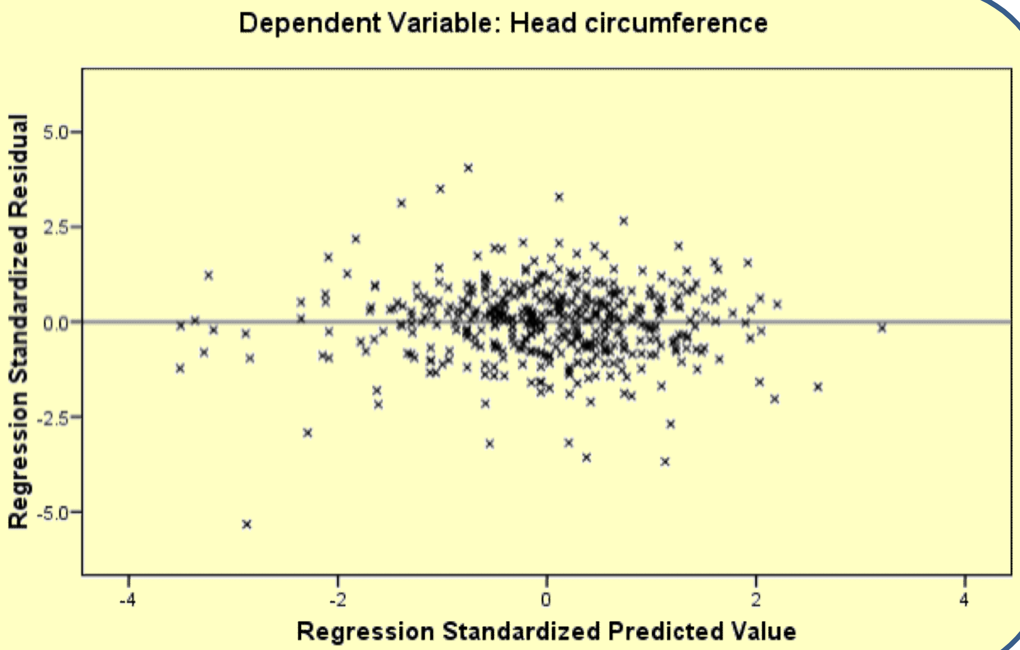
- An alternative measure that is often reported is the **adjusted *r^2***

- The **adjusted *r^2*** corrects the ***r^2*** for the number of covariates in the model (model complexity)
- More covariates => smaller ***r^2***

Assumptions

- To check the linear model assumptions we can use the same tools as before – the analysis of the residuals





- The assumptions seem to hold

Categorical covariates

- The categorical variables play a "special" role in regression.
- Lets consider the same example but with the covariates birthweight and baby sex (categorical, binary variable)
- Sex is coded as 0 – female 1 – male
- The model is written as:

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 sex_i + \varepsilon_i$$

Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 sex_i + \varepsilon_i$$

- For **baby girls**, sex = 0, so the model becomes

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$

- For **baby boys**, sex = 1, so the model becomes

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$

Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

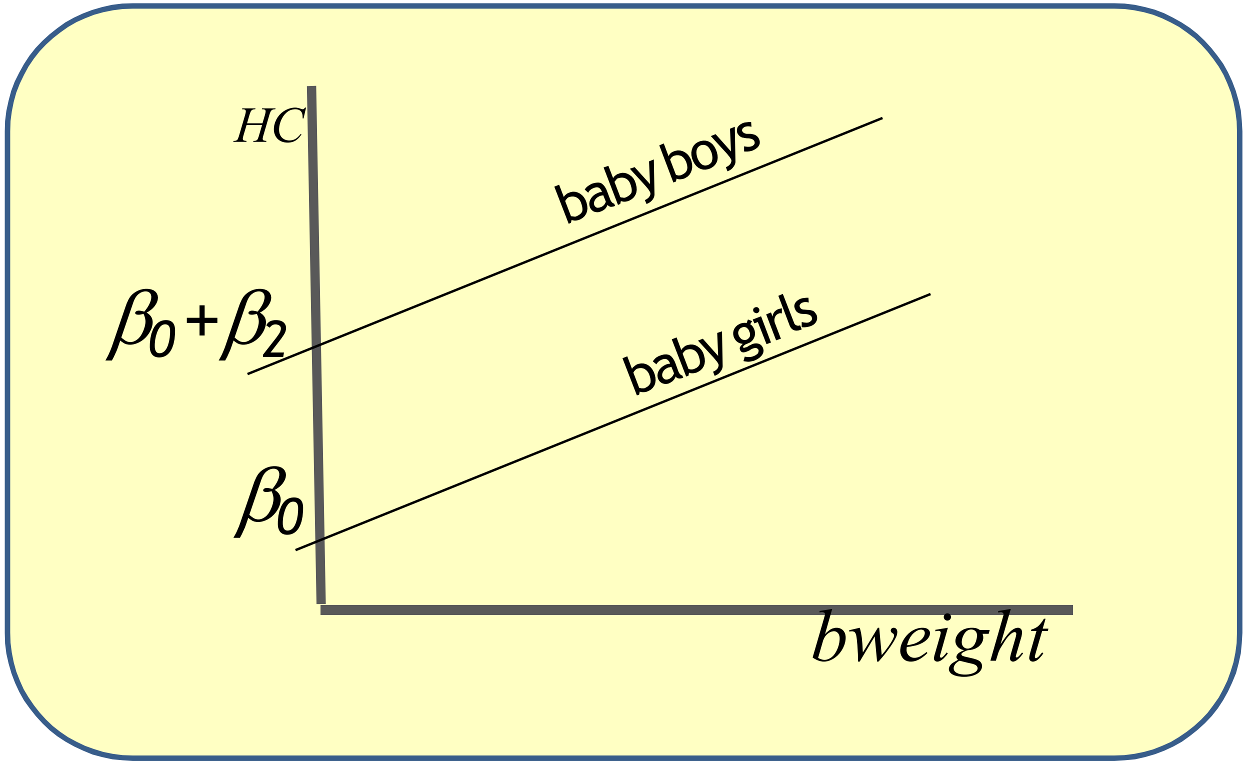
$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **baby boys**

- Each equation corresponds to a straigh line with the same slope (β_1) but different intercept (β_0 for girls *and* $\beta_0 + \beta_2$ for boys)

Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **baby boys**

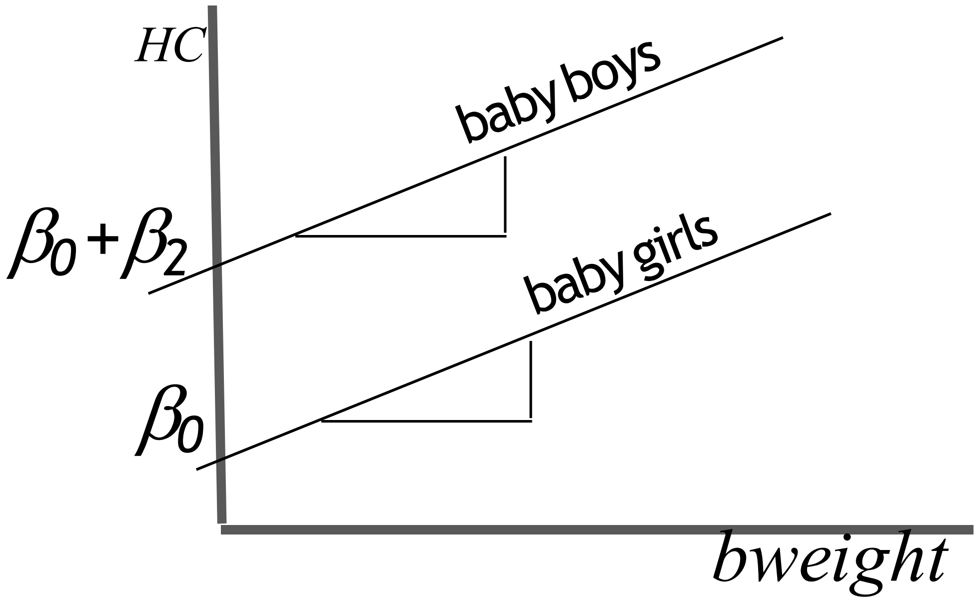


Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **baby boys**

Girls have in average a constant difference in HC but the effect of birthweigth on HC is the same for both sexes



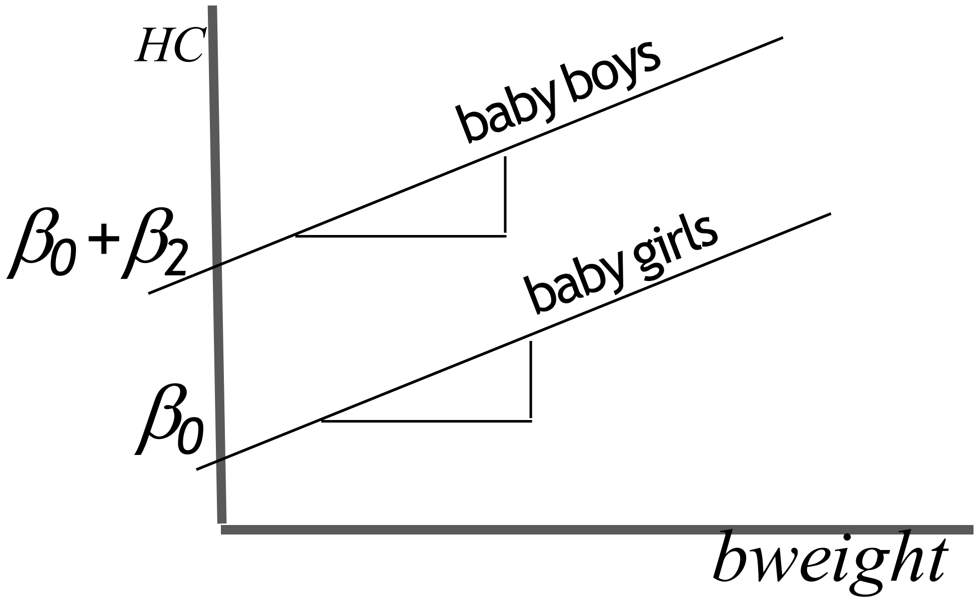
Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **baby boys**

Notice that we are imposing this by chosing this model

In fact, the effect of birthweight on HC could be different for each sex



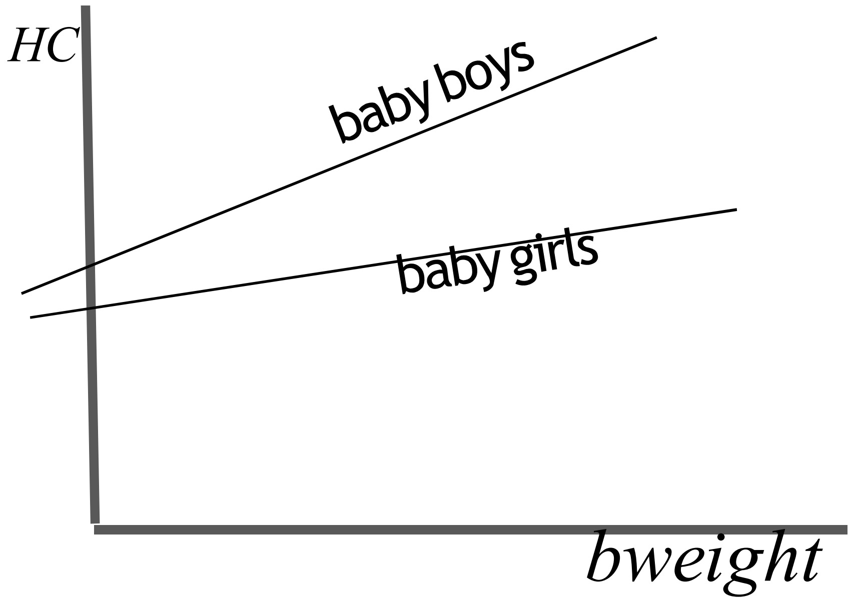
Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **baby boys**

But the model we chose **does not allow** different slopes!

We will come back to this point later.



Categorical covariates

- The result for the previous model is:

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
birthwt	44622	1	44622	244.46	<.001
sex	505	1	505	2.76	0.097
Residuals	82321	451	183		

Note. Type 3 sum of squares

[3]

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	280.9274	4.01369	69.99	<.001
birthwt	0.0189	0.00121	15.64	<.001
sex:				
Male – Female	2.1198	1.27486	1.66	0.097

- Sex is non-significant, so there is no evidence that supports differences between the genders regarding HC, after adjusting for birthweight

Categorical covariates

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

- Consider the model for HC using gestational age and mother weight at admission as covariates

- Mother weight at admission is a categorical variable coded as following:

  0: <=65

  1: 66 to 75

  2: >76

Categorical covariates

- We may use themodel

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 mweight_i + \varepsilon_i$$

- This defines three straightlines

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

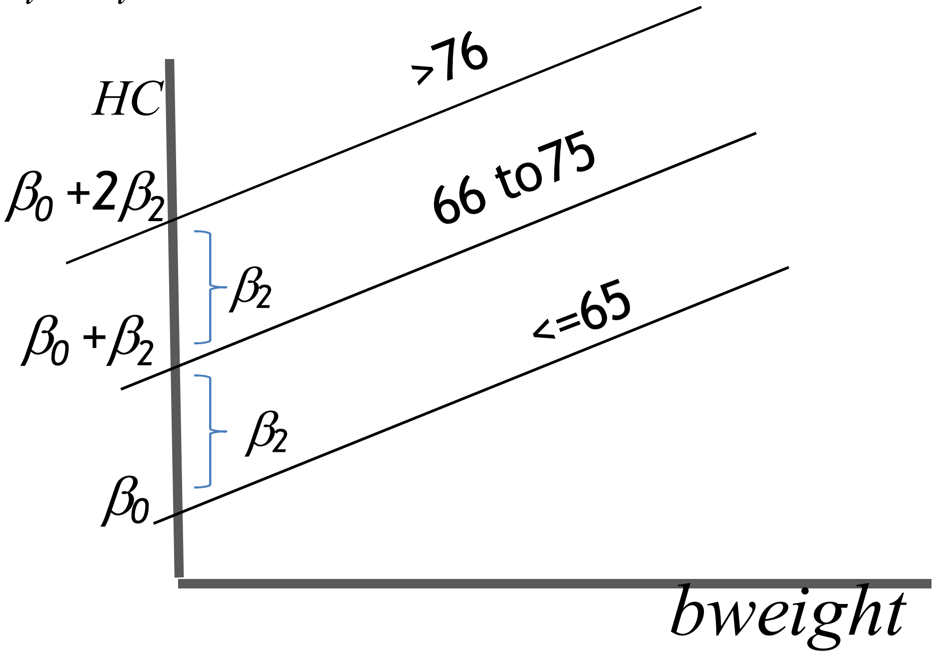
Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **For mweight = 0**

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

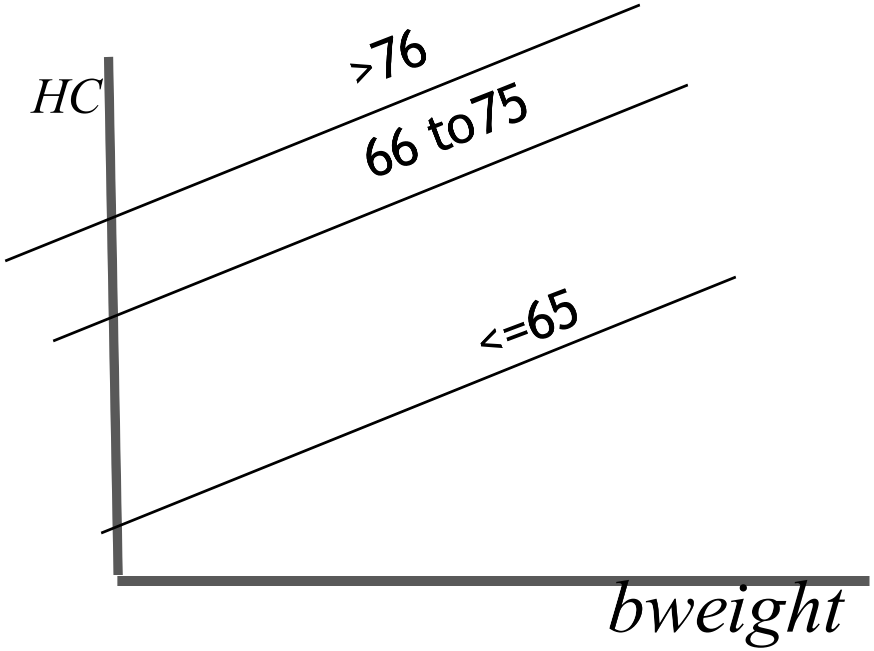
$$HC_i = (\beta_0 + 2\beta_2) + \beta_1 bweight_i + \varepsilon_i$$ **For mweight = 2**

Additionally to the parallel lines we are imposing that the difference between the 3rd and 2nd group is the same as the 2nd and 1st group



Categorical covariates

Another approach is to allow different differences between the groups we have to create indicator (dummy) variables for the categories



Let I_1 and I_2 be twoindicator variables defined as:

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

Categorical covariates

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

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

And let's use I_1 and I_2 in the model instead of mother weight

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

Categorical covariates

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

The model again defines separate straight lined for each group, but now:

$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$	**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,** $I_1=0,\ I_2=1$

Categorical covariates

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$

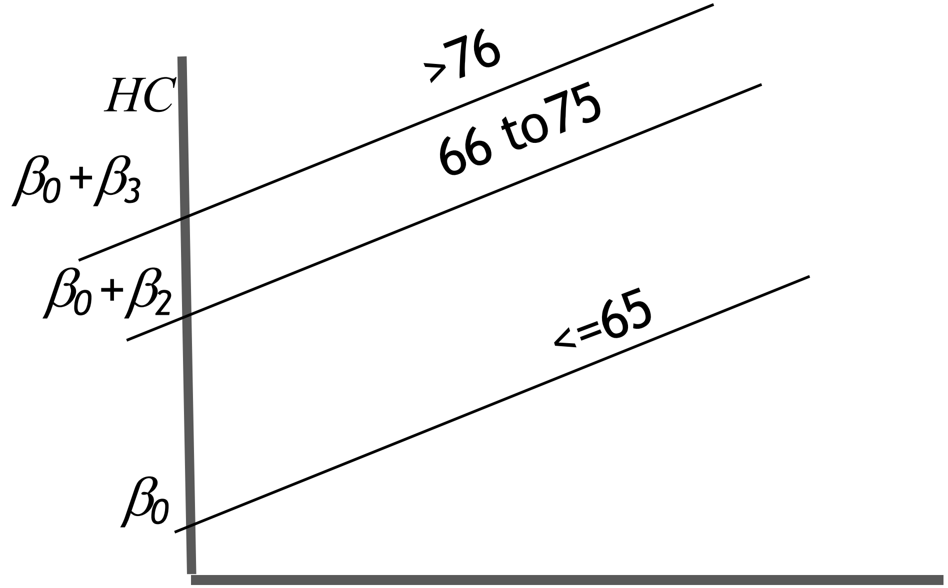
For mweight = 0, i.e, $I_1=0,\ I_2=0$

$$HC_i = (\beta_0 + \beta_2) + \beta_1 bweight_i + \varepsilon_i$$

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, $I_1=0,\ I_2=1$



Categorical covariates

- Fitting the previous model to the data, we obtain:

Model Coefficients - HC

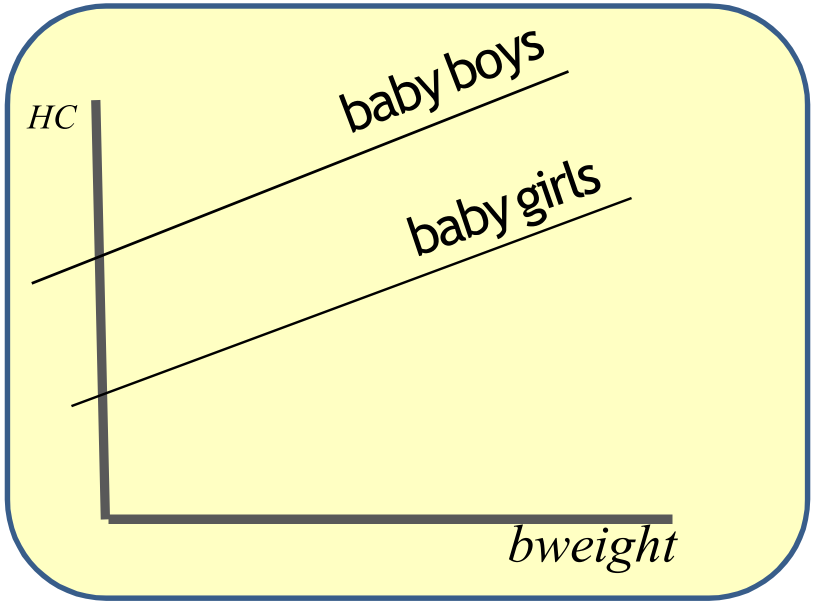
Predictor	Estimate	SE	t	p
Intercept	228.03	19.871	11.475	<.001
matwr:				
66 - 75 – <= 65	1.48	1.996	0.742	0.459
76+ – <= 65	5.94	2.127	2.793	0.006
gestimp	2.94	0.506	5.812	<.001

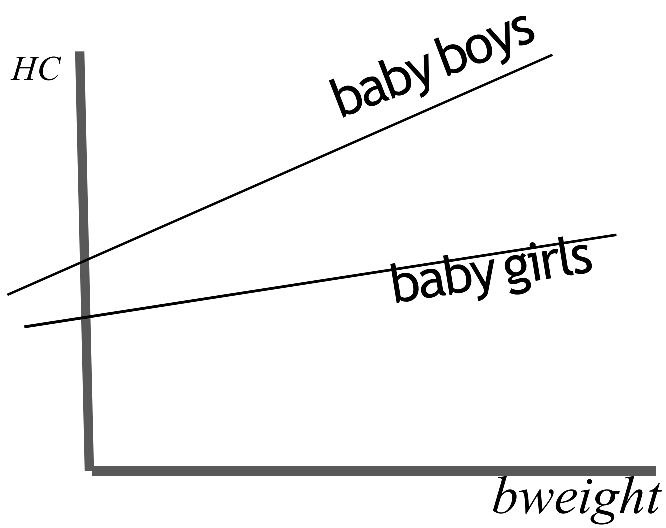
- For a fixed gestational age, the HC increases in average 1.48mm for mother weight 66-75kg in comparison with mother weight <=65kg (non-significant)

- For a fixed gestational age, the HC increases in average 5.94 mm for mother weight >76kg in comparison with mother weight <=65kg

Interactions

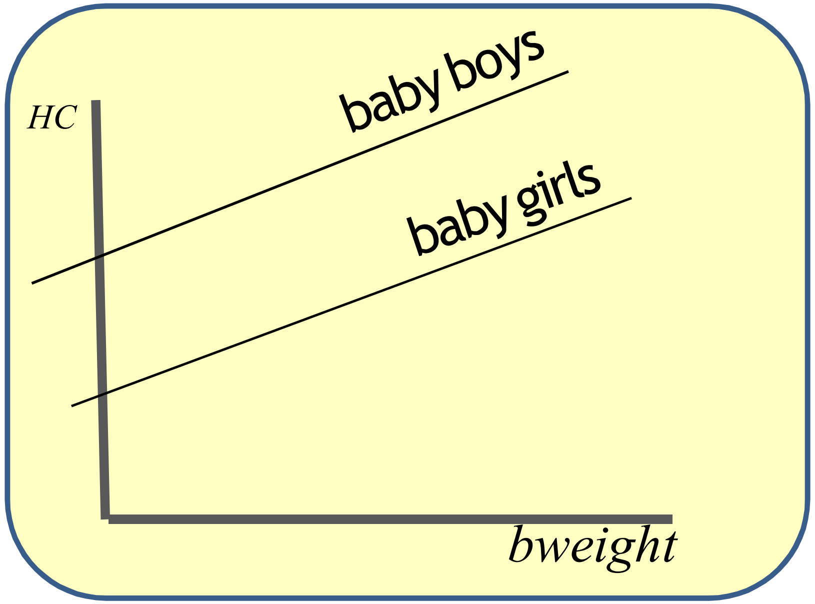
- We have seenthat $HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 sex_i + \varepsilon_i$ assumes parallel lines for bothgender

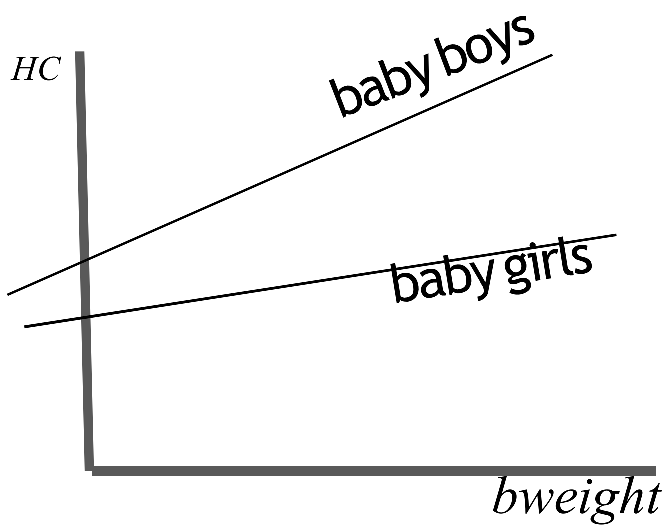




Interactions

- Recall that the parallel lines mean that the effect of birthweight is the same for both sex, i.e., the increase of 1 gr. in weight leads to the same increase in HC whether it is a girl or a boy

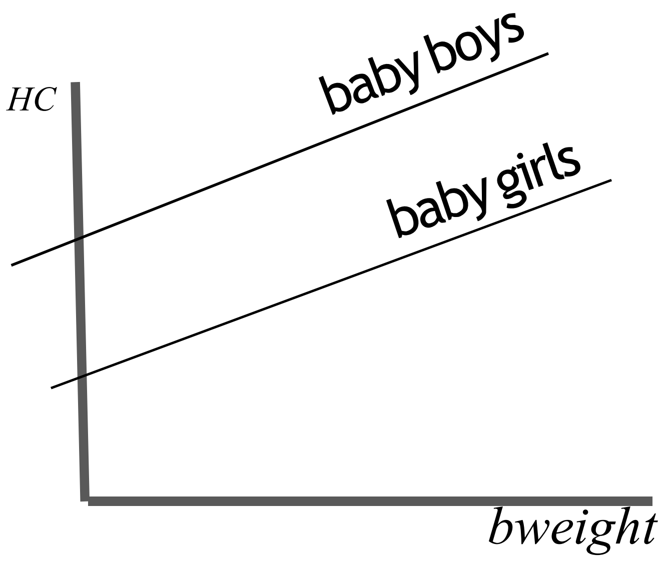


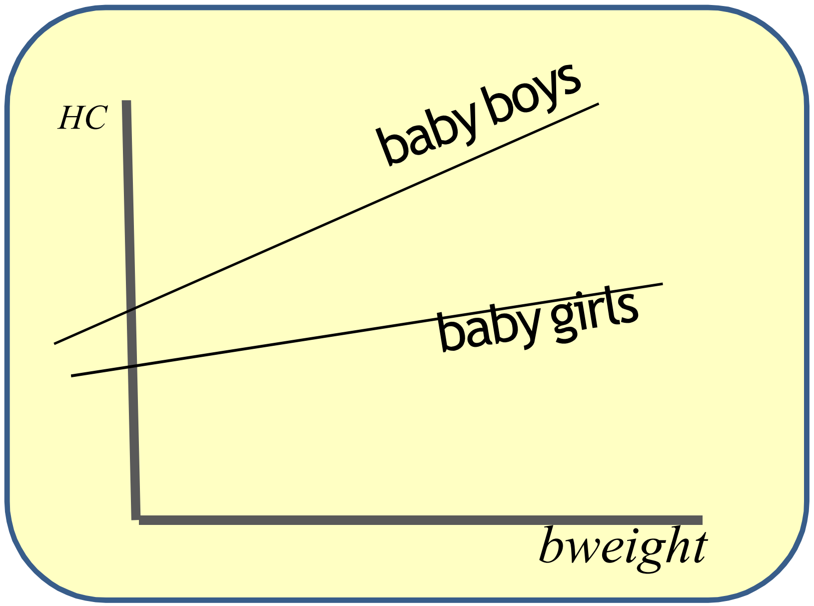


Interactions

- If the effect of birthweight is different for each sex we say that there is an interaction between birthweight and sex

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 sex_i + \beta_3 bweight_i \times sex_i + \varepsilon_i$$





Interactions

This defines two lines withdifferent intercept and different slopes

$$HC_i = \beta_0 + \beta_1 bweight_i + \beta_2 sex_i + \beta_3 bweight_i \times sex_i + \varepsilon_i$$

- For **baby girls**, sex = 0, so the model becomes

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$

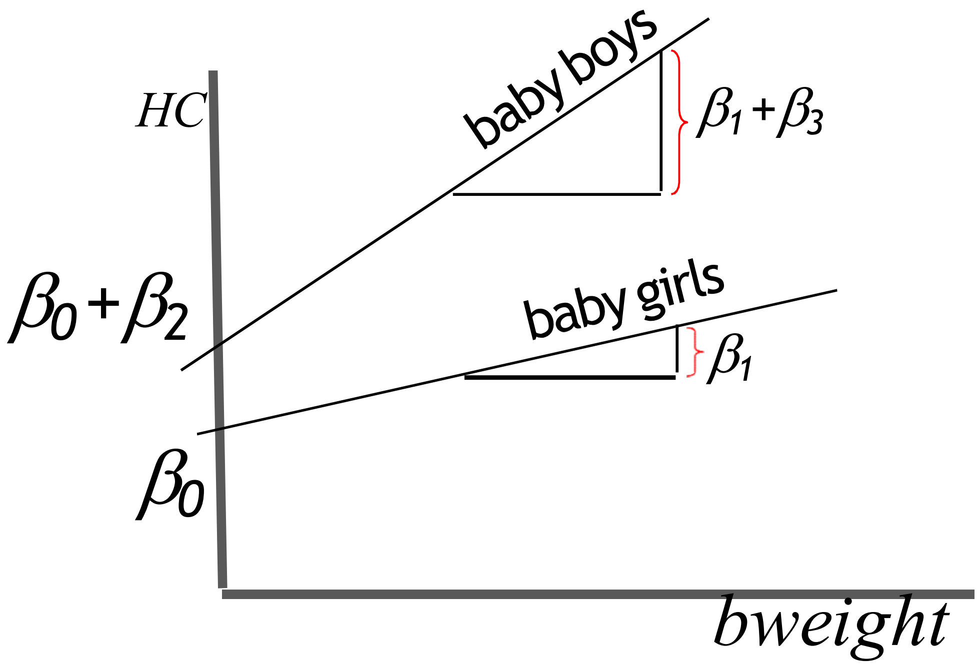
- For **baby boys**, sex = 1, so the model becomes

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

Interactions

$$HC_i = \beta_0 + \beta_1 bweight_i + \varepsilon_i$$ **baby girls**

$$HC_i = \beta_0 + \beta_2 + (\beta_1 + \beta_3) bweight_i + \varepsilon_i$$ **baby boys**

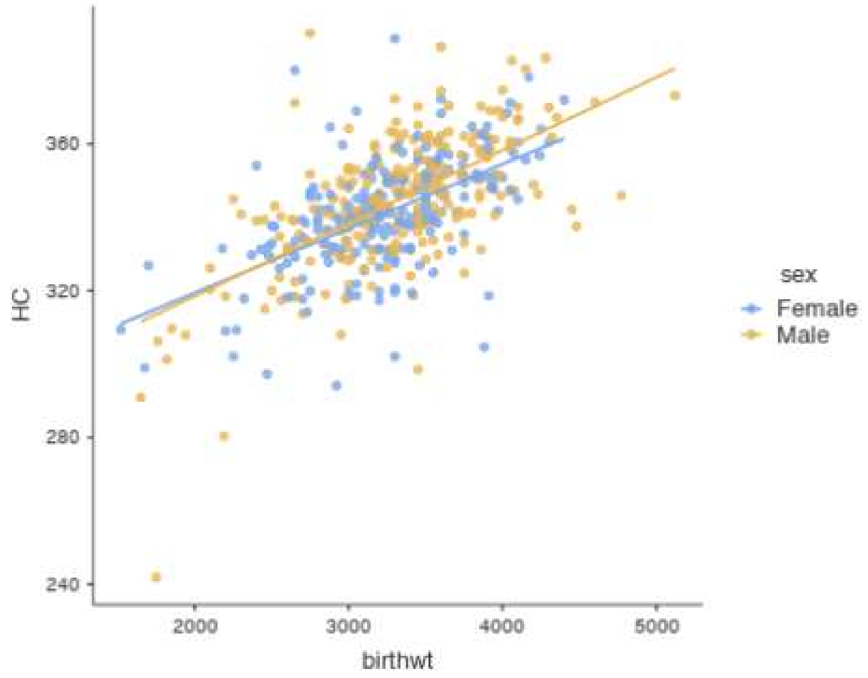


Interactions

- In the example:

•The lines represent the regression line for each sex

•Graphically, the effect of birthweight on HC does not seem to be different for boys and girls



Interactions

- In the example:

Omnibus ANOVA Test

	Sum of Squares	df	Mean Square	F	p
birthwt	42305.6	1	42305.6	231.727	<.001
sex	84.5	1	84.5	0.463	0.497
birthwt ✻ sex	165.7	1	165.7	0.908	0.341
Residuals	82155.1	450	182.6		

Note. Type 3 sum of squares

[3]

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	281.55865	4.06841	69.206	<.001
birthwt	0.01871	0.00123	15.223	<.001
sex:				
Male – Female	–5.53732	8.13682	–0.681	0.497
birthwt ✻ sex:				
birthwt ✻ (Male – Female)	0.00234	0.00246	0.953	0.341

Interactions

- In the example:

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	281.55865	4.06841	69.206	<.001
birthwt	0.01871	0.00123	15.223	<.001
sex:				
Male – Female	−5.53732	8.13682	−0.681	0.497
birthwt ∗ sex:				
birthwt ∗ (Male – Female)	0.00234	0.00246	0.953	0.341

- The effect of birthweight on HC is 0.018 for girls (sex=0) and 0.018+0.002 for boys(sex=1)

- However, the interaction is not significant (p=0.341)

Interactions

- In the example:

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	281.55865	4.06841	69.206	< .001
birthwt	0.01871	0.00123	15.223	< .001
sex:				
Male – Female	-5.53732	8.13682	-0.681	0.497
birthwt * sex:				
birthwt * (Male – Female)	0.00234	0.00246	0.953	0.341

- Note that the main effect of sex refers to the difference on the intercept and it is not (usually) of interest

- For the same reason, the test for the main effect is not of interest

Model Selection

In the previous result we observed that the interaction between sex and birthweight is non-significant.

Should we remove it from the model?

And more generally how do we decide what variables to include or exclude from the model?

There are several approaches to this problem

It is hard to argue which one is the correct approach

Probably a mix of several techniques, together with (lots of) common sense might be the right way to go

Model Selection

It is generally assumed that the researcher has some knowledge about the research topic so that he/she can identify a group of variables that are "candidates" for the model

Note that this assumption is also made at the data collection level when the researcher chooses what data to collect

Given the subset of candidate variables, we now want to build a good (best!)model

We designate this process as model building, model selection or simply, modeling

Model Selection

- Even with a small subset of covariates the number of possible models is quite large

  (including the possible interactions and polynomial terms)

- Three possible strategies are commonly used
  - Forward selection
  - Backward elimination
  - Forward and backward (stepwise)

Model Selection

- **Forward selection**
  - The first variable considered for entry into the equation is the one with the largest positive or negative correlation with the dependent variable.
  - This variable is entered into the equation only if it satisfies the criterion for entry (normally base on the p-value).
  - If the first variable is entered, the independent variable not in the equation that has the largest partial correlation is considered next.
  - The procedure stops when there are no variables that meet the entry criterion.

Model Selection

- **Forward selection**
  - In the example let's consider the covariates birthweight,gestational age, sex and mother'sage

Model Specific Results Model 1

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	282.1108	4.60404	61.3	<.001
birthwt	0.0186	0.00138	13.4	<.001

Model Specific Results Model 2

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	255.0002	17.24910	14.78	<.001
birthwt	0.0176	0.00153	11.49	<.001
gestImp	0.7793	0.47794	1.63	0.104

Model Specific Results Model 3

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	250.6210	17.33462	14.46	<.001
birthwt	0.0172	0.00153	11.21	<.001
gestImp	0.9186	0.48162	1.91	0.057
sex:				
Male – Female	2.7983	1.45383	1.92	0.055

Model Specific Results Model 4

Model Coefficients - HC

Predictor	Estimate	SE	t	p
Intercept	250.4210	17.42224	14.3736	<.001
birthwt	0.0172	0.00161	10.6597	<.001
gestImp	0.9235	0.48247	1.9141	0.056
sex:				
Male – Female	2.6768	1.47138	1.8192	0.070
matwr:				
66 - 75 – <= 65	−1.3247	1.76297	−0.7514	0.453
76+ – <= 65	0.0528	1.93097	0.0273	0.978

Model Selection

- Backward elimination
  - A variable selection procedure in which all variables are entered into the equation and then sequentially removed.
  - The variable with the smallest partial correlation with the dependent variable is considered first for removal.
  - If it meets the criterion for elimination, it is removed.
  - After the first variable is removed, the variable remaining in the equation with the smallest partial correlation is considered next.
  - The procedure stops when there are no variables in the equation that satisfy the removal criteria (based on the p-value).

Model Selection

- Forward and backward(stepwise)
  - At each step, the independent variable not in the equation that has the smallest probability of F is entered, if that probability is sufficiently small.
  - Variables already in the regression equation are removed if their probability of F becomes sufficientlylarge.
  - The method terminates when no more variables are eligible for inclusion or removal.
  - This is very similar to the Forward selection but each variable is reevaluated at each step and can be excluded after being included

Cheers!

Title

Multiple Linear Regression

Acknowledgments

Armando Teixeira Pinto, University of Sydney, Australia

Basics of Health Intelligent Data Analysis

PhD Programme in Health Data Science

Porto, 2nd of December, 2019

Cláudia Camila Dias

Pedro Pereira Rodrigues