A tutorial for Linear Regression in R

Version 1.0.0

Konstantinos I. Bougioukas

20/01/2022

Contents

Preface

Regression analysis is at the very heart of applied statistics. It is a form of mathematical modelling that identify the associations between dependent and independent variables. In this lesson, we'll describe some of the key concepts and techniques underlying the **linear** regression analysis.

Objectives

- fit and interpret (simple and multiple) linear models
- assess the quality of a linear regression fit
- evaluate the appropriateness of the chosen linear model with diagnostic plots and statistical tests

Download and load the following packages:

```
library(summarytools)
library(GGally)
library(pmsampsize)
library(jtools)
library(moderndive)
library(rstatix)
library(skimr)
library(ggstance)
library(ggstatsplot)
library(ggpubr)
library(interactions)
library(performance)
library(gvlma)

library(here)
library(tidyverse)
```

1 Introduction

The fundamental premise of data modeling is to make explicit the association between:

- an *outcome variable* y, also called a *dependent variable* or *response variable*, and
- one or more *explanatory/predictor variables* $x_1, x_2, ..., x_p$, also called *independent variables* or covariates (Figure 1).

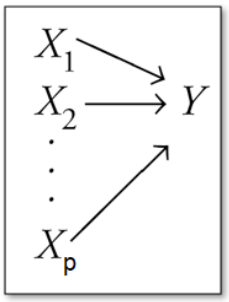


Figure 1: The basic idea of regression modeling

Another way to state this is using mathematical terminology: we will model the outcome variable y "as a function" of the explanatory/predictor variables $x_1, x_2, ..., x_p$. When we say "function" here, we aren't referring to functions in R like the `ggplot()` function, but rather as a mathematical function. But, why do we have two different labels, explanatory and predictor, for the variables $x_1, x_2, ..., x_p$? That's because even though the two terms are often used interchangeably, roughly speaking data modeling serves one of two purposes:

1. **Modeling for explanation**: When we want to explicitly describe and quantify the association between the outcome variable y and a set of explanatory variables $x_1, x_2, ..., x_p$, determine the significance of any associations, have measures summarizing these associations, and possibly identify any *causal* associations between the variables. For example, directed acyclic graphs (DAGs) provide a simple and transparent way for observational data scientists to identify and demonstrate their knowledge, theories and assumptions about the causal relationships between variables.

2. **Modeling for prediction**: When we want to predict an outcome variable y based on the information contained in a set of predictor variables $x_1, x_2, ..., x_p$. Clinical prediction models usually fall within one of two major categories: **diagnostic** prediction models that estimate an individual's probability of a specific health condition (often a disease) being currently present, and **prognostic** prediction models that estimate the probability of developing a specific health outcome over a specific time period. Unlike modeling for explanation, however, we don't care so much about understanding how all the variables relate and interact with one another, but rather only whether we can make good predictions about y using the information in $x_1, x_2, ..., x_p$. Prediction models focus on the performance of the model as a whole.

In this course, we'll focus on modeling for explanation and hence refer to $x_1, x_2, ..., x_p$ as *explanatory variables*. Furthermore, while there exist many techniques for modeling, such as tree-based models and neural networks, in this lesson we'll focus on one particular technique: *Regression*. Regression is one of the most commonly-used and easy-to-understand approaches to modeling.

2 The many faces of regression: Linear regression

Nowadays, the term "regression" includes many specialized varieties (Figure 2). This terminology is due to Sir Francis Galton (1822-1911) who noticed that tall and short men tend to have sons with heights closer to the mean. He called this "regression towards the mean."

In this course, regression analysis is used for explaining or modeling the association between a single variable Y, let's call this response variable (or just outcome), and one or more explanatory variables, $x_1, x_2, ..., x_p$. When p=1, it is called simple regression but when p>1 it is called multiple regression. We are going to deal with Linear, Logistic and Cox regression during this course. Fortunately, `R` has powerful and comprehensive features for fitting regression models.

It should be noted here that **linear** regression extents the t-test, **logistic** regression extends the chi-squared test and **Cox** regression extends the log-rank test.

	Multiple logistic regression	Multiple Cox regression	Multiple linear regression / Multiple ANOVA / ANCOVA
Dependent variable	**Dichotomous** (no information about timepoint) *Example: Treatment response (yes/no)*	**Time to event** (dichotomous with information about timepoint) *Example: Overall survival*	**Quantitative** *Example: Blood pressure*
Independent variables	2 or more quantitative or categorical variables	2 or more quantitative or categorical variables	2 or more quantitative or categorical variables[a]
Equation[b]	$\text{logit}(p) = a + b_1x_1 + b_2x_2 \ldots$	$\log(h_i(t)) = a + b_1x_1 + b_2x_2 \ldots$	$y = a + b_1x_1 + b_2x_2 \ldots$
Parameter	OR (= Exp(b))	HR (= Exp(b))	β (= b)
Interpretation	Odds for: • Category X vs reference category (if independent variable is categorical) • A 1-unit increase (if independent variable is quantitative)	Instantaneous risk/hazard (hazard per unit time) for: • Category X vs reference category (if independent variable is categorical) • A 1-unit increase (if independent variable is quantitative)	Size of the effect on the outcome (in outcome units) for: • Category X vs reference category (if independent variable is categorical) • A 1-unit increase (if independent variable is quantitative)
Example of reporting	*"… odds of treatment failure were 3 times higher in men than in women"*	*"… risk of death was 3 times higher in men versus women"*	*"… systolic blood pressure was 3 mmHg higher in men than in women"*

[a] For ANOVA and ANCOVA at least 1 categorical variable is needed

[b] $\text{logit}(p)$ is $\log(p/1\text{-}p)$, where p is the probability of the outcome; a denotes a constant, b_n denotes the coefficient for each independent variable, x_n denotes an independent variable, $h_i(t)$ is the hazard to individual i at time t, and y denotes a dependent variable

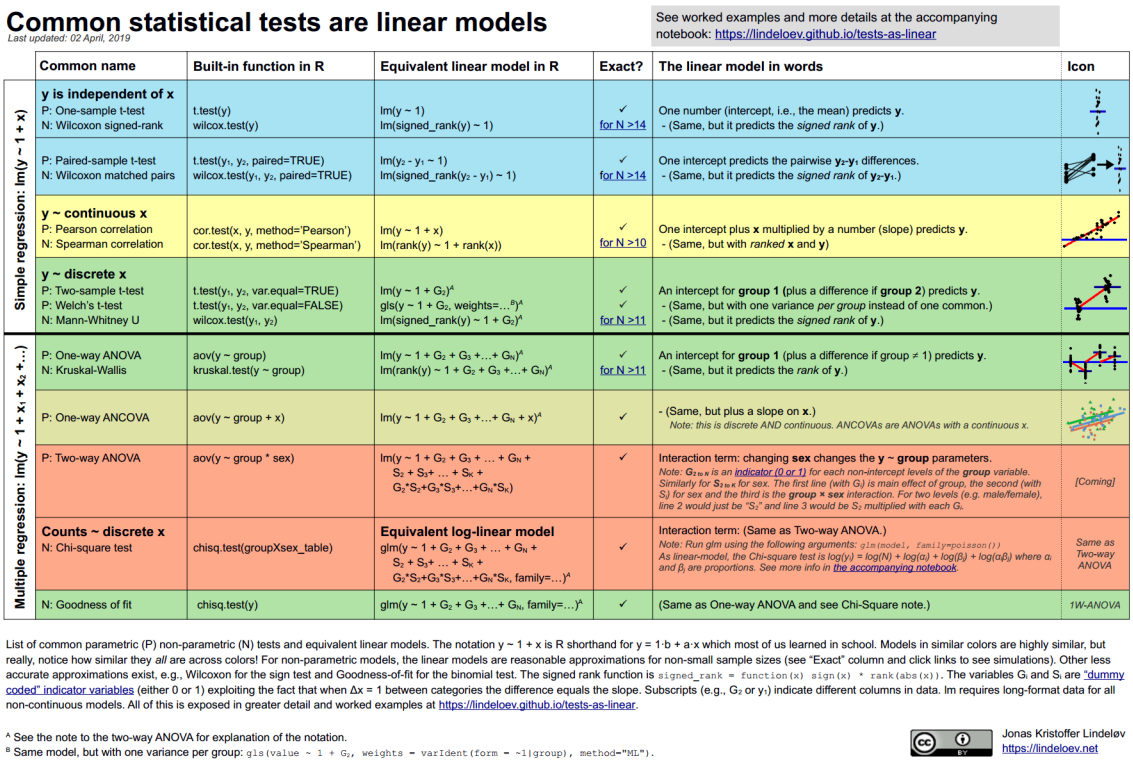
Figure 2: Types of multivariable models commonly used in biomedical studies

Linear regression

These notes are about linear regression. Linear regression involves a **numerical** outcome variable y and explanatory variables x that are either **numerical** or **categorical** (Figure 2, Figure 3). Furthermore, the association between y and $x_1, x_2, ..., x_p$ is assumed to be **linear**, or in other words, a line. However, we'll see that what constitutes a "line" will vary depending on the nature of our explanatory variables $x_1, x_2, ..., x_p$.

Firstly, we'll consider models with a single (numerical or categorical) explanatory variable x (**simple linear regression**).

Then, we'll extend the ideas behind simple regression and consider models with more explanatory variables (**multiple linear regression**).



Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon
Simple regression: lm(y ~ 1 + x)	**y is independent of x** P: One-sample t-test N: Wilcoxon signed-rank	t.test(y) wilcox.test(y)	lm(y ~ 1) lm(signed_rank(y) ~ 1)	✓ for N >14	One number (intercept, i.e., the mean) predicts **y**. - (Same, but it predicts the *signed rank* of **y**.)	
	P: Paired-sample t-test N: Wilcoxon matched pairs	t.test(y_1, y_2, paired=TRUE) wilcox.test(y_1, y_2, paired=TRUE)	lm(y_2 - y_1 ~ 1) lm(signed_rank(y_2 - y_1) ~ 1)	✓ for N >14	One intercept predicts the pairwise $\mathbf{y_2}$-$\mathbf{y_1}$ differences. - (Same, but it predicts the *signed rank* of $\mathbf{y_2}$-$\mathbf{y_1}$.)	
	y ~ continuous x P: Pearson correlation N: Spearman correlation	cor.test(x, y, method='Pearson') cor.test(x, y, method='Spearman')	lm(y ~ 1 + x) lm(rank(y) ~ 1 + rank(x))	✓ for N >10	One intercept plus **x** multiplied by a number (slope) predicts **y**. - (Same, but with *ranked* **x** and **y**)	
	y ~ discrete x P: Two-sample t-test P: Welch's t-test N: Mann-Whitney U	t.test(y_1, y_2, var.equal=TRUE) t.test(y_1, y_2, var.equal=FALSE) wilcox.test(y_1, y_2)	lm(y ~ 1 + G_2)[A] gls(y ~ 1 + G_2, weights=…)[B][A] lm(signed_rank(y) ~ 1 + G_2)[A]	✓ ✓ for N >11	An intercept for **group 1** (plus a difference if **group 2**) predicts **y**. - (Same, but with one variance *per group* instead of one common.) - (Same, but it predicts the *signed rank* of **y**.)	
Multiple regression: lm(y ~ 1 + x_1 + x_2 +…)	P: One-way ANOVA N: Kruskal-Wallis	aov(y ~ group) kruskal.test(y ~ group)	lm(y ~ 1 + G_2 + G_3 +…+ G_N)[A] lm(rank(y) ~ 1 + G_2 + G_3 +…+ G_N)[A]	✓ for N >11	An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**. - (Same, but it predicts the *rank* of **y**.)	
	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +…+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: $G_{2 \text{ to } N}$ is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for $S_{2 \text{ to } K}$ for sex. The first line (with G_2) is main effect of group, the second (with S_2) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + … + G_N + S_2 + S_3+ … + S_K + G_2*S_2+G_3*S_3+…+G_N*S_K, family=…)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments: `glm(model, family=poisson())`* *As linear-model, the Chi-square test is log(y) = log(N) + log(α) + log(β) + log(αβ) where α_i and β_i are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +…+ G_N, family=…)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for $y = 1 \cdot b + a \cdot x$ which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when $\Delta x = 1$ between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G2, weights = varIdent(form = ~1|group), method="ML")`.

Figure 3: Basic statistical tests <-> linear models

Common statistical tests are linear models!

3 Dataset preparation

Data of 550 infants at 1 month age was collected (BirthWeight). The following variables were recorded (Table 1):

- Body weight of the infant in kg (weight)
- Body height of the infant in cm (height)
- Head circumference in cm (headc)
- Gender of the infant (gender: Female, Male)
- Birth order in their family (parity: Singleton, One sibling, 2 or more siblings)
- Education of the mother (education: tertiary, year10, year12)

We import the data:

```
library(readxl)
BirthWeight <- read_excel(here("data", "BirthWeight.xlsx"))
```

Then we inspect the data:

Table 1: Birth Weight Data (first and last 5 rows)

id	weight	height	headc	gender	education	parity
L001	3.95	55.5	37.5	Female	tertiary	2 or more siblings
L003	4.63	57	38.5	Female	tertiary	Singleton
L004	4.75	56	38.5	Male	year12	2 or more siblings
L005	3.92	56	39	Male	tertiary	One sibling
L006	4.56	55	39.5	Male	year10	2 or more siblings
NA	...	...	...	NA	NA	NA
W319	5.35	57	39.5	Male	tertiary	2 or more siblings
W320	5.39	60	40	Male	tertiary	Singleton
W321	3.88	52	36	Male	year10	One sibling
W322	5.23	57.5	40	Male	year10	2 or more siblings
W323	4.57	53.5	37.5	Female	tertiary	2 or more siblings

A useful function that presents descriptive statistics for our variables in the Viewer panel is the following:

```
summarytools::view(dfSummary(BirthWeight))
```

First, we need to make some transformations to the variables (`weight` in grams, `gender` to factor, `education` to factor with reference category `year10`, `parity` to factor with reference category the `singletons`):

```
BirthWeight <- BirthWeight %>%
  dplyr::select(-id) %>%  # remove id variable
  mutate(weight = weight*1000,  # multiply by 1000
    gender = factor(gender),
    education = fct_relevel(education, "tertiary", after = Inf), # tertiary to the end
    parity = factor(parity, levels = c("Singleton", "One sibling",
                                       "2 or more siblings"))  # singleton first
  )
```

We inspect the variables again:

```
summarytools::view(dfSummary(BirthWeight))
```

```
# or with skimr
skimr::skim(BirthWeight)
```

4 Simple linear regression

Simple linear regression refers to linear regression models with a single explanatory variable x.

You may recall from secondary/high school algebra that the equation of a line is $y = a \cdot x + \beta$. It is defined by two coefficients α and β. The slope coefficient α for x is the change in y for every one unit increase in x. The intercept coefficient β is the value of y when $x = 0$ (the point where the fitted line crosses the y-axis).

However, when defining a regression line, we use slightly different notation: the equation of the regression line is $\widehat{y} = b_0 + b_1 \cdot x$. The intercept coefficient is b_0, so b_0 is the value of $\widehat{y}$ when $x = 0$. The slope coefficient for x is b_1, i.e., the change in $\widehat{y}$ for every one unit increase in x. Why do we put a "hat" on top of the y? It's a form of notation commonly used in regression to indicate that we have a "fitted value," or the value of y on the regression line for a given x value.

4.1 Simple linear regression with a continuous explanatory variable

Let's say that we want to explore the association between weight and height for the sample of 550 infants of 1 month age.

A first step that is usually useful in studying the association between two continuous variables is to prepare a scatter plot of the data (Figure 4). The pattern made by the points plotted on the scatter plot usually suggests the basic nature and strength of the association between two variables.

```
# correlation graph applying ggscatmat() function from GGally package
BirthWeight %>%
  select(weight, height) %>%
  ggscatmat(corMethod = "pearson")
```

```
BirthWeight %>%
  select(weight, height) %>%
  cor_test(method="pearson")
```

Table 2: Correlation table for weight and height

var1	var2	cor	statistic	p	conf.low	conf.high	method
weight	height	0.71	23.813	0	0.669	0.752	Pearson

There is a significant high positive linear correlation (r=0.71, 95%CI: 0.67 to 0.75, p<0.001) between weight and height for infants of 1 month age (Table 2).

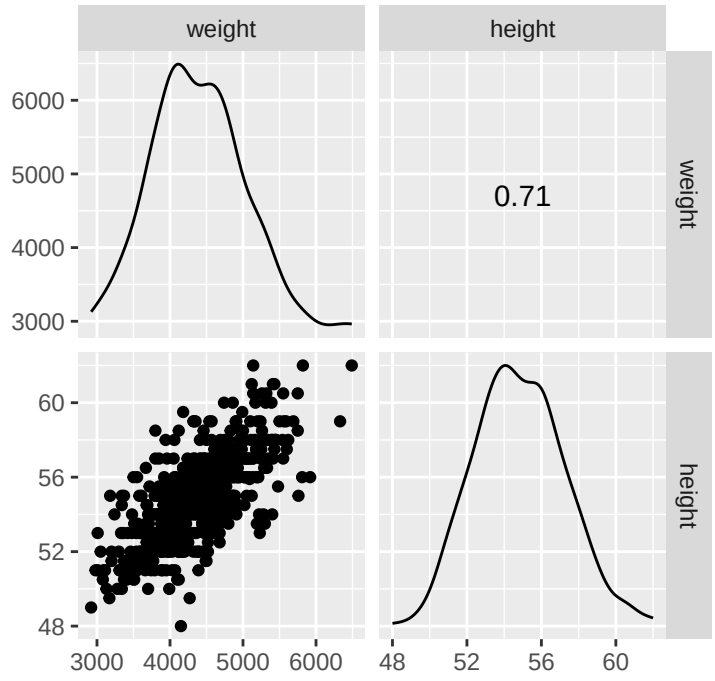


Figure 4: Correlation graph between weight and height

As you can see, the points seem to be scattered around an invisible line (Figure 4). The scatter plot also shows that, in general, infants with high height tend to have high weight (positive association). The Pearson's correlation coefficient r, quantifies the strength of this association. This coefficient ranges from -1 **to** 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases (perfect positive association). A value of -1 implies that all data points lie on a line for which Y decreases as X increases (perfect negative association). A value of **zero** implies that there is no linear correlation between the variables.

Now, we are interested in finding the regression equation of the line (Figure 5):

$$\widehat{y} = b_0 + b_1 \cdot x$$

We can obtain the values of the intercept b_0 and the slope for `height`, b_1, by outputting a *linear regression table*. This is done in two steps:

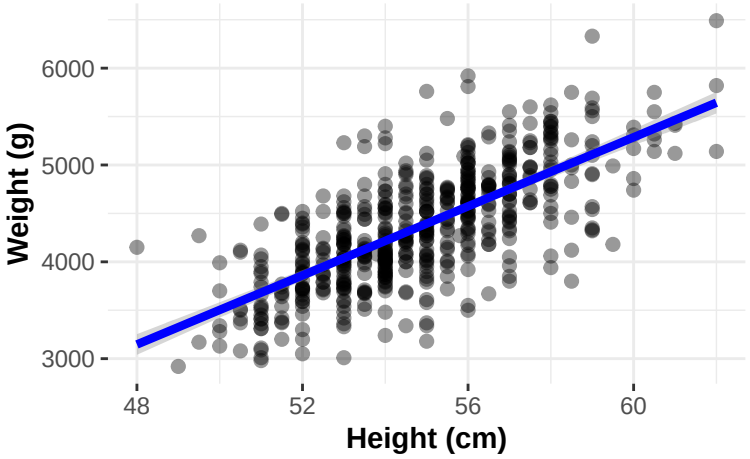


Figure 5: Scatter plot with the regression line

1. We first "fit" the linear regression model using the `lm()` function and save it in `model_height`.
2. We get the regression table by applying the `summary()` and `confint()` functions to `model_height`.

```
# Fit regression model:
model_height <- lm(weight ~ height, data = BirthWeight)

summary(model_height)
```

```
##
## Call:
## lm(formula = weight ~ height, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1218.86  -263.13   -24.02   282.29  1365.21
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5412.145    411.040  -13.17   <2e-16 ***
```

```
## height        178.308      7.488   23.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 422.3 on 548 degrees of freedom
## Multiple R-squared:  0.5085, Adjusted R-squared:  0.5076
## F-statistic:   567 on 1 and 548 DF,  p-value: < 2.2e-16
```

```
confint(model_height, level=0.95)
```

```
##                   2.5 %     97.5 %
## (Intercept) -6219.5518 -4604.7374
## height        163.5992   193.0164
```

An alternative to obtain the results is by applying the `get_regression_table()` function from the `{moderndive}` package:

```
# Get regression table:
get_regression_table(model_height)
```

Table 3: Linear regression table: height

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-5412.145	411.040	-13.167	0	-6219.552	-4604.737
height	178.308	7.488	23.813	0	163.599	193.016

Let's first focus on interpreting the regression Table 3. In the `estimate` column are the intercept b_0 = -5412.145 and the slope b_1 = 178.308 for `height`. Thus the equation of the regression line follows:

$$\begin{aligned} \widehat{y} &= b_0 + b_1 \cdot x \\ \widehat{\text{weight}} &= b_0 + b_1 \cdot \text{height} \\ &= -5412.145 + 178.308 \cdot \text{height} \end{aligned}$$

The intercept b_0 = -5412.145 is the average weight $\widehat{y}$ = $\widehat{\texttt{weight}}$ for those infants with `height` of 0. Or in graphical terms, it's where the line intersects the y axis when x = 0 (Figure 6). Note, however, that while the intercept of the regression line has a mathematical interpretation, it has no *physical* interpretation here, since observing a `weight` of 0 is impossible.

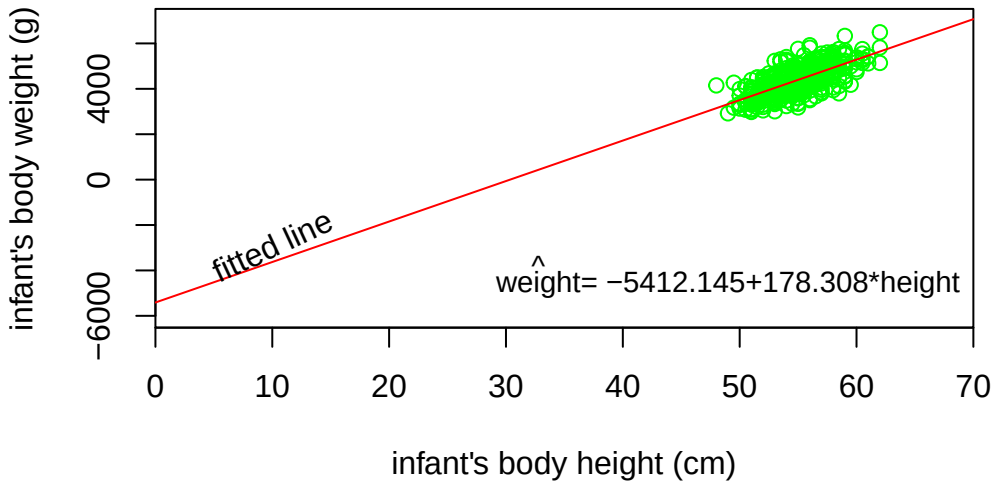


Figure 6: Scatter plot of infants' body height-body weight with fitted line crossing the y-axis

Of greater interest is the slope b_1 for `height` of 178.308, as this summarizes the association between the height and weight variables.

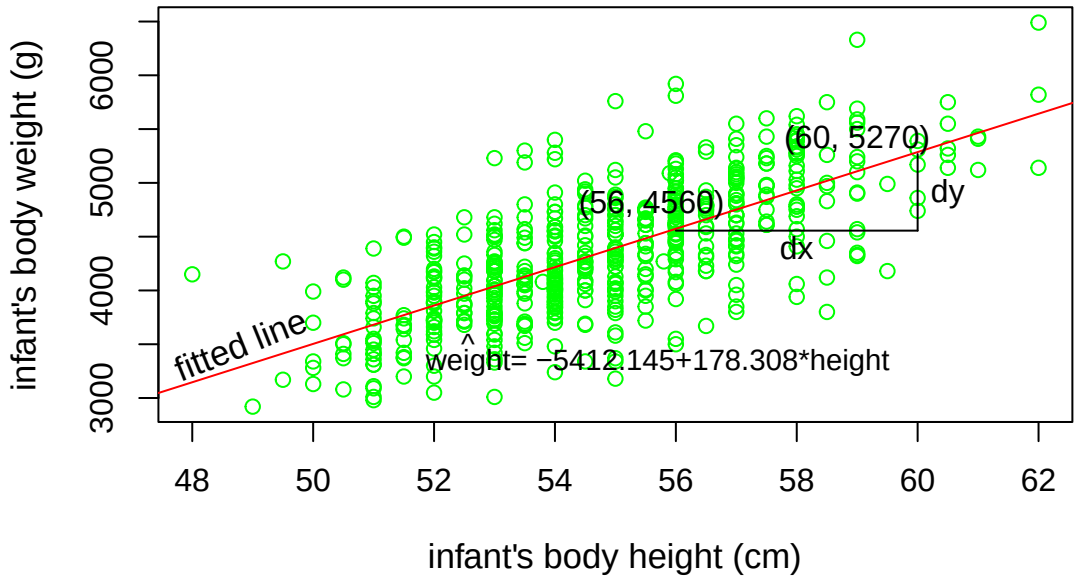


Figure 7: Scatter plot of infants' body height-body weight and graphically calculation of the slope

The graphical calculation of the slope from two points of the fitted line is (Figure 7):

$$b_1 = \frac{dy}{dx} = \frac{5270 - 4560}{60 - 56} = \frac{710}{4} \approx 178$$

Note that, in this example, the coefficient has units g/cm.

Additionally, note that the sign is positive, suggesting a positive association between these two variables, meaning infants with higher height also tend to have higher weight. Recall from earlier that the correlation coefficient is r = 0.71. They both have the same positive sign, but have a different value. Recall further that the correlation's interpretation is the "strength of linear association". The slope's interpretation is a little different:

For every 1 unit increase in 'height', there is **on average** an **associated** increase of 178.308 units of 'weight'.

We only state that there is an *associated* increase and not necessarily a *causal* increase. In other words, just because two variables are strongly associated, it doesn't necessarily mean that one causes the other. This is summed up in the often quoted phrase, "correlation is not necessarily causation."

Furthermore, we say that this associated increase is **on average** 178.308 units of weight, because we might have two infants whose height differ by 1 unit, but their difference in weight won't necessarily be exactly 178.308. What the slope of 178.308 is saying is that across all possible infants, the *average* difference in weight between two infants whose height differ by 1 cm is 178.308 g.

In summary, we can say that the regression coefficient of the height (178) is significantly different from zero (p < 0.001) and indicates that there's on average an increase of 178 g (95%CI: 164 to 193) in weight for every 1 cm increase in height. Note that the 95%CI does not include the hypothesized null value of zero for the slope.

How much is the average increase in weight for each 10 cm increase in height?

Standard error (SE)

The third column of the regression table in Table 1 `std_error` corresponds to the *standard error* of our estimates.

> Say we hypothetically collected 1000 samples of pairs of weight and height, computed the 1000 resulting values of the fitted slope b_1, and visualized them in a histogram. This would be a visualization of the *sampling distribution* of b_1. The standard deviation of the *sampling distribution* of b_1 has a special name: the *standard error*.

The *standard error* of b_1 quantifies how much variation in the fitted slope b_1 one would expect between different samples. So in our case, we can expect about 7.5 units of variation in the **slope** of `height` variable.

Test statistic and confidence intervals

The fourth column of the regression Table 1 `statistic` corresponds to a *test statistic* relating to the following *hypothesis test*:

$$H_0 : \beta_1 = 0$$
$$\text{vs } H_A : \beta_1 \neq 0.$$

The null hypothesis states that the coefficient of the explanatory variable (slope) is equal to zero, and the alternative hypothesis states that the coefficient of the explanatory variable is not equal to zero.

The `statistic` column in the regression table is a tricky one, however. It corresponds to a standardized *t-test statistic*, much like the *two-sample* t *statistic* we saw in `Introductory Statistics` where we used a theory-based method for conducting hypothesis tests. In both these cases, the *null distribution* can be mathematically proven to be a t*-distribution*.

The t-statistic, here, is defined by the following equation:

$$t = \frac{b_1}{\mathsf{SE}_{b_1}}$$

In our example:

$$t = \frac{b_1}{\mathsf{SE}_{b_1}} = \frac{178.308}{7.488} = 23.81$$

The 95%CI (confidence interval) of the coefficient b_1 for a significance level a = 0.05, $df = n - 2$ degrees of freedom and for a two-tailed t-test is given by

$$95\%\mathsf{CI}_{b_1} = b_1 \pm 1.96 \cdot \mathsf{SE}_{b_1}.$$

In our example:

$$95\%\mathsf{CI}_{b_1} = 178.308 \pm 1.96 \cdot 7.488 \Rightarrow 95\%\mathsf{CI}_{b_1} = (163.6, 193).$$

Observed, fitted values and residuals

We define the following three concepts:

1. **Observed** values y, or the observed value of the outcome variable for a given x value
2. **Fitted** values $\widehat{y}$, or the value on the regression line for a given x value
3. **Residuals** $y - \widehat{y}$, or the error between the observed value and the fitted value for a given x value

We obtained these values for our dataset using the `get_regression_points()` function from the `{moderndive}` package (Table 4).

```
regression_points <- get_regression_points(model_height)
regression_points
```

Table 4: Regression points (First 10 out of 550 infants)

ID	weight	height	weight_hat	residual
1	3950	55.5	4483.939	-533.939
2	4630	57.0	4751.401	-121.401
3	4750	56.0	4573.093	176.907
4	3920	56.0	4573.093	-653.093
5	4560	55.0	4394.785	165.215
6	3640	51.5	3770.708	-130.708
7	3550	56.0	4573.093	-1023.093
8	4530	57.0	4751.401	-221.401
9	4970	58.5	5018.863	-48.863
10	3740	52.0	3859.862	-119.862

Observe in the above table that `weight_hat` contains the fitted values $\widehat{y}$ = $\widehat{\text{weight}}$.

The `residual` column is simply $y - \widehat{y}$ = `weight - weight_hat`.

Let's see, for example, the values for the first infant and have a visual representation (Figure 8):

- **Circle**: The *observed value* y = 3950 is infant's weight for x = 55.5.

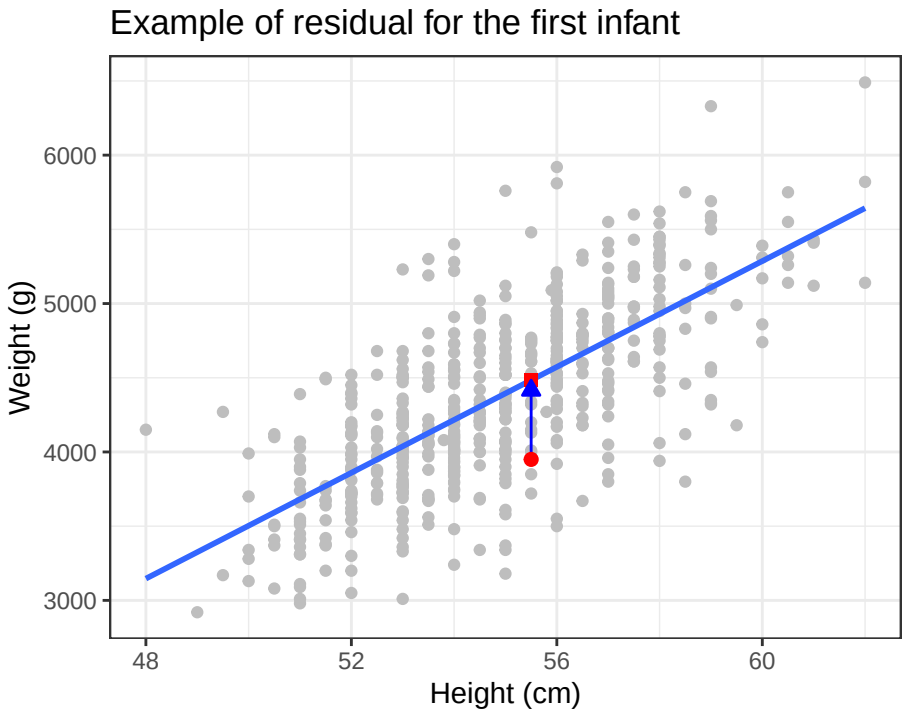


Figure 8: Example of observed value, fitted value, and residual.

- **Square**: The *fitted value* $\widehat{y}$ is the value 4483.939 on the regression line for x = 55.5. This value is computed using the intercept and slope in the previous regression Table 3:

$$\widehat{y} = b_0 + b_1 \cdot x = -5412.145 + 178.308 \cdot 55.5 = 4483.9$$

- **Arrow**: The length of this arrow is the *residual* and is computed by subtracting the fitted value $\widehat{y}$ from the observed value y. The residual can be thought of as a model's error or "lack of fit" for a particular observation. In the case of this infant, it is $y - \widehat{y}$ = 3950 - 4483.9 = -533.9 .

The residuals are exactly the **vertical** distance between the observed data point and the associated point on the regression line (Figure 9). Positive residuals have associated y values above the fitted line and negative residuals have values below. We want the residuals to be **small** in magnitude, because large negative residuals are as bad as large positive residuals.

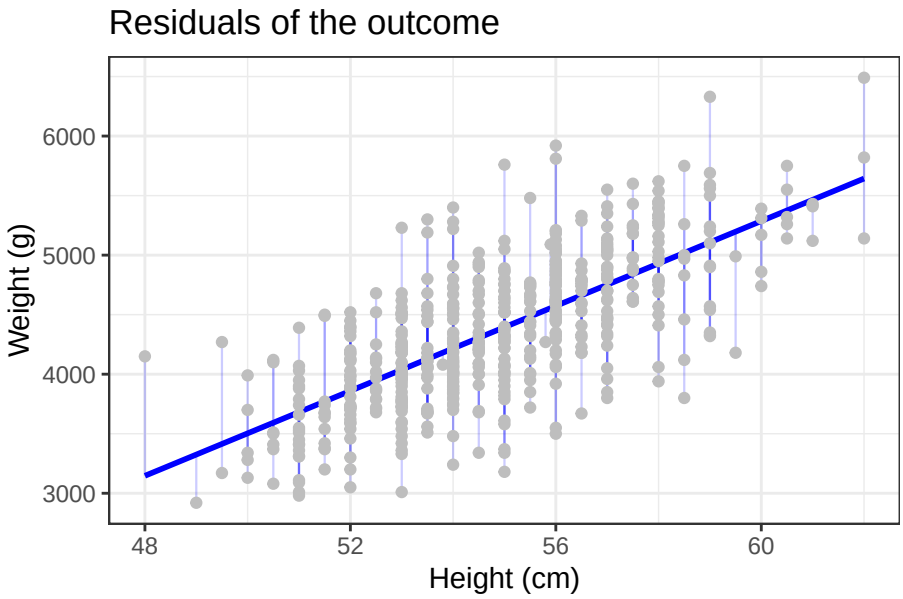


Figure 9: Example of observed value, fitted line, and residuals.

A “best-fitting” line refers to the line that **minimizes** the sum of squared residuals (RSS), also known as sum of squared estimate of errors (SSE) out of all possible lines we can draw through the points.

$$minimize(RSS) = minimize \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

In Figure 10, we have found the minimum value of RSS (it turns out to be 97723317) and have drawn a horizontal dashed green line. At the point where this minimum touches the graph, we have read down to the x axis to find the best value of the slope (the red arrow). This is the value 178.

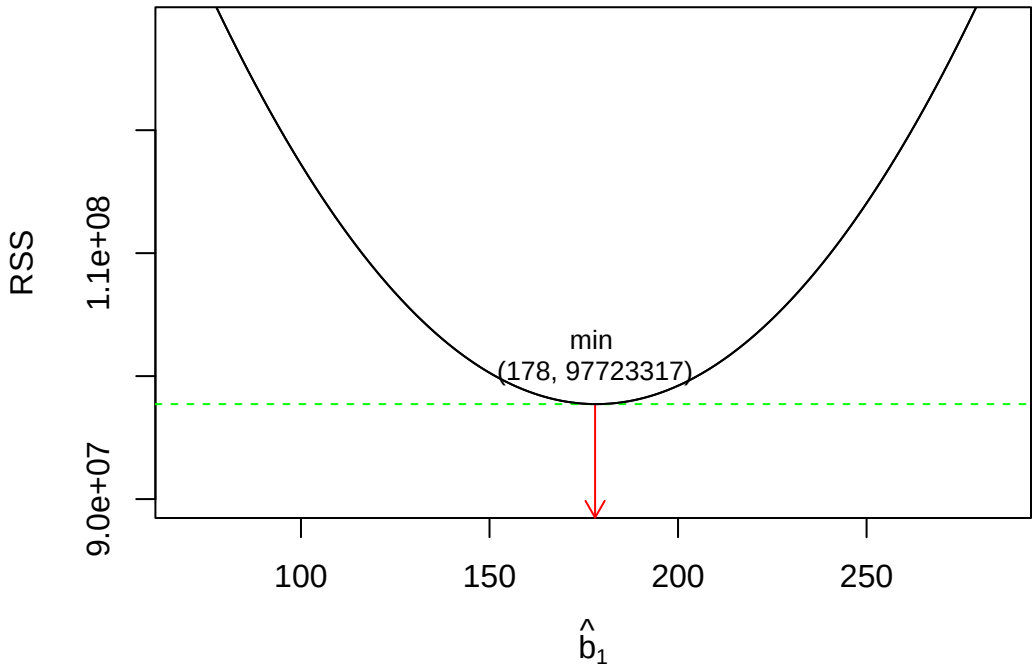


Figure 10: The sum of the squares of the residuals against the value of the coefficient of the slope which we are trying to estimate.

Quality of a linear regression fit

The quality of a linear regression fit is typically assessed using two related quantities: residual standard error (RSE) and the coefficient of determination R^2.

- **Residual standard error** (RSE)

RSE represents the average distance that the observed values fall from the regression line. Conveniently, it tells us how wrong the regression model is on average using the units of the response variable.

Smaller values are better because it indicates that the observations are closer to the fitted line. In our example,

$$RSE = \sqrt{\frac{RSS}{n-2}} = \sqrt{\frac{97723317}{550-2}} = 422.3$$

- **Coefficient of determination,** R^2

The R^2 is the fraction of the total variation in y that is explained by the regression.

$$R^2 = \frac{explained\ \ variation}{total\ \ variation}$$

The R^2 value is called the **coefficient of determination** and indicates the per cent of the variance in the outcome variable that can be explained or accounted for by the explanatory variable(s). Hence, it is a measure of the **'goodness of fit'** of the regression line to the data. It ranges between 0 and 1 (it won't be negative). An R^2 statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response.

This statistic is part of the R output labeled "Multiple R-squared" and in our example takes the value 0.5085. It indicates that about 51% of the variation in infant's body weight can be explained by the variation of the infant's body height.

Note

In simple linear regression

$$\sqrt{0.5085} = 0.713$$

which equals to the Pearson's correlation coefficient, r.

ANOVA table

The idea of partitioning the variability in y into that accounted for by the model (explained) and that which is not (unexplained) should not be new to us, remember that

we had seen that when we first discussed analysis of variance (ANOVA):

```
anova(model_height)
```

```
## Analysis of Variance Table
##
## Response: weight
##            Df    Sum Sq   Mean Sq F value    Pr(>F)
## height      1 101118823 101118823  567.04 < 2.2e-16 ***
## Residuals 548  97723016    178327
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We have a pretty small p-value, meaning we would reject the null hypothesis and in this case rejecting the null hypothesis means that the data provided convincing evidence that the explanatory variable "height" is a significant explanatory variable of the response variable "weight".

From ANOVA table we can also calculate R^2. We have defined previously that R^2 is the proportion of variability explained by the model (Sum Sq for height) to total variability of the response variable (Sum Sq for height and residuals), so:

$$R^2 = 101118823/(101118823 + 97723016) = 101118823/198841839 = 0.5085$$

4.2 Simple linear regression with a binary explanatory variable

Using the same sample of 550 infants of 1 month age we want to examine how body weight is associated with the gender of the infant. Now we have an explanatory variable x that is **binary** (Male/Female), as opposed to the numerical explanatory variable model (height) that we used previously. A graphical comparison of the weight between the two groups is presented below (Figure 11):

```
ggbetweenstats(
   data = BirthWeight,
   x = gender,
   y = weight,
   xlab = "gender",
   ylab = "Weight (g)",
   bf.message = FALSE,
   messages = F,
   marginal = FALSE,
   results.subtitle = F)
```

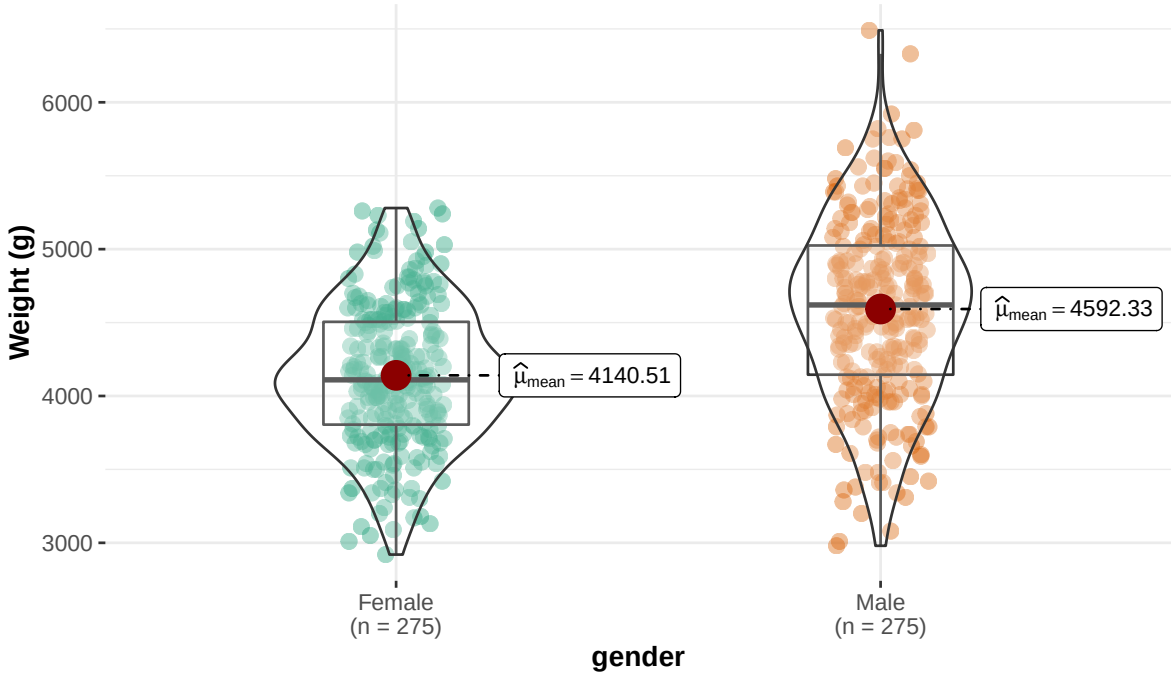


Figure 11: Comparison of the weight between the females and males.

How can we handle this variable in a mathematical equation?
Well, we will use a trick. All cases in which the respondent is `Male` will be coded as 1 and all other cases, in which the respondent is `Female`, will be coded as 0 (reference category). This allows us to enter in the `gender` values as numerical (remember, these numbers are just indicators).

$$genderMale = \begin{cases} 1 & \text{if infant is Male} \\ 0 & \text{otherwise (ref.)} \end{cases}$$

The equation of the regression line will have the following form:

$$\widehat{y} = b_0 + b_1 \cdot x$$
$$\widehat{y} = \widehat{\text{weight}} = b_0 + b_1 \cdot \ genderMale$$

Let's output the regression table for this model. Recall that this is done in two steps:

1. We first "fit" the linear regression model using the `lm()` function and save it in `model_gender`.
2. We get the regression table by applying the `summary()` and `confint()` function to `model_gender`.

```
# Fit regression model:
model_gender <- lm(weight ~ gender, data = BirthWeight)

summary(model_gender)
```

```
##
## Call:
## lm(formula = weight ~ gender, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1612.33  -371.87     3.58   379.49  1897.67
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4140.51      33.66 122.999   <2e-16 ***
## genderMale    451.82      47.61   9.491   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 558.2 on 548 degrees of freedom
## Multiple R-squared:  0.1412, Adjusted R-squared:  0.1396
## F-statistic: 90.07 on 1 and 548 DF,  p-value: < 2.2e-16
```

```
confint(model_gender, level=0.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 4074.3848 4206.633
## genderMale   358.3044  545.332
```

Let's apply the `get_regression_table()` function from the `{moderndive}` package and focus on the values in the `estimate` column (Table 5).

```
# Get regression table:
get_regression_table(model_gender)
```

Table 5: Linear regression table: gender

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4140.509	33.663	122.999	0	4074.385	4206.633
gender: Male	451.818	47.607	9.491	0	358.304	545.332

1. `intercept` corresponds to the mean weight 4140.509 g of a female infant which is the reference category.
2. `genderMale` corresponds to male infants and the value 451.818 is the mean difference in weight for a male infant relative to a female infant.

Therefore, the mean weight of a male infant is (4140 + 452) 4592 g which is significantly higher (on average) about 452 g relative to a female infant (p<0.001). The 95% confidence interval for this estimation (the difference in means) is 358 to 545 g.

Note that the model without the intercept term gives the mean for each group (Table 6):

```
# Fit regression model:
model_gender2 <- lm(weight ~ 0 + gender, data = BirthWeight)
```

Table 6: Linear regression table without intercept: gender

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
genderFemale	4140.509	33.663	122.999	0	4074.385	4206.633
gender: Male	4592.327	33.663	136.421	0	4526.203	4658.452

The above analysis is equivalent to run a two-sample t-test.

Let's perform the two-sample t-test.

```
BirthWeight %>%
  mutate(gender = fct_relevel(gender, "Male")) %>%  # for calculating Male - Female
  t_test(weight ~ gender, var.equal = T, detailed = T)
```

Table 7: An equivalent two-sample t-test

estimate	estimate1	estimate2	.y.	group1	group2	statistic	p	conf.low	conf.high
451.818	4592.327	4140.509	weight	Male	Female	9.491	0	358.304	545.332

The calculated difference between weight means equals to 452 g (4592 - 4140) which is the coefficient b_1 of the regression model_gender (Table 7). The value of the t-test (9.491) is the same as the t-test for b_1 in regression analysis. In addition, note that the 95% confidence interval of the difference in weight means (358 to 545) is the same as the confidence interval for b_1 in regression analysis.

4.3 Simple linear regression with a categorical explanatory variable (> 2 categories)

Suppose that infants are categorized into three categories based on `parity`: singletons, having one sibling or having 2 or more siblings. We choose as the reference category the singleton infants (we have already reorder the factor levels). A graphical investigation is presented below (Figure 12):

```
ggbetweenstats(
   data = BirthWeight,
   x = parity,
   y = weight,
   xlab = "parity",
   ylab = "Weight (g)",
   bf.message = FALSE,
   messages = F,
   marginal = FALSE,
   results.subtitle = F)
```

How will we form the equation of the regression line?

Well, we will use the previous trick and we will create 2 `dummy variables` to assign numerical values to the levels of parity. So each dummy variable will represent one category of the explanatory variable and will be coded with 1 if the case falls in that category and with 0 if not.

$$parityOne\ sibling = \begin{cases} 1 & \text{if infant has one sibling} \\ 0 & \text{otherwise} \end{cases}$$

$$parity\ \geqslant 2\text{siblings} = \begin{cases} 1 & \text{if infant has 2 or more siblings} \\ 0 & \text{otherwise} \end{cases}$$

$$\widehat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

$$\widehat{y} = \widehat{\text{weight}} = b_0 + b_1 \cdot\ parityOne\ sibling + b_2 \cdot\ parity \geqslant 2\ \text{siblings}$$

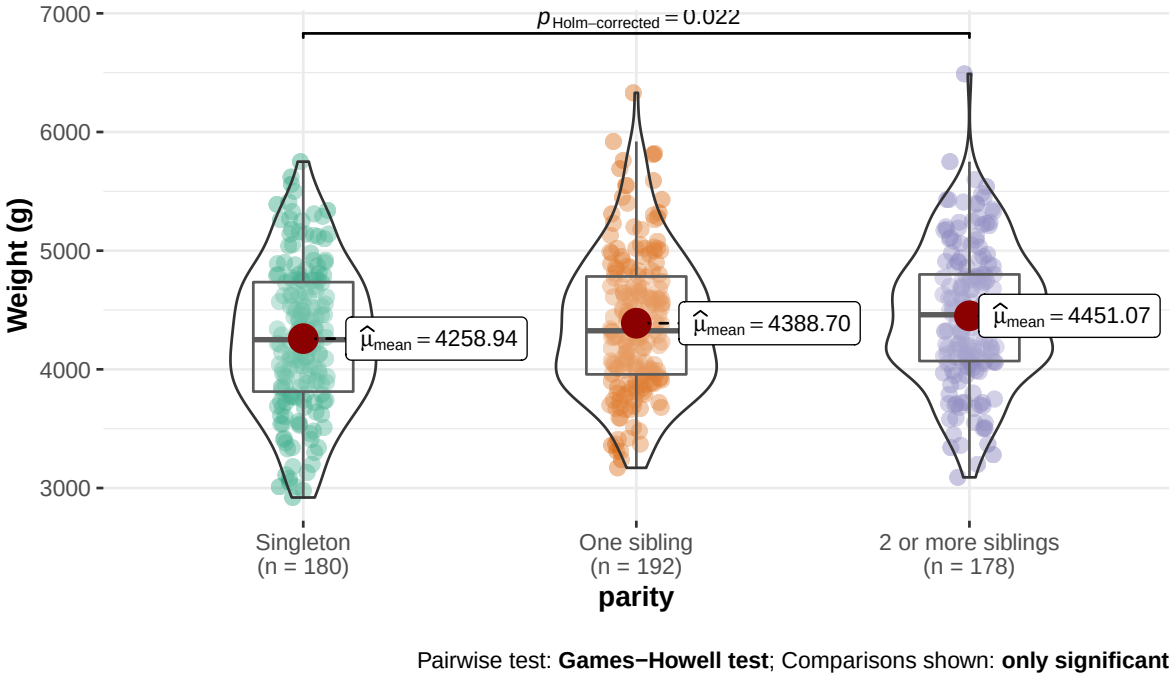


Figure 12: Comparison of the weight between the categories of the parity variable.

Therefore, we are including all the categories to the linear regression model except the one which is going to be used as the reference category (here is the `Singleton`). Actually, we create a multiple regression model which we will examine later analytically.

In general, a categorical explanatory variable with k-levels or categories requires (k-1) dummy variables to represent it. The explanatory variable here has three categories so we need to create two dummy variables (3-1=2) (Figure 13).

parity	One sibling	2 or more siblings
Singleton (ref.)	0	0
One sibling	1	0
2 or more siblings	0	1

Figure 13: Binary coding for dummy variables and reference category

Let's output the regression table for this model in two steps:

1. We first "fit" the linear regression model using the `lm()` function and save it in

model_parity.

2. We get the regression table by applying the `summary()` and `confint()` function to `model_parity`.

```
# Fit regression model:
model_parity <- lm(weight ~ parity, data = BirthWeight)

summary(model_parity)
```

```
##
## Call:
## lm(formula = weight ~ parity, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1361.07  -408.88   -23.82   411.24  2038.93
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4258.94      44.55  95.605  < 2e-16 ***
## parityOne sibling           129.75      62.01   2.093  0.03685 *
## parity2 or more siblings    192.12      63.18   3.041  0.00247 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 597.7 on 547 degrees of freedom
## Multiple R-squared:  0.01735,    Adjusted R-squared:  0.01376
## F-statistic: 4.829 on 2 and 547 DF,  p-value: 0.008339
```

```
confint(model_parity, level=0.95)
```

```
##                                2.5 %    97.5 %
## (Intercept)              4171.439324 4346.4496
```

```
## parityOne sibling           7.951467  251.5554
## parity2 or more siblings   68.024861  316.2210
```

Let's apply the `get_regression_table()` function from the `{moderndive}` package and focus on the values in the `estimate` column (Table 8).

```
# Get regression table:
get_regression_table(model_parity)
```

Table 8: Linear regression table: parity

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4258.944	44.548	95.605	0.000	4171.439	4346.450
parity: One sibling	129.753	62.008	2.093	0.037	7.951	251.555
parity: 2 or more siblings	192.123	63.176	3.041	0.002	68.025	316.221

1. `intercept` corresponds to the mean weight 4258.944 g for a singleton infant which is the reference category.

2. `parityOne sibling` corresponds to an infant with one sibling. The value 129.753 is the mean difference in weight for an infant with one sibling relative to a singleton infant.

Therefore, the mean weight of an infant with one sibling is 4389 g which is significantly higher (on average) about 130 g relative to a singleton infant (p=0.037<0.05). The 95% confidence interval for this estimation (the difference in means) is 8 to 252 g.

3. `parity2 or more siblings` corresponds to an infant with 2 or more siblings. The value 192.123 is the mean difference in weight for an infant with 2 or more siblings relative to a singleton infant.

Therefore, the mean weight of an infant with 2 or more siblings is 4451 g which is significantly higher (on average) about 192 g relative to a singleton infant (p=0.002<0.05). The 95% confidence interval for this estimation (the difference in means) is 68 to 316 g.

Note that the model without the intercept term gives the mean for each group (Table 9):

```
# Fit regression model:
model_parity2 <- lm(weight ~ 0 + parity, data = BirthWeight)
```

Table 9: Linear regression table without intercept: parity

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
paritySingleton	4258.944	44.548	95.605	0	4171.439	4346.450
parity: One sibling	4388.698	43.133	101.748	0	4303.971	4473.424
parity: 2 or more siblings	4451.067	44.797	99.361	0	4363.072	4539.063

ANOVA Table

We can obtain the ANOVA table:

```
anova(model_parity)
```

```
## Analysis of Variance Table
##
## Response: weight
##             Df    Sum Sq Mean Sq F value   Pr(>F)
## parity       2   3449872 1724936   4.829 0.008339 **
## Residuals 547 195391968  357207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since our p-value is less than 0.05, we say that the model (with the two dummy variables) as a whole is significant. We reject the null hypothesis, and the alternative hypothesis is suggesting that there is at least something interesting to look for here.

4.4 Similarly for variables `headc` and `education`

For the continuous variable `headc`:

```
# Fit regression model:
model_headc  <- lm(weight ~ headc, data = BirthWeight)

summary(model_headc)
```

```
##
## Call:
## lm(formula = weight ~ headc, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1420.63  -318.62    -4.02   287.47  1819.63
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -6059.87     560.36  -10.81   <2e-16 ***
## headc         275.13      14.78   18.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 471.4 on 548 degrees of freedom
## Multiple R-squared:  0.3875, Adjusted R-squared:  0.3863
## F-statistic: 346.6 on 1 and 548 DF,  p-value: < 2.2e-16
```

```
confint(model_headc, level=0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) -7160.5907 -4959.142
## headc         246.1064   304.162
```

```r
# Get regression table:
get_regression_table(model_headc)
```

Table 10: Linear regression table: headc

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-6059.866	560.364	-10.814	0	-7160.591	-4959.142
headc	275.134	14.778	18.618	0	246.106	304.162

The `headc` is a significant explanatory variable for the weight of the infants (Table 10).

For the categorical variable `education` (reference category `year10`):

```r
# Fit regression model:
model_education  <- lm(weight ~ education, data = BirthWeight)

summary(model_education)
```

```
##
## Call:
## lm(formula = weight ~ education, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1439.57  -422.09   -31.92   407.93  2079.09
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4352.929     42.821 101.653   <2e-16 ***
## educationyear12     57.980     74.169   0.782    0.435
## educationtertiary    6.636     57.173   0.116    0.908
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 602.6 on 547 degrees of freedom
## Multiple R-squared:  0.001226,   Adjusted R-squared:  -0.002425
## F-statistic: 0.3359 on 2 and 547 DF,  p-value: 0.7149
```

```
confint(model_education, level=0.95)
```

```
##                          2.5 %    97.5 %
## (Intercept)         4268.81473 4437.0439
## educationyear12      -87.71091  203.6705
## educationtertiary   -105.66914  118.9409
```

```
# Get regression table:
get_regression_table(model_education)
```

Table 11: Linear regression table: education

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4352.929	42.821	101.653	0.000	4268.815	4437.044
education: year12	57.980	74.169	0.782	0.435	-87.711	203.671
education: tertiary	6.636	57.173	0.116	0.908	-105.669	118.941

We can see that the education of the mother is not a significant explanatory variable for the weight (Table 11).

5 Multiple linear regression

The concepts and techniques discussed here are useful when the researcher wishes to consider simultaneously the associations among more than two explanatory variables. Although the concepts, computations, and interpretations associated with analysis of multiple-variable data may seem complex, they are natural extensions of material already explored previously.

We can write the model as:

$$\widehat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + ... + b_p \cdot x_p$$

The goal is to obtain coefficient estimates, which are also known as partial regression slopes, $b_1, b_2, b_3..b_p$ such that the linear model fits the available data well. In other words, we want to find estimators of these parameters such that the resulting line is as close as possible to the *n* data points.

Note that the model called "linear" because it is linear in the b's not necessarily in the x's. For example, the following model is considering a general linear model:

$$\widehat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2 + b_4 \cdot x_1^2$$

In Figure 14 we present a model consisted of one response variable (`weight`) and two continuous explanatory variables (`height`, `headc`)

$$\widehat{weight} = b_0 + b_1 \cdot height + b_2 \cdot headc$$

We have visualized some of the points as being located above the plane and some as being located below the plane. The deviation of a point from the plane is represented by the dashed red line and is called **residual**. When the model contains more than two independent variables, it is described geometrically as a hyperplane.

Note The residuals in linear regression are assumed to be independent and identically distributed following the normal distribution with mean equals to zero and constant variance.

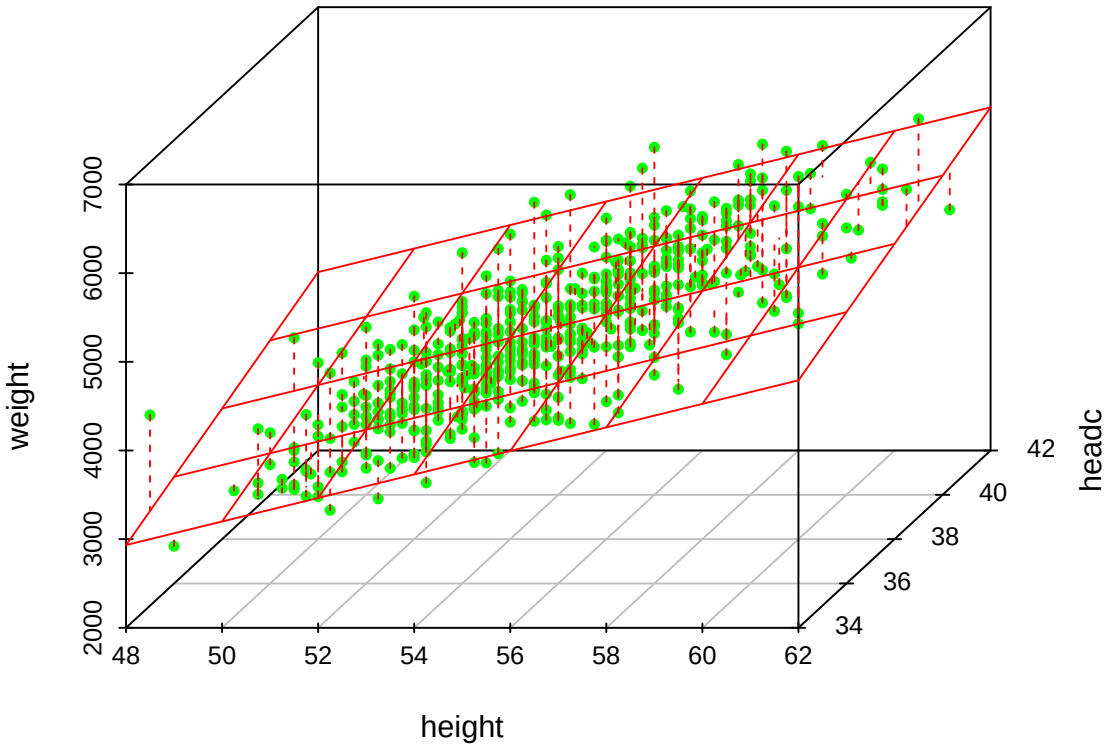


Figure 14: For a model consisted of one response variable and two explanatory variables a plane in three-dimensional space may be fitted to the data points.

5.1 Sample size calculation

Sample size calculations is a fundamental part of designing a study to develop a new regression model. The sample size should be **pre-specified in the protocol** of the study taking into account the candidate variables.

Many researchers suggested that 10 subjects per variable (SPV) is the minimum required sample size for linear regression models to ensure accurate prediction in subsequent subjects. Other researchers described a rule, that specifies a minimum of 200 subjects for any regression analysis. In a recent paper (Riley et al. 2019) was proposed how to ascertain the minimum sample size needed to develop a prediction model using linear regression. The authors also provided the `pmsampsize` package in R for the calculation. The authors described in their technical article four criteria that the sample size should meet:

i) small overfitting defined by an expected shrinkage of predictor effects by 10% or less

ii) small absolute difference of 0.05 in the model's apparent and adjusted R-squared value

iii) precise estimation of the residual standard deviation

iv) precise estimation of the average outcome value.

Note The average outcome and the standard deviation values in the population of interest are often informed by the literature base or prior research, or, simply by knowledge of the researcher about the field in which he or she works.

Note Researchers should be conservative with their chosen R-squared value; for example, by taking the R-squared value from a previous model, even if they hope their new model will improve performance.

Let's apply them in our working example. Initially we had 5 candidate variables

(height, gender, parity, headc, and education). Parity and education have dummy variables so we need to calculate the sample size with a total of seven parameters:

```
pmsampsize(type = "c",          # continuous outcome
           rsquared = 0.4,      # variance in outcome explained by the model
           parameters = 7,      # the number of candidate predictor parameter
           shrinkage = 0.9,     # measure of overfitting (range from 0 to 1)
           intercept = 4000,    # the average outcome value in the population
           sd = 700,            # the st. deviation of outcome in the population
           mmoe = 1.1)          # multiplicative margin of error for intercept (10%)
```

```
## NB: Assuming 0.05 acceptable difference in apparent & adjusted R-squared
## NB: Assuming MMOE <= 1.1 in estimation of intercept & residual standard deviation
## SPP - Subjects per Predictor Parameter
##
##              Samp_size Shrinkage Parameter Rsq   SPP
## Criteria 1          84     0.901         7 0.4 12.00
## Criteria 2          85     0.902         7 0.4 12.14
## Criteria 3         241     0.962         7 0.4 34.43
## Criteria 4*        241     0.962         7 0.4 34.43
## Final              241     0.962         7 0.4 34.43
##
##  Minimum sample size required for new model development based on user inputs = 241
##
##  * 95% CI for intercept = (3997.4, 4002.6), for sample size n = 241
```

Therefore, according to the above estimation we need at least 241 infants for running our regression model with seven paremeters.

5.2 Basic Criteria for Model Selection

When we are conducting a simple linear regression, there is, of course, only one possibility for the model, since there is only a single explanatory variable. However, when

we are conducting multiple linear regression, thereby having more than a single explanatory variable, the question then becomes that of choosing or selecting the best model. The problem is how to best define best! Which model shall we adopt? Should we adopt a model with very few explanatory variables, or should we adopt a model with many more explanatory variables? For instance, if we have 20 explanatory variables available, should we use all of them, or only a subset of them in our quest to define the best model? How should that subset be chosen as to optimize some function of the data?

Surprising to most novices in the area, these questions are not easy for even the most experienced of researchers or data analysts. Selecting variables in regression models is a complicated problem, and there are many conflicting views on which type of variable selection procedure is best. Indeed, as we will see, how a "best" model is arrived at will often depend on **substantive** and **scientific issues** as it will on **statistical criteria**.

Generally, a guiding principle in model selection is that simplicity rules the day. That is, given a choice between a more complex model and a simpler one, all else equal, the simpler one is usually preferable to the more complex. This principle often goes by the name of Occam's razor (law of parsimony) and is applicable not only to statistical models, but to virtually all narrative explanations.

Parcimonious model: we typically prefer models that are simple if they account for similar amounts of variance as do more complex ones.

Next we present some commonly used strategies for model selection:

1. **Simultaneous Regression**

The first and most obvious way to select a model is to simply include all available explanatory variables into the model and assess model fit based on this complete model. We may call this approach **full entry** or **simultaneous** regression. In this approach, we build the regression model by simultaneously estimating all parameters at the same time. However, there are instances where full-entry or simultaneous regression may not be considered the best option for model-building, and where the researcher may

be more interested in adopting a more complex algorithm to building his or her regression model.

2. **Hierarchical Regression**

In hierarchical regression, in contrast to simultaneous regression where all explanatory variables are entered into the model at the same time, a researcher usually has a designated prespecified order in which he or she would like to enter explanatory variables into the model. This order of entry is usually theoretically driven, based on the prior knowledge base of the researcher. The technique of hierarchical regression is popular among social scientists in testing mediational hypotheses.

3. **Automated regression methods (Best subset selection)**

These methods combine the explanatory variables in all possible ways. The best subset selection (using backward elimination, forward selection, or both[stepwise selection]) seek to find the best model according to statistical criteria in many steps (stepwise method, the model is re-avaluated in each step). However, as you might imagine, the number of possible models quickly becomes quite large.

Note There are several statistical criteria and indices that can be used to assess the efficiency or fit of a model that are penalized by the number of explonatory variables. These criteria are calculated and compared for a set of competing models thereby providing an objective basis on which to select the 'best' regression model (e.g., adjusted R square, Akaike Information Criterion, AIC: the smaller the value of AIC the better the model, Bayesian Information Criteria, BIC).

At first glance, especially to the newcomer in statistics, automated selection methods seem like a great idea and the best way of proceeding with almost any multiple regression problem. They at first can appear as the "panacea" to the modeling problem. After all, it would appear a great idea to just leave model selection to the computer so that it "figures out," in all of its complex computing capacities, the most "correct" model. Such an algorithmic approach must be the best solution, right? However, the situation is not as clear-cut as this, unfortunately, and there are many issues and problems,

some statistical, others substantive, that surround selection approaches. Automation does have its downside, and not everything can be left to a computer to decide.

Statistically, automated selection methods based on algorithmic approaches such as backward and forward regression have been shown to bias parameter estimates, and essentially, make the resulting inferential model suspect. After repeated testing the probability of type I error is far greater than the nominal α (usually 0.05).

Aside from statistical, there are also, and perhaps more importantly, substantive cautions that must be exercised when considering model selection. The chosen model at the last step of automated regression may not be one that has maximum utility.

Maximizing statistical criteria is not the same as maximizing utility of the model.

4. **Purposeful selection process**

The purposeful selection process begins by a univariable analysis of each candidate variable. Then, a general decision rule to include a candidate variable is applied. For example:

> Any variable with a **p < 0.20** in univariable analysis is selected for the multivariable analysis.

According to this rule, in our example, education of the mother is the only variable that will not be selected for the mutivariable analysis because in the univariable analysis we found that p > 0.2 .

However, given that the goal of the multivariable model is usually to assess the effect of the study intervention, while controlling for putative confounding variables, variable selection should take into account the existing knowledge and clinical important of the variables (if this is the case, we can break the previous rule). For a example, a potential confounder could be included in the model if it changes the coefficient of the primary exposure variable by 10 percent in the multivariable model.

5.3 Final model

Our multivariable final model will be based on purposeful selection process. Therefore, it will include the following explanatory variables: height, headc, gender, and parity. We excluded the education variable because in the univariable analysis we found that p >0.2).

```
final_model <- lm(weight ~ height + headc + gender + parity, data = BirthWeight)
summary(final_model)
```

```
##
## Call:
## lm(formula = weight ~ height + headc + gender + parity, data = BirthWeight)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1326.02  -247.49     9.35   250.91  1239.58
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -7072.398    499.623 -14.155  < 2e-16 ***
## height                      129.752      8.535  15.201  < 2e-16 ***
## headc                       109.831     15.582   7.049 5.50e-12 ***
## genderMale                  196.897     34.916   5.639 2.75e-08 ***
## parityOne sibling            82.078     40.096   2.047   0.0411 *
## parity2 or more siblings    104.874     41.020   2.557   0.0108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 384.3 on 544 degrees of freedom
## Multiple R-squared:  0.596,  Adjusted R-squared:  0.5922
## F-statistic: 160.5 on 5 and 544 DF,  p-value: < 2.2e-16
```

```
confint(final_model, level=0.95)
```

```
##                                 2.5 %     97.5 %
## (Intercept)              -8053.823624 -6090.9717
## height                     112.985121   146.5181
## headc                       79.222608   140.4395
## genderMale                 128.310973   265.4826
## parityOne sibling            3.315166   160.8400
## parity2 or more siblings    24.296956   185.4515
```

```
# or applying the summ() function from the jtools package
summ(final_model, confint = TRUE, vifs = TRUE, digits = 3)
```

Observations	550
Dependent variable	weight
Type	OLS linear regression

F(5,544)	160.478
R^2	0.596
Adj. R^2	0.592

	Est.	2.5%	97.5%	t val.	p	VIF
(Intercept)	-7072.398	-8053.824	-6090.972	-14.155	0.000	NA
height	129.752	112.985	146.518	15.201	0.000	1.569
headc	109.831	79.223	140.439	7.049	0.000	1.673
genderMale	196.897	128.311	265.483	5.639	0.000	1.135
parityOne sibling	82.078	3.315	160.840	2.047	0.041	1.023
parity2 or more siblings	104.874	24.297	185.451	2.557	0.011	1.023

Standard errors: OLS

Figure 15: Results of multiple regression table

```
# forest plot
jtools::plot_summs(final_model, scale = F, color.class = "darkgreen") +
  theme_classic2(base_size = 11)
```

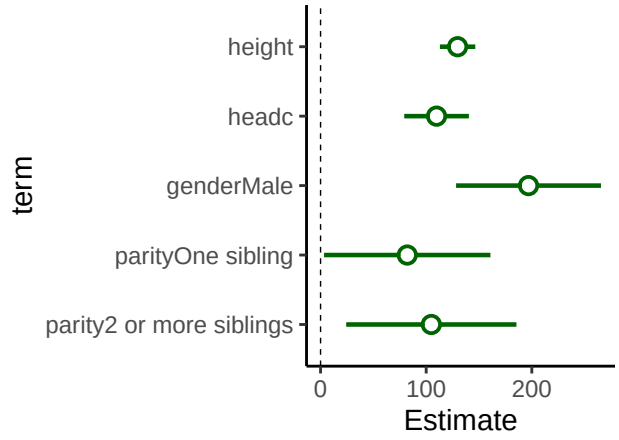


Figure 16: Forest plot with the coefficients of the linear model.

The Interpretation of the coefficients is similar with the univariable analysis but here the result is **adjusted or controlling for the other explanatory variables**. For example, for every 1 cm increase in infant's height, we expect, on average, the body weight to increase significantly about 130 g **adjusted (or controlling) for the gender, parity and head circumference**.

Note that the multiple R-squared value is the squared correlation between observed and predicted values of y. To demonstrate its computation, let's generate this number manually in R, where ^2 represents squaring:

```
cor(BirthWeight$weight, fitted(final_model))^2
```

```
## [1] 0.5959569
```

However, we are interested in the adjusted R^2 value that penalizes the multiple R-squared value somewhat in the sense of potentially fitting more parameters than absolutely necessary (Figure 15). The adjusted R^2 = 0.592 for the model indicates that

about 59.2% of the variation in infant's body weight can be explained by the variation of the explanatory variables.

Note The adjusted R-squared value is the R-squared value adjusted for the number of explanatory variables included in the model and can therefore be compared between models that include **different numbers** of explanatory variables.

Fisher global test (F-statistic)

The Fisher global test statistic (F-statistic) and the associated p-value are used to test the global joint contribution of all explanatory variables in the model to explain the variation of y. The hypotheses are:

$$H_0 : \beta_1 = \beta_2 = ...\beta_p = 0$$

vs H_A : at least one of the β_i coefficients is different from zero $\neq 0$.

The F-test, yielding a significant result (like in our example: F= 160.5, p <0.001) doesn't mean the model fits the data well. It just means that at least one of the estimates is non-zero.

The F-test, on the other hand, not yielding a significant result doesn't mean individual variables included in the model are not good predictors of y. It just means that the combination of these variables doesn't yield a good model.

Presentation of the results

The presentation of the results for the univariabe and multivariable analysis for the final model should be reported in one table as following (Table 12):

Table 12: Univariabe and multivariable analysis

Dependent: weight		Coefficient (univariable)	Coefficient (multivariable)
height	[48.0,62.0]	178.31 (163.60 to 193.02, p<0.001)	129.75 (112.99 to 146.52, p<0.001)
headc	[34.0,41.2]	275.13 (246.11 to 304.16, p<0.001)	109.83 (79.22 to 140.44, p<0.001)
gender	Female	-	-
	Male	451.82 (358.30 to 545.33, p<0.001)	196.90 (128.31 to 265.48, p<0.001)
parity	Singleton	-	-
	One sibling	129.75 (7.95 to 251.56, p=0.037)	82.08 (3.32 to 160.84, p=0.041)
	2 or more siblings	192.12 (68.02 to 316.22, p=0.002)	104.87 (24.30 to 185.45, p=0.011)

The results of the multivariable model, here, are all significantly important ($p < 0.05$). Generally, we report the results from our final multivariable model either the coefficients of the variables are significant or non-significant. In other words, **it is not necessary the final multivariable model to include only significant coefficients**.

6 Verifying Model Assumptions

There are certain assumptions that need to be met in order for the results of our hypothesis tests and confidence intervals we described previously to have valid meaning. These assumptions must be met for the assumed underlying mathematical and probability theory to hold true.

For inference for regression, there are four assumptions that need to be met to maximize the reliability of hypothesis tests. Note the first four letters of these assumptions are highlighted in bold in what follows: **LINE**. This can serve as a nice reminder of what to check for whenever you perform linear regression.

1. **L**inearity of relationship between variables
2. **I**ndependence of the residuals
3. **N**ormality of the residuals
4. **E**quality of variance of the residuals

Assumptions **L**, **N**, and **E** can be verified through what is known as a *residual analysis*. Assumption **I** can be verified through an understanding of how the data was collected.

6.1 Check Model Assumptions with statistical tests

- **I**ndependence of the residuals

We can perform a Durbin-Watson-Test to check for autocorrelated residuals (a p-value < 0.05 indicates autocorrelated residuals).

```
set.seed(126)
check_autocorrelation(final_model, nsim = 1000)
```

```
## OK: Residuals appear to be independent and not autocorrelated (p = 0.052).
```

- **N**ormality of the residuals

```
check_normality(final_model)
```

```
## Warning: Non-normality of residuals detected (p = 0.043).
```

The function performs a `shapiro.test` and checks the standardized residuals for normal distribution. Note that this formal test almost always yields significant results for the distribution of residuals for large samples and visual inspection (e.g., histogram, Q-Q plot) are preferable.

- **E**quality of variance of the residuals

The most common test for checking equality of variance of the residuals (homoskedasticity) is the Breush-Pagan test (a p-value < 0.05 indicates presence of heteroscedasticity).

The original version of Breusch-Pagan test:

```
lmtest::bptest(final_model, studentize = F)
```

```
##
##  Breusch-Pagan test
##
## data:  final_model
## BP = 11.302, df = 5, p-value = 0.04571
```

However, the **studentized** BP test (which is the default) is more robust than the original one:

```
lmtest::bptest(final_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  final_model
## BP = 8.7457, df = 5, p-value = 0.1196
```

Caution!

Variations or different versions of the previous statistical tests may have different results. Visual approaches of the residuals should also be employed.

6.2 Diagnostic plots

A residual analysis is used to verify conditions L, N, and E and can be performed using appropriate data visualizations. We will describe some built-in diagnostic plots in R for testing the assumptions underlying linear regression model (Figure 17).

```
par(mfrow = c(3, 2), mar = c(6.5, 6.5, 8.5, 6.5))
plot(final_model, 1:5, cex.main = 8, cex.lab = 3, cex.axis = 2)
```

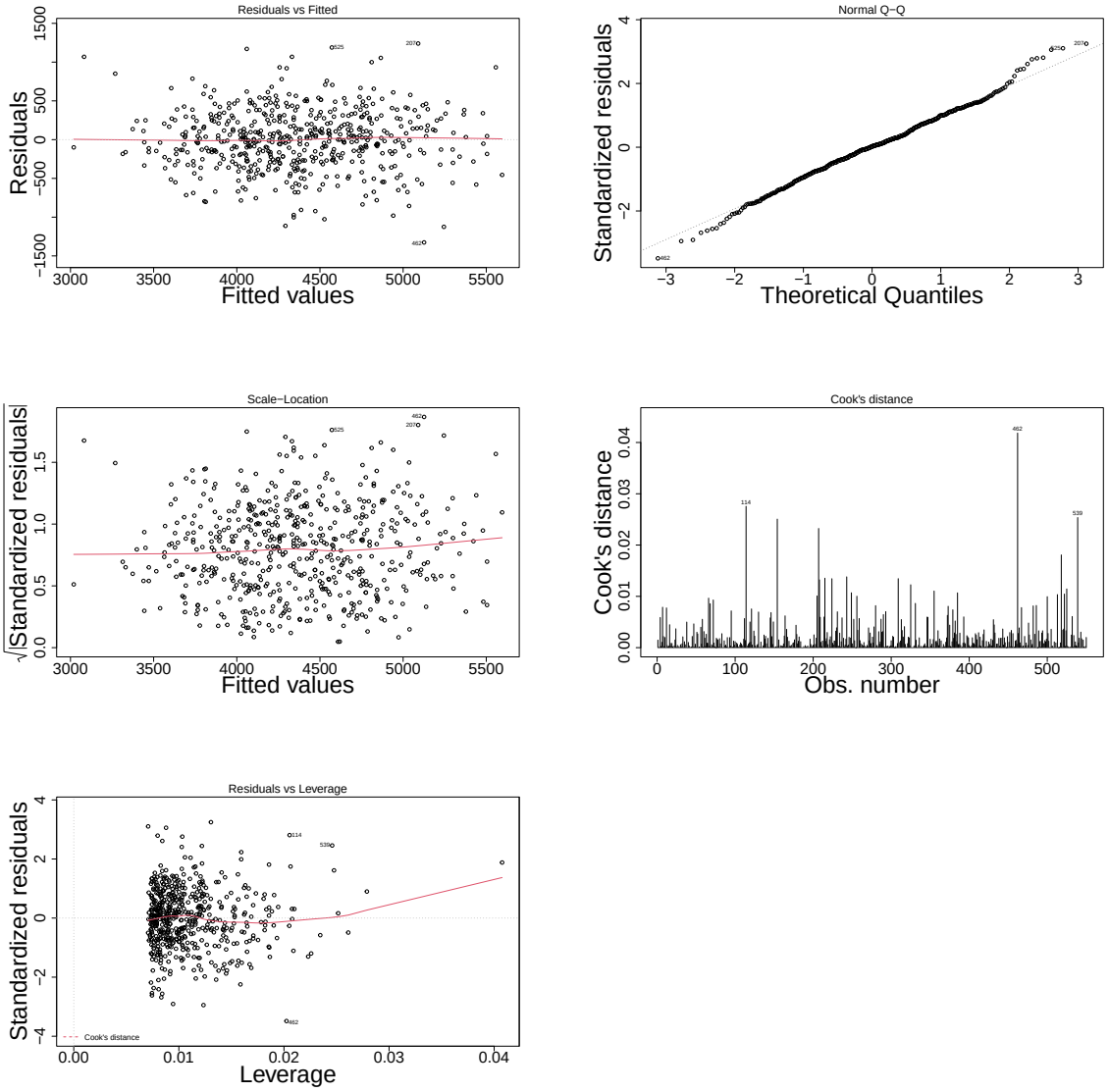


Figure 17: Diagnostic Plots.

The diagnostic plots show residuals in five different ways (Figure 17):

1. **Residuals vs Fitted**. Used to check the linear association assumption. The red line is just a scatterplot smoother, showing the average value of the residuals at each value of fitted value. **A horizontal red line, without distinct patterns is an indication for a linear association**.

2. **Normal Q-Q**. Used to examine whether the residuals are roughly normally distributed. We can standardize the residuals (mean zero and scale variance to 1) and then "percentile match" against a standard normal distribution. **It's good if the standardized residuals points follow the straight dashed line which is the perfectly percentile-matched line** . Of note, for large sample sizes, the normality assumption for the residuals is weaker (due to Central Limit Theorem).

3. **Scale-Location (or Spread-Location)**. Used to check the homogeneity of variance of the residuals (homoscedasticity). To verify the assumption, it suffices to plot the (square root of standardized) residuals against predicted or "fitted" values from the regression. In this case, the residuals are rescaled and all values are positive. **Horizontal line with equally spread points is a good indication of homoscedasticity**.

4. **Cook distance**. A metric to determine the influence of a value to the regression fit. Values ≥ 1 correspond to highly influential observations.

5. **(Standardized) Residuals vs Leverage**. Used to identify influential cases, that is extreme values that might influence the regression results when included or excluded from the analysis. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line.

As you can see from the Figure 15, the model assumptions appear to be well satisfied. There is no obvious "wedge" pattern evident in the residual plot (confirming that the assumption of homogeneity of variance is likely to be met). The Q-Q normal plot does not deviate greatly from normal. Finally, none of the points approach the high Cook's D contours suggesting that none of the observations are overly influential on the final fitted model.

Note The standardized residual is the residual divided by its standard deviation. The standardization (mean of zero and a variance of one) allows the residuals to be compared on the "standard scale": +/- 2 indicates something unusual, +/- 3 indicates something really out of the ordinary.

In summary, as the residuals are the differences between the observed and predicted values along a vertical plane, they provide a measure of how much of an outlier each point is in y-space (on y-axis). Outliers are identified by relatively large residual values. The patterns of residuals against predicted y values (residual plot) are also useful diagnostic tools for investigating linearity and homogeneity of variance assumptions (Figure 18).

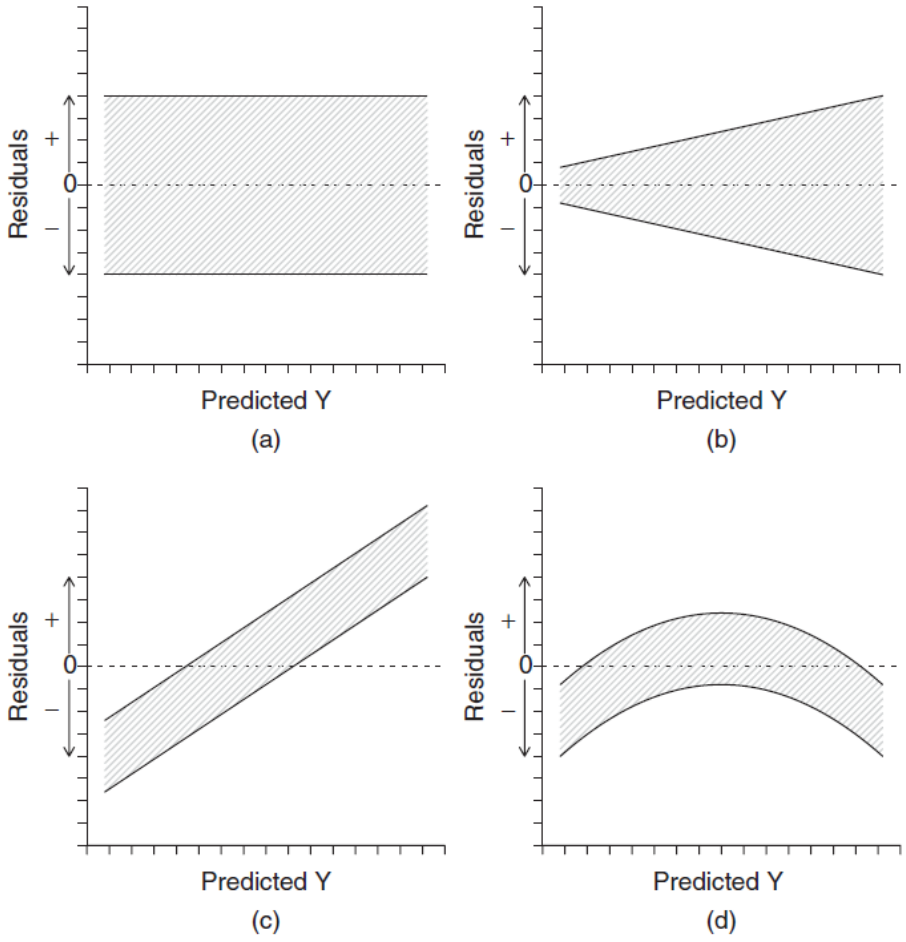


Figure 18: Stylised residual plots depicting characteristic patterns of residuals (a) random scatter of points - homogeneity of variance and linearity met, (b) 'wedge-shaped' - homogeneity of variance not met, (c) linear pattern remaining - erroneously calculated residuals or additional variable(s) required, and (d) curved pattern remaining - linear function applied to a curvilinear association.

We have only scratched the surface here with regard to diagnostics for regression. In most cases, simple checks as we did above will be adequate for most data to ensure

there are no serious violations. However, one can delve much further into residual analysis and obtain a whole slew of additional diagnostics, including partial residual plots that plot the association between a given explanatory variable and the response while considering all other explanatory variables in the regression model.

6.3 (Multi)collinearity

Though we expect **explanatory** variables in multiple regression to be associated to some degree, high association among them is not favorable.

Collinearity
It refers to the situation in which two or more **explanatory** variables are closely related to one another.

The problems with collinearity among explanatory variables are twofold:

1. Substantively, if one variable is highly collinear with another, then they are accounting for similar proportions of variance, and hence though they are not replications of one another, it suggests that one of the two variables may be sufficient for study.

2. Collinearity has major detrimental effects on model fitting: (a) instability of the estimated partial regression slopes (small changes in the data or variable inclusion can cause dramatic changes in parameter estimates), and (b) inflated standard errors and confidence intervals of model parameters. The importance of a variable can be masked due to the presence of collinearity.

A simple way to detect collinearity is to look at the correlation matrix of the explanatory variables. An element of this matrix that is large in absolute value indicates a pair of highly correlated variables, and therefore a collinearity problem in the data. Unfortunately, not all collinearity problems can be detected by inspection of the correlation matrix: it is possible for collinearity to exist between three or more variables

even if no pair of variables has a particularly high correlation. We call this situation **multicollinearity**.

Multicollinearity

It occurs when two or more **explanatory** variables are highly associated, conditional on the other explanatory variables in the model.

Multicollinearity can be diagnosed with the **variance inflation factor, or VIF for short,** which is computed for each explanatory variable entered into the model.

Variance Inflation Factor (VIF): the extent the variance of the estimated coefficients are inflated compared to the variance when the explanatory variables are NOT correlated.

Note

The smallest possible value for VIF is 1, which indicates the complete absence of collinearity. A VIF less than 5 indicates a low correlation of that explanatory variable with other variables (in practice no-multicollinearity). A value between 5 and 10 indicates a moderate correlation, while VIF values larger than 10 are a sign for high, not tolerable correlation of model explanatory variables.

As a rule of thumb, a VIF value that exceeds 5 indicates a problematic amount of collinearity [according to James, G., et al. (2013). An introduction to statistical learning: with applications in R. New York: Springer. page 101].**(Note: Other textbooks use a VIF cut off point of 4).**

The colomn `VIF` of the multiple linear table indicates that, in our example, the assumption is well-satisfied (VIF < 5) (Table 13).

Table 13: VIF of the multiple linear table

variables	VIF
(Intercept)	NA
height	1.568948
headc	1.673260
genderMale	1.135022
parityOne sibling	1.023174
parity2 or more siblings	1.023174

Some proposed remedies for multicollinearity from the literature

In a statistical sense, there is no way to "fix" multicollinearity. However, methods have been developed to mitigate its effects. Perhaps the most effective way to remedy multicollinearity is to make a priori judgements about the association between explanatory variables and remove or consolidate variables that have known correlations. This is not always possible however, especially when the true functional forms of associations are not known.

- Larger sample size: This will decrease standard errors and improve the precision.

- Model respecification: Remove some of the highly associated explanatory variables or replace them with a linear combination of them (if possible).

- Regularization (Tolerant) methods: Some regression techniques may be more sensitive to multicollinearity than others. Recent developments in model selection methods have introduced new methods for balancing model complexity and fit. For example two special linear regression model — Lasso and Ridge regression. Although not necessarily designed to be tolerant of collinearity, they offer approaches that may be less sensitive.

- Principal Component Regression

- The situation of multicollinearity may be averted by changing the reference category for categorical variables.

The solution lies not in more clever statistical methods, but in strategies such as larger sample sizes or experimental manipulation of the variables.

6.4 Modern Diagnostic plots using {performance} package

The model diagnostic plots that can also be easily generated using the `check_model()` function in the `{performance}` package. For each plot (Figure 19) there is an explanation title:

```
check_model(final_model, check = "all")
```

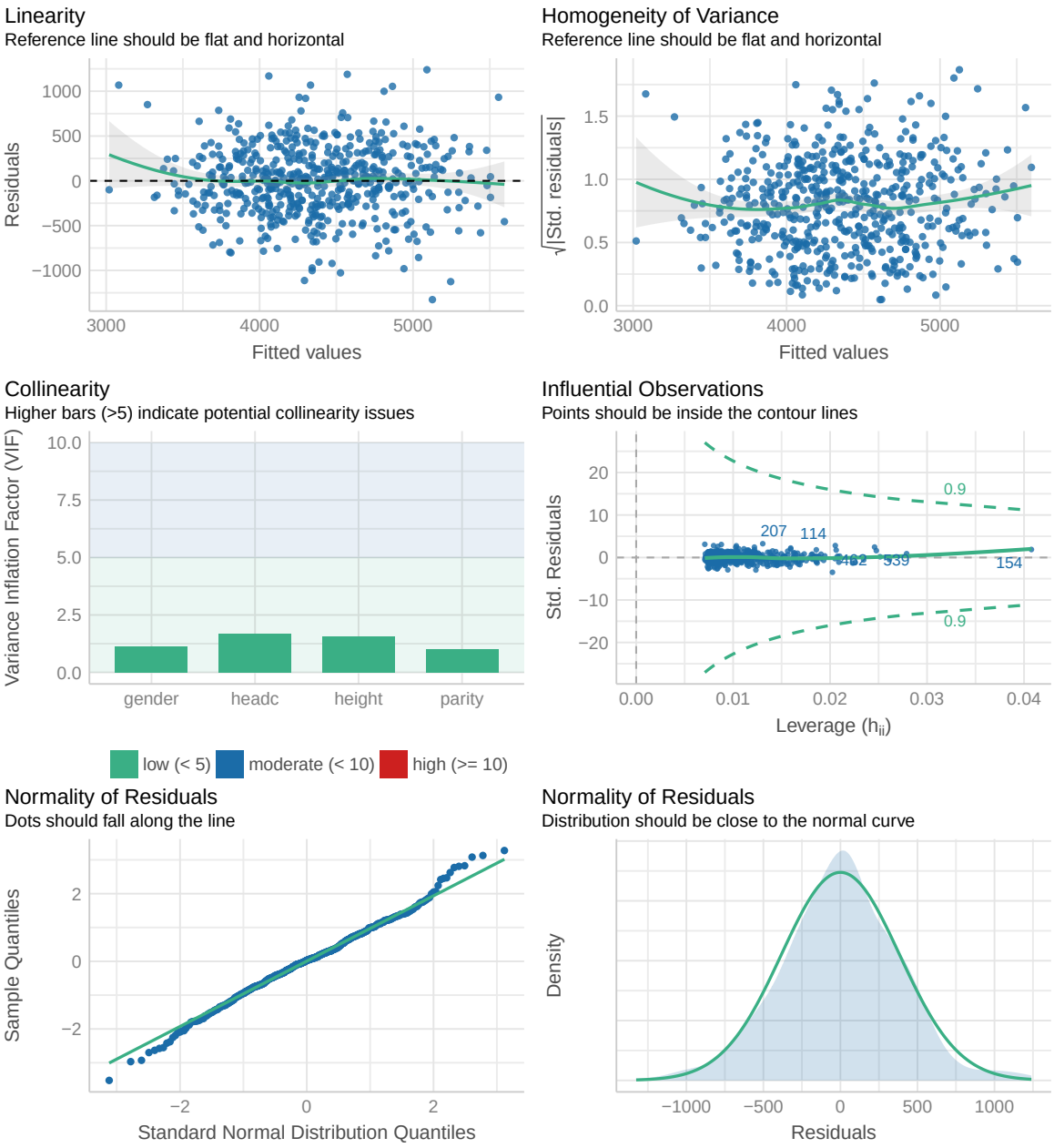


Figure 19: Modern diagnostic Plots

7 Partial Fisher Test for nested model

Fisher's partial test is used to test the contribution of a subset of explanatory variables (reduced model) in a model which already includes other explanatory variables (full model).

Nested models Two models are nested if one model contains all the terms of the other, and at least one additional term. The larger model is the complete (or full) model, and the smaller is the reduced (or restricted) model.

The hypotheses are:

H_0 : models do not signifficantly differ (not significant difference in RSS))
vs H_A : the full model is signifficantly better than the reduced model (significantly lower RSS)

In our example, consider a **reduced model** without `headc` variable (weight ~ height + gender + parity) obtained from our final model (weight ~ height + headc + gender + parity) which is consideed the full model:

```
reduced_model <- lm(weight ~ height + gender + parity, data = BirthWeight)
anova(reduced_model, final_model)
```

```
## Analysis of Variance Table
##
## Model 1: weight ~ height + gender + parity
## Model 2: weight ~ height + headc + gender + parity
##   Res.Df      RSS Df Sum of Sq      F    Pr(>F)
## 1    545 87677949
## 2    544 80340663  1   7337286 49.682 5.502e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Does including `headc` in the model significantly decrease the error and significantly increase the predictive power of the model? Does the `headc` variable significantly improve the model?

To determine if the full model is significantly better, we will check if the Residual Sum of Squares (RSS) also known as Sum of Squared Residuals (SSE) is significantly **lower** in the full model as compared to the reduced model. Partial F-test compares the RSS of the full and reduced models to see if there has been a significant change in RSS due to the removal of a term and hence a significant change in how well the model fits or predicts the observed data.

The RSS=80340663 for the full model is significantly **lower** than that of the reduced model in which RSS = 87677949. The p-value < 0.001 of the test indicates that `headc` gives extra information (better model) to predict infant's weight, when the other explanatory variables have already been taken into account.

Partial Fisher Test It tests a linear sub-hypothesis for nested models.

8 Stepwise models (AIC or BIC selection)

The base function step() is used for backward, forward and stepwise selection. These automated methods may be useful when the number of explanatory variables is large.

8.1 Backward elimination

Backward elimination starts with all candidate explanatory variables in the regression model. Then, at each step, it deletes the explanatory variable such that the resulting model has the lowest value of an information criterion, such as AIC. This process is continued until all variables have been deleted from the model or the information criterion increases. The results of this procedure for our model are the following:

```
model_back <- step(lm(weight ~ ., data = BirthWeight),
                   direction = "backward")
```

```
## Start:  AIC=6555.48
```

```
## weight ~ height + headc + gender + education + parity
##
##             Df Sum of Sq       RSS    AIC
## - education  2    152710  80340663 6552.5
## <none>                    80187954 6555.5
## - parity     2    856466  81044420 6557.3
## - gender     1   4658303  84846257 6584.5
## - headc      1   7266968  87454922 6601.2
## - height     1  34091381 114279335 6748.3
##
## Step:  AIC=6552.53
## weight ~ height + headc + gender + parity
##
##          Df Sum of Sq       RSS    AIC
## <none>                 80340663 6552.5
## - parity  2   1074173  81414836 6555.8
## - gender  1   4696517  85037180 6581.8
## - headc   1   7337286  87677949 6598.6
## - height  1  34127789 114468452 6745.2
```

Since the exclusion of `education` gives a lower AIC value (6552.5) than the model with all variables (AIC=6555.5) the final model will be: weight ~ length + headc + gender + parity and the selection process is completed in one step.

8.2 Forward selection

Forward selection starts with no potential explanatory variables in the regression equation. Then, at each step, it adds the explanatory variable such that the resulting model has the lowest value of AIC. This process is continued until all variables have been added to the model or the information criterion increases. The results of this procedure for our model is the following:

```r
# Forward selection
model_forward <- step(lm(weight ~ 1, data = BirthWeight),
                      direction = "forward",
                      scope = ~ height + headc + gender + parity + education)
```

```
## Start:  AIC=7040.96
## weight ~ 1
##
##             Df Sum of Sq       RSS    AIC
## + height     1 101118823  97723016 6652.3
## + headc      1  77043810 121798029 6773.4
## + gender     1  28069205 170772634 6959.3
## + parity     2   3449872 195391968 7035.3
## <none>                   198841839 7041.0
## + education  2    243873 198597967 7044.3
##
## Step:  AIC=6652.25
## weight ~ height
##
##             Df Sum of Sq      RSS    AIC
## + headc      1  11875897 85847119 6583.0
## + gender     1   8115868 89607148 6606.6
## + parity     2   1667481 96055535 6646.8
## <none>                   97723016 6652.3
## + education  2    689626 97033390 6652.4
##
## Step:  AIC=6582.99
## weight ~ height + headc
##
##             Df Sum of Sq      RSS    AIC
## + gender     1   4432283 81414836 6555.8
## + parity     2    809939 85037180 6581.8
## <none>                   85847119 6583.0
```

```
## + education  2    386386 85460733 6584.5
##
## Step:  AIC=6555.83
## weight ~ height + headc + gender
##
##             Df Sum of Sq      RSS    AIC
## + parity     2   1074173 80340663 6552.5
## <none>                   81414836 6555.8
## + education  2    370416 81044420 6557.3
##
## Step:  AIC=6552.53
## weight ~ height + headc + gender + parity
##
##             Df Sum of Sq      RSS    AIC
## <none>                   80340663 6552.5
## + education  2    152710 80187954 6555.5
```

Therefore since the model weight ~ height + headc + gender + parity has lower AIC (6552.53) than the model with all variables (AIC=6555.5) this is the preferred model. Here it took 4 steps to find the best model.

Note Backward elimination and forward selection method may not result in the same regression equation (in this example the two automatic methods give the same final model which is the model with all explanaory variables except the 'education' variable). Backward method is generally preferrable to forward method as it includes less steps.

8.3 Stepwise selection (AIC selection)

The stepwise method is the combination of backward and forward procedures. AIC is used as the marker for selecting best model. Lower AIC, the better model.

```r
model_both <- step(lm(weight ~ ., data = BirthWeight),
                   direction = "both")
```

```
## Start:  AIC=6555.48
## weight ~ height + headc + gender + education + parity
##
##             Df Sum of Sq       RSS    AIC
## - education  2    152710  80340663 6552.5
## <none>                    80187954 6555.5
## - parity     2    856466  81044420 6557.3
## - gender     1   4658303  84846257 6584.5
## - headc      1   7266968  87454922 6601.2
## - height     1  34091381 114279335 6748.3
##
## Step:  AIC=6552.53
## weight ~ height + headc + gender + parity
##
##             Df Sum of Sq       RSS    AIC
## <none>                    80340663 6552.5
## + education  2    152710  80187954 6555.5
## - parity     2   1074173  81414836 6555.8
## - gender     1   4696517  85037180 6581.8
## - headc      1   7337286  87677949 6598.6
## - height     1  34127789 114468452 6745.2
```

8.4 Stepwise selection (BIC selection)

The `step()` takes the argument k as 2 (default) or logn, where n is the sample size. With k = 2 it uses the AIC criterion and with k = log(n) it considers the BIC.

```r
model_stepBIC <- step(lm(weight ~ ., data = BirthWeight),
                   direction = "both", k = log(nrow(BirthWeight)))
```

```
## Start:  AIC=6589.96
## weight ~ height + headc + gender + education + parity
##
##             Df Sum of Sq       RSS    AIC
## - education  2    152710  80340663 6578.4
## - parity     2    856466  81044420 6583.2
## <none>                    80187954 6590.0
## - gender     1   4658303  84846257 6614.7
## - headc      1   7266968  87454922 6631.4
## - height     1  34091381 114279335 6778.5
##
## Step:  AIC=6578.39
## weight ~ height + headc + gender + parity
##
##             Df Sum of Sq       RSS    AIC
## - parity     2   1074173  81414836 6573.1
## <none>                    80340663 6578.4
## + education  2    152710  80187954 6590.0
## - gender     1   4696517  85037180 6603.3
## - headc      1   7337286  87677949 6620.1
## - height     1  34127789 114468452 6766.8
##
## Step:  AIC=6573.07
## weight ~ height + headc + gender
##
##             Df Sum of Sq       RSS    AIC
## <none>                    81414836 6573.1
## + parity     2   1074173  80340663 6578.4
## + education  2    370416  81044420 6583.2
## - gender     1   4432283  85847119 6595.9
## - headc      1   8192312  89607148 6619.5
## - height     1  33973742 115388578 6758.6
```

The best model based on BIC is: weight ~ height + headc + gender.

Note Although labelled AIC, since k was changed from its default value the step function is calculating the BIC.

9 Interaction Between Variables (optional reading)

Up to this point, each explanatory variable has been incorporated into the regression function through an additive term $b_i \cdot X_i$. Such a term is called a main effect. For a main effect, a variable changes the average response by b_i for each unit increase in X_i, controlling for the other explanatory variables.

Some of the most interesting research findings are those involving **interactions** among explanatory variables. Interaction is also referred to as **effect modification** or **moderation**, and must be distinguished from confounding.

We say that there is interaction between two explanatory variables X_1 and X_2 if the association between one of the variables and the response variable `Y` is not the same depending on the values of the other variable.

An interaction between two variables X_i and X_j is an additive term of the form $b_k \cdot X_i \cdot X_j$ in the regression function. For example, if there are two variables, the main effects and interactions give the following regression function:

$$\widehat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_1 \cdot x_2$$

In order to include interaction term to a model, the interaction term should satisfy the bellow:

- It should make sense conceptually.
- It should be statistically significant. This can be judged by the p-value

9.1 Interaction between a numeric variable and a binary variable

To make things clearer, suppose we wish to determine whether the effect of quantitative variable height is modified by the binary variable gender. We then consider the model:

$$\widehat{\text{weight}} = b_0 + b_1 \cdot height + b_2 \cdot \ genderMale + b_3 \cdot height \cdot \ genderMale$$

- For females, $genderMale = 0$, and the model is:

$$\widehat{\text{weight}} = b_0 + b_1 \cdot height$$

and the effect of height is measured by b_1 .

- For males, $genderMale = 1$, and the model is:

$$\widehat{\text{weight}} = b_0 + b_1 \cdot height + b_2 \cdot 1 + b_3 \cdot height \cdot 1$$

$$\widehat{\text{weight}} = (b_0 + b_2) + (b_1 + b_3) \cdot height$$

and the effect of height is measured by $b_1 + b_3$.

There is interaction between height and gender (or modification of the effect of height by gender) if the effect of height is different in females and males. We therefore need to perform a test for the interaction between the variables. If we reject H_0, we keep the interaction term in the model: there is modification of effect.

Let's see this example in practice. First, we will build the model with the interaction:

```
# model with interaction
 interact_model<-lm(weight ~ height + gender + height:gender, data=BirthWeight)
```

A versatile and sometimes the most interpretable method for understanding interaction effects is via plotting. The next plot (Figure 20) shows two different slopes for the data, illustrating the notion of interaction.

```
# Graphical investigation: interaction between height and gender (non-parallel lines)
interact_plot(interact_model, plot.points = T, pred = height, modx = gender) +
theme_bw()
```

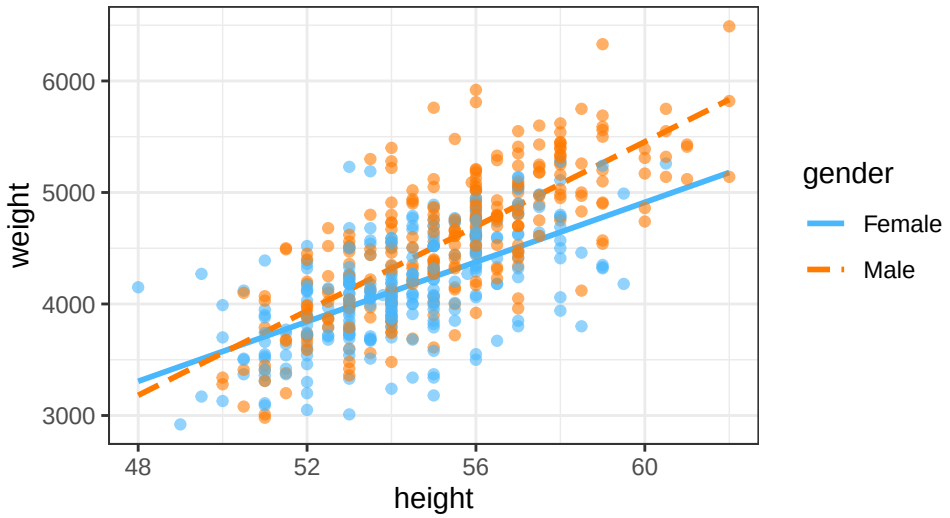


Figure 20: Graphical investigation: interaction between height and gender (non-parallel lines)

The slopes of the regression lines are not parallel which indicates that there is interaction between height and gender. In epidemiological terms, this is referred as there is an effect modification, the effect of height is modified by gender.

Now, we can test the **significance** of the interaction term by analysing the results of the model with the interaction term (Table 14):

```
# Get regression table:
get_regression_table(interact_model)
```

Table 14: Linear regression table: height + gender + interaction

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-3112.864	601.802	-5.173	0.000	-4294.995	-1930.733
height	133.744	11.088	12.062	0.000	111.965	155.524
gender: Male	-2800.039	810.025	-3.457	0.001	-4391.185	-1208.893
height:genderMale	55.714	14.777	3.770	0.000	26.687	84.741

The coefficient **55.71** is significantly different from zero (p-value < 0.001). We therefore conclude that the effect of height on infant weight at birth is depending on the gender of the infant (as we illustrated in Figure 20).

Therefore, the results on association between height and infant weight can be presented separately for the females and males.

- For females, $genderMale = 0$, and the model is:

$$\widehat{\text{weight}} = -3112.9 + 133.7 \cdot height$$

Note that the above equation can also be obtained by running the analysis in R only for the females (Table 15):

```
# filter the data to include only female infants
 BirthWeight_female <- BirthWeight %>%
   dplyr::filter(gender=="Female")

# model for females
model_females<-lm(weight ~ height, data=BirthWeight_female)

# Get regression table:
get_regression_table(model_females)
```

Table 15: Linear regression table for females

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-3112.864	577.045	-5.394	0	-4248.887	-1976.841
height	133.744	10.632	12.580	0	112.814	154.675

- For males, $genderMale = 1$, and the model is:

$$\widehat{\text{weight}} = -3112.9 + 133.7 \cdot height - 2800 \cdot 1 + 55.71 \cdot height \cdot 1 = -5912.9 + 189.4 \cdot height$$

Note that the above equation can also be obtained by running the analysis in R only for males (Table 16):

```
# filter the data to include only male infants
  BirthWeight_male <- BirthWeight %>%
   dplyr::filter(gender=="Male")

# model for males
model_males<-lm(weight ~ height, data=BirthWeight_male)

# Get regression table:
get_regression_table(model_males)
```

Table 16: Linear regression table for males

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-5912.903	563.617	-10.491	0	-7022.490	-4803.315
height	189.458	10.155	18.657	0	169.467	209.450

We present the two linear models in the following scatter plots (Figure 21):

```
ggscatter(BirthWeight, x = "height", y = "weight", digits = 3,
  color = "gender", palette = "jco", add = "reg.line") +
  facet_wrap(~gender) +
```

```
  stat_cor(label.y = 6000, digits = 2) +
  stat_regline_equation(label.y = 6200)
```

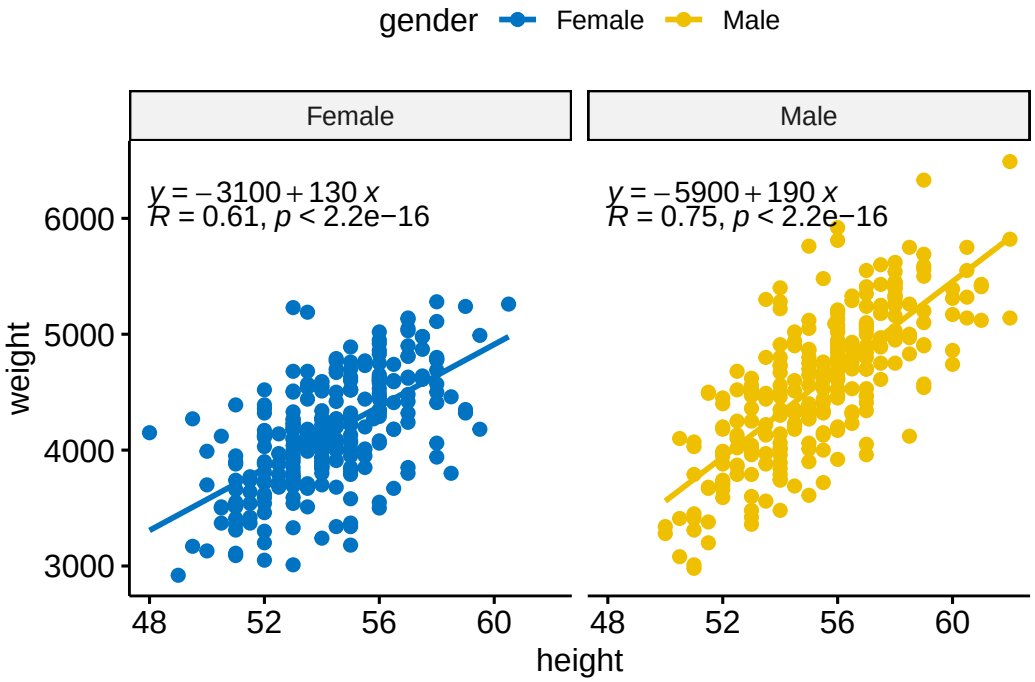


Figure 21: Correlation plots with the regression lines

9.2 Interaction between two numeric variables

Let's suppose we wish to determine whether there is an interaction between height and headc. We then consider the model:

```
# model with interaction
 interact_model2<-lm(weight ~ height + headc + height:headc, data=BirthWeight)

# Get regression table:
get_regression_table(interact_model2)
```

Table 17: Linear regression table: height + headc + interaction

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-12265.165	9804.246	-1.251	0.211	-31523.825	6993.494
height	210.460	179.560	1.172	0.242	-142.254	563.174
headc	246.564	258.165	0.955	0.340	-260.555	753.682
height:headc	-2.045	4.715	-0.434	0.665	-11.308	7.218

We can see in Table 17 that the interaction term is not significant (p-value = 0.665) and should not be included in the model.

We can plot the regression line between weight and height for 1 standard deviation above and below the mean and the mean itself of the headc (Figure 22).

```
interact_plot(interact_model2, plot.points = T, pred = height, modx = headc) +
theme_bw()
```

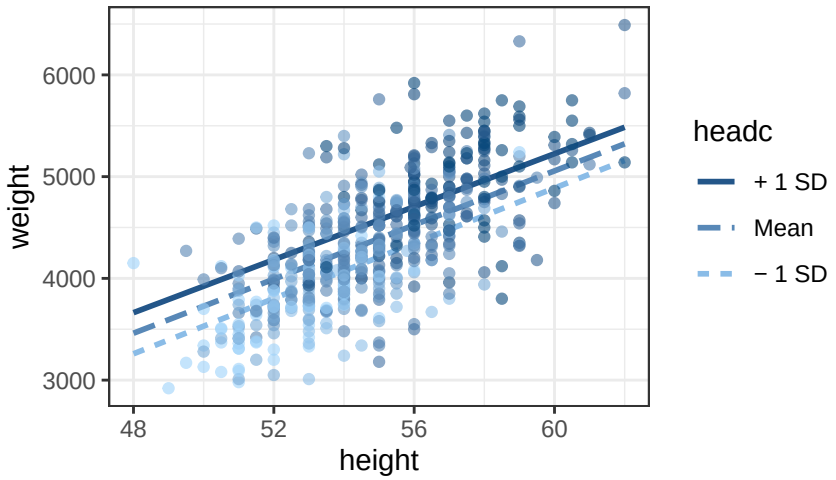


Figure 22: Plot for interaction.

or we can represent the regression line between weight and height for four equally spaced values of headc (Figure 23).

```
interact_plot(interact_model2, pred = height, plot.points = T, modx = headc,
              modx.values = c(35, 37, 39, 41)) +
theme_bw()
```

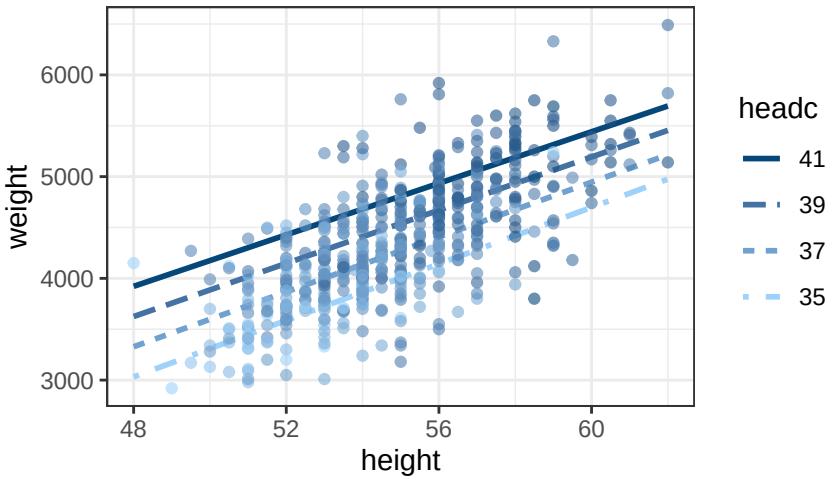


Figure 23: Plot for interaction.

As you can see, there is not much of an interaction. The slopes of the regression lines are parallel which indicates that there is not interaction between height and headc. The model should include only the main effects (parallel slopes model), as follows (Table 18):

```
# parallel slopes model
 parallel_model<-lm(weight ~ height + headc, data=BirthWeight)

# Get regression table:
get_regression_table(parallel_model)
```

Table 18: Linear regression table: height + headc

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-8018.683	488.341	-16.420	0	-8977.937	-7059.429
height	132.684	8.767	15.135	0	115.463	149.904
headc	134.808	15.497	8.699	0	104.367	165.250

The equation of the model is:

$$\widehat{\text{weight}} = -8018.7 + 132.7 \cdot height + 134.8 \cdot headc$$

9.3 Examples of common interactions in model development

In the fields of biostatistics and epidemiology, some types of interactions that have consistently been found to be important in model development (Figure 24):

Interaction	Effect
Severity of disease × treatment	Less benefit with less severe disease
Place × treatment	Benefit varies by treament centre
Place × predictors	Predictor effects vary by centre/region
Calender time × treatment	Learning curves for some treatments
Calender time × predictors	Increasing or decreasing impact of predictors over the years
Age × predictors	Older subjects less affected by risk factors; or more affected by certain types of disease
Follow-up time × predictors	Non-proportionality of survival effects, often a decreasing effect over time
Season × predictors	Seasonal effect of predictors

Figure 24: Examples of interactions to consider in clinical multivariable models

(Sources: Ewout W. Steyerberg, Clinical Prediction Models; Frank E. Harrell, Regression Modeling Strategies)