Analyzing data using linear models

Stéphanie M. van den Berg

Versie 0.1
(October 12, 2018)

Abstract

This book is intended to be of use to bachelor students in social, behavioural and management sciences that want to learn how to analyze their data, with the specific aim to answer research questions. The book has a practical take on data analysis: how to do it, how to interpret the results, and how to report the results. All techniques are presented within the framework of linear models: this includes simple and multiple regression models, to linear mixed models and generalized linear models. All methods can be carried out within one supermodel: the generalized linear mixed model. This approach is illustrated using SPSS.

Contents

Chapter 1

Moderation: testing interaction effects

1.1 Interaction with one numeric and one dichotomous variable

Suppose there is a linear relationship between age (in years) and vocabulary (the number of words one knows): the older you get, the more words you know. Suppose we have the following linear regression equation for this relationship:

$$\widehat{vocab} = 205 + 500 \times age \tag{1.1}$$

So according to this equation, the expected number of words for a newborn baby (age=0) equals 205. This may sound silly, but suppose this model is a very good model for vocabulary size in children between 2 and 5 years of age. Then this equation tells us that the expected increase in vocabulary size is 500 words per year.

This model is meant for everybody in the Netherlands. But suppose that one researcher expects that the increase in words is much faster in children from high SES families than in children from low SES families. First he believes that vocabulary will be larger in higher SES children than in low SES children. In other words, he expects an effect of SES, over and above the effect of age:

$$\widehat{vocab} = b_0 + b_1 \times age + b_2 \times SES \tag{1.2}$$

This *main effect* of SES is yet unknown and denoted by b_2. Note that this linear equation is an example of multiple regression.

Let's use some numerical example. Suppose age is coded in years, and SES is dummy coded, with a 1 for high SES and a 0 for low SES. Let b_2, the effect

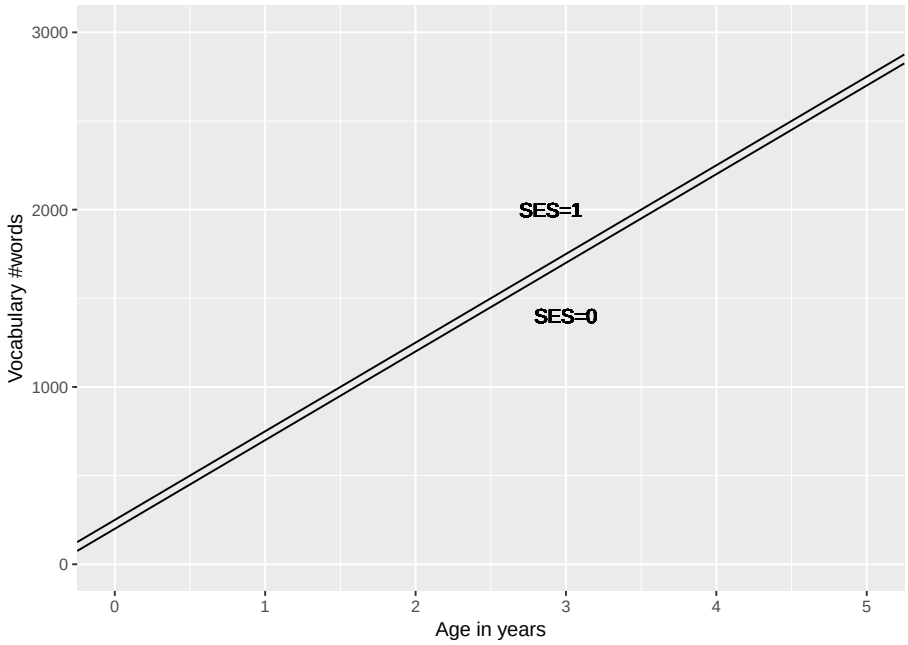

Figure 1.1: Two regression lines: one for low SES children and one for high SES children.

of SES over and above age, be 10. Then we can write out the linear equation for low SES and high SES separately.

$$\begin{aligned} lowSES: \widehat{vocab} &= 200 + 500 \times age + 10 \times 0 && (1.3)\\ &= 200 + 500 \times age && (1.4)\\ highSES: \widehat{vocab} &= 200 + 500 \times age + 10 \times 1 && (1.5)\\ &= (200 + 10) + 500 \times age && (1.6)\\ &= 210 + 500 \times age && (1.7) \end{aligned}$$

Figure 1.1 depicts the two regression lines for the high and low SES children separately. So we see that the effect of SES involves a change in the intercept: the intercept equals 200 for low SES children and the intercept for high SES children equals 210. The difference in intercept is indicated by the coefficient for SES. Note that the two regression lines are parallel: for every age, the difference between the two lines is equal to 10. For every age therefore, the predicted number of words is 10 words more for high SES children than for low SES children.

So far, this ordinary multiple regression. But suppose that such a model does not describe the data that we actually have, or does not make the right

predictions based on on our theories. Suppose our researcher also expects that the *yearly increase* in vocabulary is a bit lower than 500 words in low SES families, and a little bit higher than 500 words in high SES families. In other words, he believes that SES might *moderate* (affect or change) the slope coefficient for age. Let's call the slope coefficent in this case b_1. In the above equation this slope parameter is equal to 500, but let's now let itself have a linear relationship with SES:

$$b_1 = \alpha + b_3 \times SES \tag{1.8}$$

In words: the slope coefficient for the regression of vocabulary on age, is itself linearly related to SES: we predict the slope on the basis of SES. We model that by including a slope b_3, but also an intercept a. Now we have *two* linear equations for the relationship between vocabulary, age and SES:

$$\begin{aligned} \widehat{vocab} &= b_0 + b_1 \times age + b_2 \times SES & (1.9) \\ b_1 &= a + b_3 \times SES & (1.10) \end{aligned}$$

We can rewrite this by plugging the second equation into the first one (substitution):

$$\widehat{vocab} = b_0 + (a + b_3 \times SES) \times age + b_2 \times SES \tag{1.11}$$

Multiplying this out gets us:

$$\widehat{vocab} = b_0 + a \times age + b_3 \times SES \times age + b_2 \times SES \tag{1.12}$$

If we rearrange the terms a bit, we get:

$$\widehat{vocab} = b_0 + a \times age + b_2 \times SES + b_3 \times SES \times age \tag{1.13}$$

Now this very much looks like a regression equation with one intercept and *three* slope coefficients: one for age (a), one for SES (b_2) and one for SES$\times$ age (b_3).

We might want to change the label a into b_1 to get a more familiar looking form:

$$\widehat{vocab} = b_0 + b_1 \times age + b_2 \times SES + b_3 \times SES \times age \tag{1.14}$$

So the first slope coefficient is the increase in vocabulary for every year that age increases (b_1), the second slope coefficient is the increase in vocabulary for an increase of 1 on the SES variable (b_2), and the third slope coefficient is the

increase in vocabulary for every increase of 1 on the *product* of age and SES (b_3).
So what does this mean exactly?

Suppose we find the following solution for the regression equation:

$$\widehat{vocab} = b_0 + b_1 \times age + b_2 \times SES + b_3 \times SES \times age \tag{1.15}$$

$$\widehat{vocab} = 200 + 450 \times age + 125 \times SES + 100 \times SES \times age \tag{1.16}$$

If we code low SES children as SES=0, and high SES children as SES=1, we can write the above equation into two regression equations, one for low SES children (SES=0) and one for high SES chilrden (SES=1):

$$\begin{aligned} lowSES : \widehat{vocab} &= 200 + 450 \times age && (1.17) \\ highSES : \widehat{vocab} &= 200 + 450 \times age + 125 + 100 \times age && (1.18) \\ &= (200 + 125) + (450 + 100) \times age \\ &= 325 + 550 \times age \end{aligned}$$

So for low SES children, the intercept is 200 and the regression slope for age is 450, so they learn 450 words per year. For high SES children, we see the same intercept of 200, with an extra 125 (this is the main effect of SES). So effectively their intercept is now 325. For the regression slope, we now have $450 \times age + 100 \times age$ which is of course equal to $550 \times age$. So we see that the high SES group has both a different intercept, and a different slope: the increase in vocabulary is 550 per year: somewhat steeper than in low SES children. So yes, the researcher was right: vocabulary increase per year is faster in high SES children than in low SES children.

These two different regression lines are depicted in Figure 1.2. It can be clearly seen that the lines have two different intercepts and two different slopes. That they have two different slopes can be seen from the fact that the lines are not parallel. One has a slope of 450 words per year and the other has a slope of 550 words per year. This difference in slope of 100 is exactly the size of the slope coefficient pertaining to the product $SES \times age$, b_3. Thus, the interpretation of the regression coefficient for a product of two variables is that it represents *the difference in slope*.

The observation that the slope coefficient is different for different groups is called an *interaction effect*, or *interaction* for short. Other words for this phenomenon are *modification* and *moderation*. In this case, SES is called the *modifier variable*: it modifies the relationship between age on vocabulary. Note however that you could also interpret age as the modifier variable: the effect of SES is larger for older children than for younger children. In the plot you see that the difference between vocabulary for high and low SES children of age 6 is larger than it is for children of age 2.

So, what do you have to do if you want to know if there is an interaction effect between age and SES on vocabulary size?

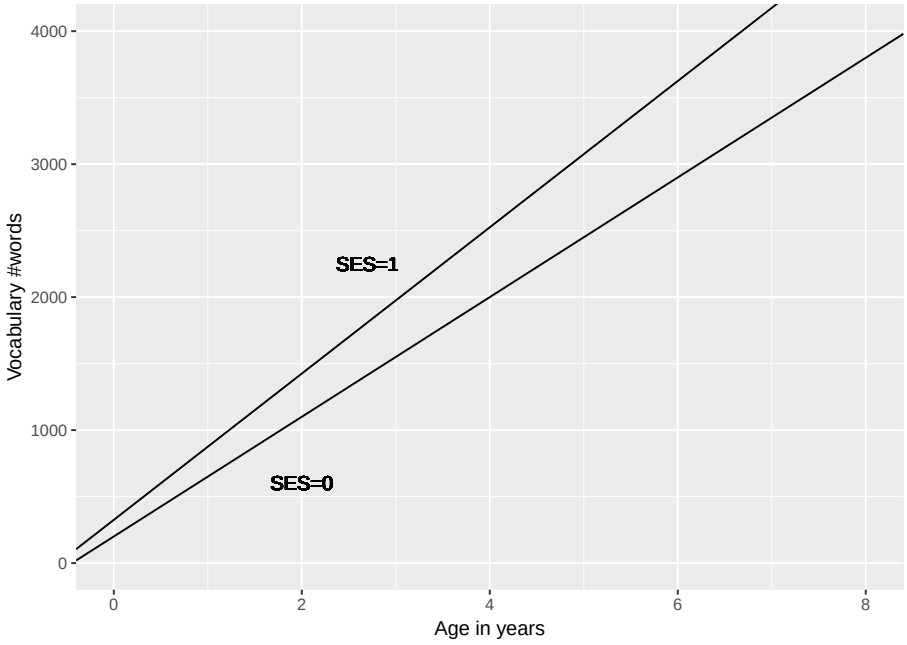

Figure 1.2: Two regression lines for the relationship between Age and Vocabulary Size, one for low SES children (SES=0) and one for high SES children (SES=1).

Table 1.1: Height of children as a function of age and location.

child	location	age	height
001	city	5	120
002	country	14	160
003	city	4	121
004	city	6	125
005	country	9	140
...	...	...	...

First you make sure that you dummy-code the grouping variable SES:

```
RECODE SES ('low'=0) ('high'=1) INTO SEShigh.
EXECUTE.
```

Next we compute a new variable, that is, the product $SES \times age$ (but use the dummy variable):

```
COMPUTE SEShighage = SEShigh * age .
EXECUTE.
```

This means that for every child in your data set, we take the age of the child (say 4), take the SEShigh value, say 1, and multiply these numbers: $4 * 1 = 4$.

So now you have three variables that we can use in a multiple regression analysis:

```
UNIANOVA vocab WITH age SEShigh SEShighage
/ design=age SEShigh SEShighage.
```

Note there is also a faster way of analyzing interaction effects in SPSS. The following syntax is exactly equivalent, but does not require the computation of the interaction variable $SEShighage$:

```
UNIANOVA vocab WITH age SEShigh
/ design = age SEShigh age*SEShigh
/ print = parameter.
```

With this design specification of **age*SEShigh**, SPSS computes the product automatically for you.

Let's look at some example output for another data set. A researcher is interested in childrens' height. She has data on children between the ages of 4 and 8, with measures on their height. She wants to know whether children growing up in the city grow just as fast as in the countryside. Part of the data are shown in Table 1.1.

The general regression of height on age might look like as shown in Figure 1.3. This regression line for the entire sample of children has a slope of around 6

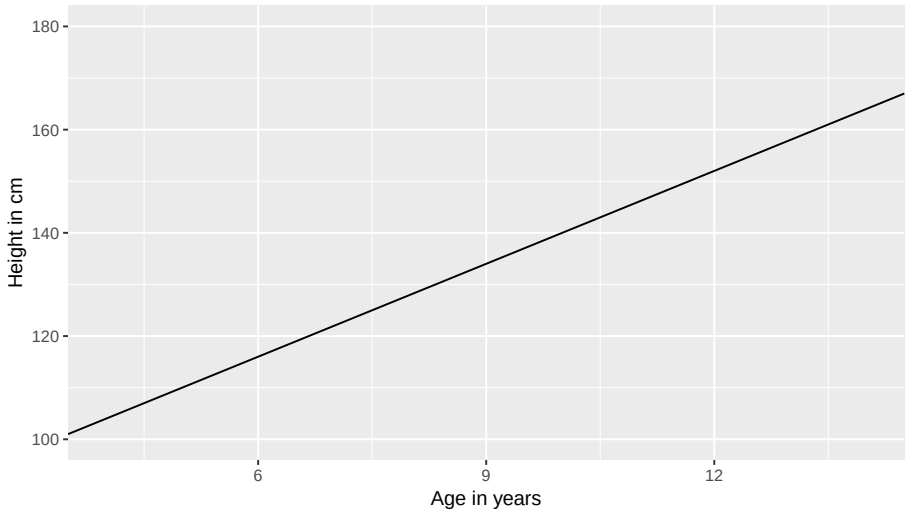

Figure 1.3: The effect of age on height.

cm per year. Now the researcher wants to know whether this slope is the same for children in the cities and in the countryside, in other words, do children grow as fast in the city as in the countryside? We might expect that location (city vs countryside) *moderates* the effect of age on height. We use the following SPSS syntax to study this $location \times age$ effect, using a dummy variable **countryside**, that codes countryside children as 1 and city children as 0:

```
UNIANOVA height WITH age countryside
 /design age countryside age*countryside
  /PRINT=PARAMETER.
```

In Figure 1.1 we find the corresponding SPSS output. So the null-hypothesis is that the two slopes are equal, in other words, that the interaction effect equals zero. In the output, this is the age * countryside effect.

In the table with the parameter estimates, we find the regression coefficients. We can now fill in the regression equation:

$$\widehat{height} = 96 + 4.6 \times age + 3.8 \times countryside - 0.368 \times age \times countryside$$

If we fill in 0s for the location dummy, we get the equation for city children:

$$\widehat{height} = 96 + 4.6 \times age$$

So the intercept equals 96 and the slope equals 4.6.

If we fill in 1s for the countryside dummy variable, we get the equation for countryside children:

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	96.652	6.068	15.928	.000	81.053	112.251
age	4.626	.704	6.568	.001	2.815	6.436
countryside	3.848	11.633	.331	.754	-26.055	33.751
age * countryside	-.368	1.621	-.227	.829	-4.535	3.799

Figure 1.4: Output with main effects of age and location_dummy, and an interaction effect.

$$\begin{aligned} \widehat{height} &= 96 + 4.6 \times age + 3.8 - 0.368 \times age \\ &= (96 + 3.8) + (4.6 - 0.368) \times age \end{aligned} \tag{1.19}$$

We see that that the intercept is now equal to the intercept is $96 + 3.8$, and the slope equals $4.6 - 0.368$.

So, we know that the slope for countryside children is 0.368 less steep than for city children. In this sample, it seems that children in the city grow 4.626 centimeters per year (on average), but that children in the countryside grow $4.626 - 0.368 = 4.258$ centimeters per year (on average). Is this value of 0.368 possible if the value in the entire population of children equals 0? In other words, is the value of 0.368 significantly different from 0? No, the effect of 0.368 is not significant, $t(5) = -0.23, p > 0.05$. We therefore do not reject the null-hypothesis and conclude that there is *no* evidence that children in the city grow at a different pace than children in the countryside.

Summarizig, in this section we discussed the situation that regression slopes might be different in two groups: the regression slope might be steeper in one group than in another group. So suppose that we had a numerical predictor x for a numerical dependent variable variable y, we said that a particular dummy variable z *moderated* the effect of x on y. This moderation was quantified by an *interaction* effect.

So suppose we have the following linear equation:

$$y = b_0 + b_1 \times x + b_2 \times dummy + b_3 \times x \times dummy + e$$

Then, we call b_0 the intercept, b_1 the main effect of x, b_2 the main effect of the dummy variable, and b_3 the interaction effect of x and the dummy.

1.1.1 Exercises

1. We have the following regression equation, with y as dependent variable, x as a continuous predictor variable, and a dummy variable $dummy$.

$$y = 5.3 + 3.6 \times x + 3.8 \times dummy + 8.2 \times x \times dummy + e \qquad (1.20)$$

   Write down the regression equation in the case the dummy variable equals 0.

2. Write down the regression equation in the case the dummy variable equals 1.

3. What is the intercept if the dummy variable equals 0?

4. What is the intercept if the dummy variable equals 1?

5. What is the slope if the dummy variable equals 0?

6. What is the slope if the dummy variable equals 1?

7. How large is the difference in intercepts between the two groups?

8. Where can we find this value in the equation?

9. How large is the difference in slopes between the two groups?

10. Where can we find this value in the equation?

11. We have the following regression equation, with y as dependent variable, x as a continuous predictor variable, and a dummy variable $dummy$.

$$y = -4.1 + 1.2 \times x - 6.5 \times dummy - 1.3 \times x \times dummy + e \qquad (1.21)$$

    Write down the regression equation in the case the dummy variable equals 0.

12. Write down the regression equation in the case the dummy variable equals 1.

13. What is the intercept if the dummy variable equals 0?

14. What is the intercept if the dummy variable equals 1?

15. What is the slope if the dummy variable equals 0?

16. What is the slope if the dummy variable equals 1?

17. How large is the difference in intercepts between the two groups?

18. Where can we find this value in the equation?

19. How large is the difference in slopes between the two groups?
20. Where can we find this value in the equation?
21. Suppose we find the following linear equation:

$$\widehat{mathscore} = 16.3 + 5.5 \times age - 0.8 \times sex - 1.2 \times age \times sex \qquad (1.22)$$

    What is the main effect of *age* on mathscore?
22. What is the main effect of the *sex* on mathscore?
23. How large is the interaction effect of *age* and *sex* on mathscore?
24. What is the predicted mathscore for a girl of age 12, if sex is coded 1 for boys?
25. What is the predicted mathscore for a boy of age 22, if sex is coded 1 for boys?

1.1.2 Answers

1.

1.2 Interaction between two dichotomous variables

In the previous section we discussed the situation that regression slopes might be different in two groups. Now we discuss the situation that we have two dummy variables, and that we're interested whether there is an interaction effect. In other words, does one dummy variable moderate the effect of the other dummy variable?

Suppose in country A, men are on average taller than women. In order to study this effect, we analyze data from a random sample of inhabitants, and we come up with the following regression equation:

$$\widehat{height} = 165 + 10 \times sex$$

In this equation, sex is coded 0 for females, and 1 for males. So, the predicted height for a female from country A equals 165 and the predicted height for a male equals $165 + 10 \times 1 = 175$.

Suppose we also study height in country B. Again with a random sample of inhabitants, we find the following regression equation:

$$\widehat{height} = 175 + 15 \times sex$$

In this equation, the predicted height for a female from country B equals 175 and the predicted height for a male equals $175 + 15 \times 1 = 190$.

So it seems that in general, the people in the random sample from country B are taller than the people in the random sample from country A: both men and women show taller averages in country B. But we also see another difference between the two countries: the average difference between men and women is 10 cm in country A, but 15 cm in country B. So we can say that in these samples, the effect of sex on height is a little bit different in both countries. Now of course this difference could be a coincidence, a random result from sampling, or it could be a real thing in the populations. Suppose we'd like to know whether the effect of sex on height is different in the two countries at population level. We'd like to know whether country is a moderator of the effect of age on height. So we use the following regression equation:

$$\widehat{height} = b_0 + b_1 \times sex + b_2 \times country + b_3 \times sex \times country$$

and perform a regression equation.

The easiest option, as we have seen earlier, is to let SPSS do the dummy coding. Simply use the BY keyword to indicate that country and sex are categorical variables. Additionally, include the multiplication in the DESIGN subcommand to indicate that you want to model an interaction effect:

```
UNIANOVA height BY sex country
/ DESIGN = sex country sex*country
/ PRINT = parameter.
```

In Figure 1.2 we see the relevant output. We see that the intercept is 190. Then we see that the people from country A get an extra -15 cm, and that for those with sex 0 get an additional -15 cm. On top of that, those who come from country A *and* have sex=0 (females), have an extra -5 cm. Thus, the expected height from women from country A equals $190 - 15 - 15 - 5 = 155$ cm. The expected height of a male (sex = 1) from country A is then $190-15+0+0 = 175$. The expected height of a female from country B is $190 + 0 - 15 + 0 = 175$, and the expected height of a male from country B is $190 + 0 + 0 + 0 = 190$.

The difference of the differences (the interaction effect) equals -5. We see that women when they come from country A, have an extra height of -5 cm in comparison to women from country B. But equally we could say: We see

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	190.000	1.369	138.756	.000	187.097	192.903
[country=A]	-15.000	1.936	-7.746	.000	-19.105	-10.895
[country=B]	0[a]	.	.	.	.	.
[sex=.00]	-15.000	1.936	-7.746	.000	-19.105	-10.895
[sex=1.00]	0[a]	.	.	.	.	.
[country=A] * [sex=.00]	5.000	2.739	1.826	.087	-.806	10.806
[country=A] * [sex=1.00]	0[a]	.	.	.	.	.
[country=B] * [sex=.00]	0[a]	.	.	.	.	.
[country=B] * [sex=1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 1.5: Output with main effects of country and sex, and an interaction effect.

that women when they come from country A have an extra height of -5 cm in comparison to males from country A. Interpretation of these results is best seen in a graph showing the means of the four groups, see Figure 1.6. The mean height difference between country A and country B is larger in females.

The data could also be represented in a different way, see Figure 1.7. There we see that the mean height difference between males and females is larger in country A than in country B. Thus, there are two ways of describing the data: either you look at the effect of country and see sex as a modifier variable, or you look at the effect of sex, and see country as a modifier. Both are describing the same interaction effect: the extra -5 cms for females from country A.

Whether the interaction effect also exists at the population level, we can see from SPSS output. Here the relevant null-hypothesis is that there is no interaction effect. This means that the coefficient for the interaction effect is equal to 0:

$$H_0 : \beta_{sex*country} = 0 \tag{1.23}$$

If the effect is significant at your pre-set level of significance (i.e. $p < \alpha$), you reject the null-hypothesis and conclude that *the difference between males and females in height is different in these two countries.* Or, equivalently, we conclude that *the difference in height between the two countries is different for males and females.* If the effect is not significant, we do not reject the null-hypothesis and conclude that the difference in height between females and males is the same in country A and B. Or, equivalently, we conclude that the difference in height between the two countries is the same for males and females.

From now on, we recommond using the BY syntax for variables that you wish to analyze qualitatively (all categorical variables, and sometimes ordinal

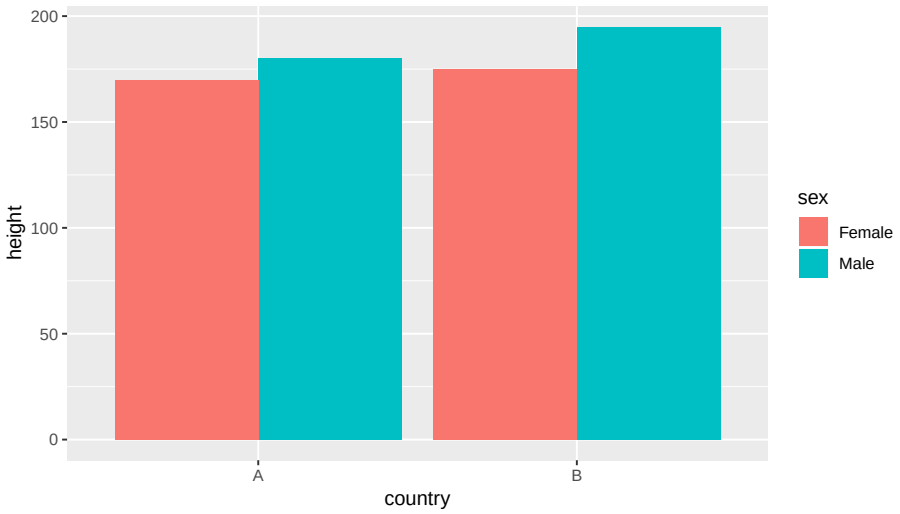

Figure 1.6: Average heights for males and females in countries A and B.

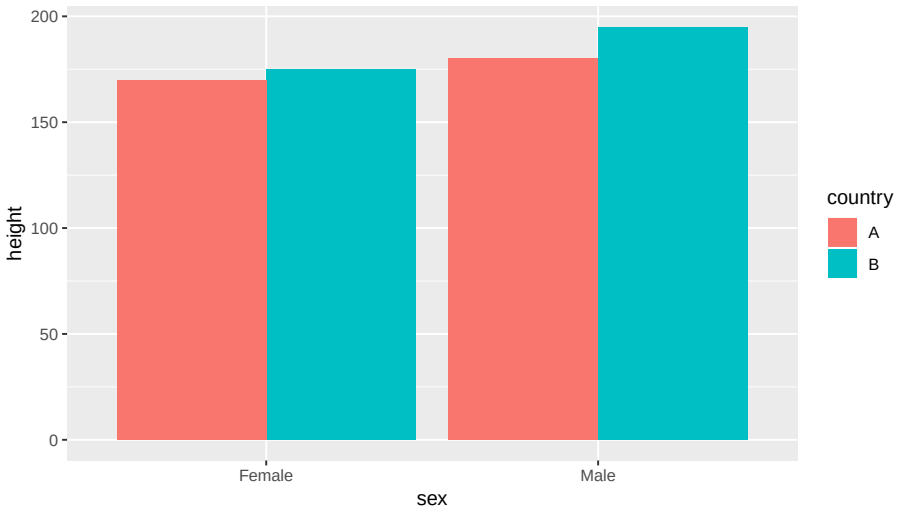

Figure 1.7: Average weights for males and females in countries A and B.

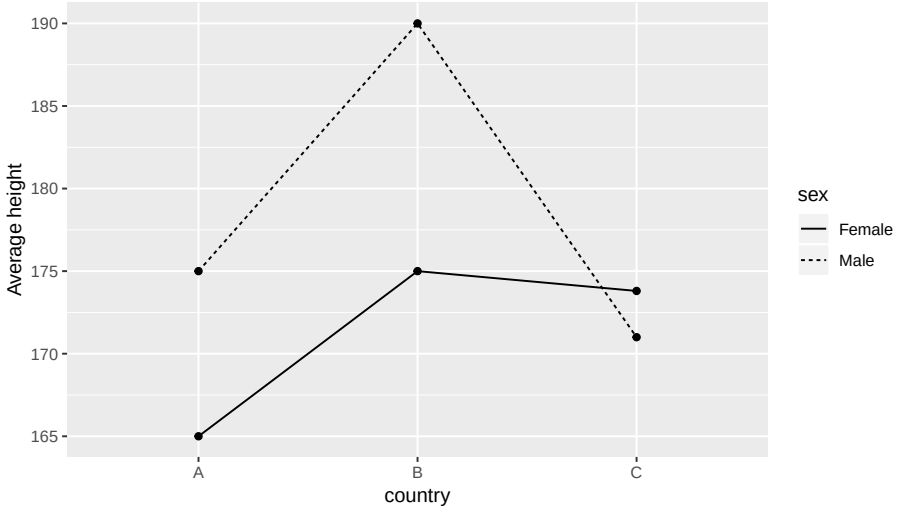

Figure 1.8: Average weights for males and females in countries A, B and C.

variables). Only when you find the output hard to interpret, make your own dummy variables and use the WITH keyword.

1.2.1 More than two groups

What happens when we have a categorical variable with more than two levels? Suppose we want to do the same study on height but now in countries A, B and C. In Figure 1.8 we see the average heights that we observe in the sample data.

Now we see a clear difference in the countries: the males are on average larger than the females, but this is only true for countries A and B. In country C the females are on average larger than the males. However, remember that this is based on a sample data. We'd like to know whether male-female differences in average height vary from country to country also in the population data. We therefore do an inferential data analysis using a linear model, including a sex by country interaction effect. Our null-hypothesis is

$$H_0 : \mu_{femaleA} - \mu_{maleA} = \mu_{femaleB} - \mu_{maleB} = \mu_{femaleC} - \mu_{maleC} \quad (1.24)$$

As we saw earlier, in SPSS we can treat variables in a regression analysis either as quantitative or qualitative. If we want to treat variable as quantitative, we use the word WITH, and if we want to treat the variable as qualitative, we use the word BY in the SPSS syntax. If you have made your own dummy variable, then use WITH. If you want SPSS to do the dummy coding for you, use BY. When you have a variable with more than two levels, say country with three levels, we generally recommend using the BY word. This makes SPSS turn the categorical variable into two dummy variables automatically.

Suppose you have the categorical variable country with levels A, B and C, and you have the sex variable dummy coded as 1 for males and 0 for females. You want to treat the dummy variable quantitatively, and the country variable qualitatively. Then with the next syntax you can run a regression analysis with a main effect of sex, a main effect of country and an interaction effect of sex by country in the following way.

```
UNIANOVA height BY country sex
/ design = sex country sex*country
/ print = parameter.
```

Tests of Between-Subjects Effects

Dependent Variable: height

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	1712.167[a]	5	342.433	21.358	.000
Intercept	918400.033	1	918400.033	57280.667	.000
sex	410.700	1	410.700	25.615	.000
country	880.067	2	440.033	27.445	.000
country * sex	421.400	2	210.700	13.141	.000
Error	384.800	24	16.033		
Total	920497.000	30			
Corrected Total	2096.967	29			

a. R Squared = .816 (Adjusted R Squared = .778)

Parameter Estimates

Dependent Variable: height

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	171.000	1.791	95.492	.000	167.304	174.696
[sex=.00]	2.800	2.532	1.106	.280	-2.427	8.027
[sex=1.00]	0[a]	.	.	.	.	.
[country=A]	4.000	2.532	1.579	.127	-1.227	9.227
[country=B]	19.000	2.532	7.503	.000	13.773	24.227
[country=C]	0[a]	.	.	.	.	.
[country=A] * [sex=.00]	-12.800	3.581	-3.574	.002	-20.192	-5.408
[country=A] * [sex=1.00]	0[a]	.	.	.	.	.
[country=B] * [sex=.00]	-17.800	3.581	-4.970	.000	-25.192	-10.408
[country=B] * [sex=1.00]	0[a]	.	.	.	.	.
[country=C] * [sex=.00]	0[a]	.	.	.	.	.
[country=C] * [sex=1.00]	0[a]	.	.	.	.	.

a. This parameter is set to zero because it is redundant.

Figure 1.9: Main effects of country (A, B, and C) and sex (0,1) and the country by sex interaction effect.

The SPSS output is in Figure 1.2.1. In the Parameter Estimates table we see that 2 dummy variables have been computed, automatically by SPSS. One for being in country A, and one for being in country B. Country C is here used as the so-called reference category. This SPSS output is therefore equivalent to the equation:

$$\begin{aligned} \widehat{height} &= 173.8 - 2.8 \times sex - 8.8 \times CountryA + 1.2 \times CountryB \\ &+ 12.8 \times CountryA \times sex + 17.8 \times CountryB \times sex \end{aligned}$$

All observations done in country C for variables CountryA and CountryB are coded as 0. So let's do the math to get the predicted heights for each subgroup. Females are coded as 0 and males as 1, so a Female from country C gets the predicted value 173.8. Let's do the computations for all subgroups:

Sex	Country	equation	height
Female	A	$173.8 - 2.8 \times 0 - 8.8 \times 1 + 1.2 \times 0 + 12.8 \times 1 \times 0 + 17.8 \times 0 \times 0$	165
Male	A	$173.8 - 2.8 \times 1 - 8.8 \times 1 + 1.2 \times 0 + 12.8 \times 1 \times 1 + 17.8 \times 0 \times 1$	175
Female	B	$173.8 - 2.8 \times 0 - 8.8 \times 0 + 1.2 \times 1 + 12.8 \times 0 \times 0 + 17.8 \times 1 \times 0$	175
Male	B	$173.8 - 2.8 \times 1 - 8.8 \times 0 + 1.2 \times 1 + 12.8 \times 0 \times 1 + 17.8 \times 1 \times 1$	190
Female	C	$173.8 - 2.8 \times 0 - 8.8 \times 0 + 1.2 \times 0 + 12.8 \times 0 \times 0 + 17.8 \times 0 \times 0$	173.8
Male	C	$173.8 - 2.8 \times 1 - 8.8 \times 0 + 1.2 \times 0 + 12.8 \times 0 \times 1 + 17.8 \times 0 \times 1$	171

Note that we now have very different values for the regression parameters than in the analysis with only countries A and B (see Figure 1.2), but nevertheless we end up with the same expected heights in Countries A and B. The difference in the parameter values stems from the fact that we have now treated country C as the reference category (coefficient fixed to 0), whereas in the previous two country analysis, we treated country B as the reference category.

Let's test the hypothesis of equal differences in heights between males and females across the three countries. In the output we see that the Country=A by sex interaction effect is significant at 0.05: there is an extra height of 12.8 cms seen in males from country A, over and above the main effects of being male in general and being from country A. In other words, the effect of being male is larger in country A than it is in Country C (the reference country). We also see this in the predicted means: male-female difference in country C is -2.8 (males shorter), but in country A it is +10 (males larger). In the output we also see that the CountryB by sex interaction effect is significant at 0.05: the effect of being male is 17.8 cm larger in country B than in Country C (the reference category). From the means we see that the male-female difference is 15 in country B, which is 17.8 cm more than the -2.8 in country C. So both these interaction effects are significant. Similarly to the previous chapter, we now have two coefficients to test one hypothesis, so again we should do an F-test to test the hypothesis that male-female differences are the same across all three countries, or, equivalently, that country differences in height are the same of males and females.

Therefore, we should look at the Analysis of Variance (ANOVA) table (Tests of Between-Subjects Effects). There we see that for the country*sex interaction effect we have an F-value of 13.141. With 2 model degrees of freedom (number of dummy variables) and 24 error degrees of freedom, the probability of getting an F-value of at least 13.141, given that the null-hypothesis is true, equals less

than 0.001. Therefore we conclude that in the populations of countries A, B and C, the difference in height between males and females is significantly different, $F(2, 24) = 13.141, MSE = 210.70, p < 0.001$. Alternatively, but equivalently, we may conclude that the differences in height across the three countries, are significantly different for males than for females, $F(2, 24) = 13.141, MSE = 210.70, p < 0.001$.[1].

1.2.2 Exercises

From a sample of data on height, country, and weight, we get the following linear equation:

$$\widehat{weight} = 40 + 30 \times CountryA + 0.4 \times height + 0.1 \times CountryA \times height$$

1. What is the expected weight for an individual from country A with a height of 1.5?

2. What is the expected weight for an individual from country B with a height of 1.0?

3. How large is the slope coefficient of height in country A?

4. How large is slope coefficient of height in country B?

Answers:

1. $$\widehat{weight} = 40 + 30 \times 1 + 0.4 \times 1.5 + 0.1 \times 1 \times 1.5 = 70.75$$

2. $$\widehat{weight} = 40 + 30 \times 0 + 0.4 \times 1.0 + 0.1 \times 0 \times 1.0 = 40.4$$

3. $0.4 + 0.1 = 0.5$

4. 0.4

[1] Note that we never report $p = 0.000$. A p-value is always greater than 0, no matter how small. Therefore, for very small values, we report $p < 0.001$.

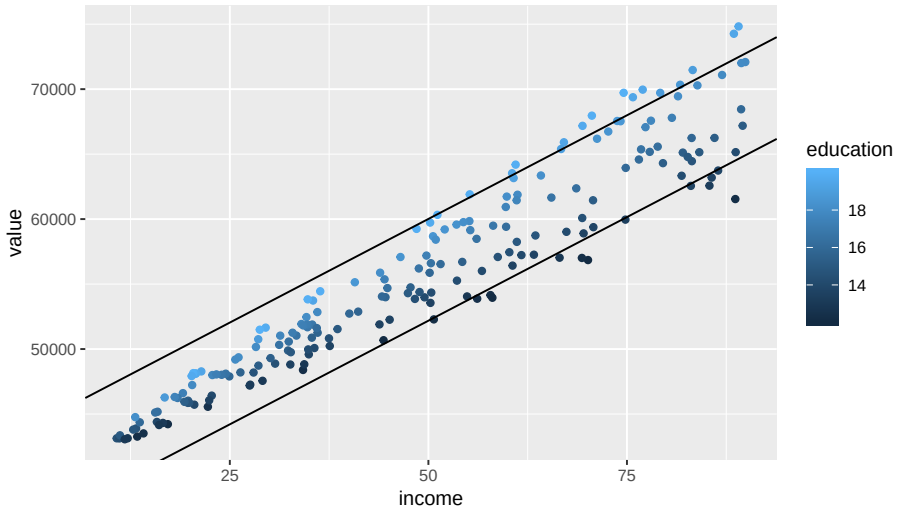

Figure 1.10: Sample data on gross yearly income, number of educational years and home market value.

1.3 Interaction between two numeric variables

Suppose we have data on current market value of housing properties. Suppose we also have data on 200 individuals, including their gross yearly income and the number of years spent in the national educational system. We'd like to see what the relationship is between income and education on the one hand, and the value of the house they live in on the other hand. Do richer people live in more valuable homes? Do people with more educational years live in more valuable homes?

Let's carry out a multiple regression analysis and find out. We use the syntax

```
UNIANOVA value WITH income  educatin
/DESIGN income  educatin
/PRINT parameter.
```

and find the output in Figure 1.3.

Based on this output, the linear equation for the relation between income and home market value is

$$\widehat{value} = 24482 + 319income + 979education \tag{1.25}$$

If education equals 20, we get the equation

$$\widehat{value} = 24482 + 319income + 979 * 20 = 44062 + 319income \tag{1.26}$$

If education equals 12, we get the equation

Parameter Estimates

Dependent Variable: value

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	24477.223	542.147	45.149	.000	23408.066	25546.379
income	319.181	3.219	99.158	.000	312.833	325.529
educatin	979.177	32.930	29.735	.000	914.237	1044.117

Figure 1.11: Main effects of income and educational years on home market value.

$$\widehat{value} = 244821 + 319income + 979 * 12 = 36230 + 319income \tag{1.27}$$

We see that for different values of education, the intercepts are different, but the slopes are equal. We can see the two regression lines in Figure 1.10. Somehow it does not seem to be a good model. For high income, we see relatively large differences between different levels of education. For low income we see small differences for educational years. Thus we could say that these sample data seem to suggest that the effect of educational years on the home market value is larger for high income people than for for low income people.

We could also look at it from a different angle. In Figure 1.10 we see that the relationships between income and value is much steeper for people with many educational years (the light blue dots), than for people with few educational years (the dark dots).

Both observations seem to suggest a moderation effect. One could say that eduation moderates the relation between income and value, or one could say that income moderates the relation between eduational years and value. We can therefore try a linear model that includes an interaction effect of income and education on home market value.

The syntax is

```
UNIANOVA value WITH income  educatin
/DESIGN income  educatin income*educatin
/PRINT parameter.
```

and the output is in Figure 1.3

Based on this output, the linear equation for the relation between income and home market value is

$$\widehat{value} = 40014 - 0.4income - .8education + 20income \times education \tag{1.28}$$

If education equals 20, we get the equation

Parameter Estimates

Dependent Variable: value

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Intercept	40013.865	11.858	3374.542	.000	39990.480	40037.250
income	-.384	.221	-1.734	.084	-.820	.053
educatin	-.802	.743	-1.080	.282	-2.266	.663
income * educatin	20.023	.014	1458.971	.000	19.996	20.050

Figure 1.12: Main effects of income and educational years on home market value.

$$\widehat{value} = 40014 - 0.4income - .8 \times 20 + 20income \times 20 = 39998 + 399.6income \tag{1.29}$$

If education equals 12, we get the equation

$$\widehat{value} = 40014 - 0.4income - .8 \times 12 + 20income \times 12 = 40003.4 + 239.6income \tag{1.30}$$

Now we see that for different values of eduation, both the intercept and the slope are different. In Figure **??** these two regression lines are plotted. These nonparallel lines seem to describe the data much better than the parallel lines in Figure **??**.

From the output, we also see that the interaction effect has very small p-value. We can therefore reject the null-hypothesis that the effect of income on home market value is the same for all levels of education. More precisely, we can reject the null-hypothesis that the *slope* of the regression line for value on income is the same for all levels of education. It seems that for people with many years of education, there is a stronger relationship between income and home market value than for people with fewer years of education.

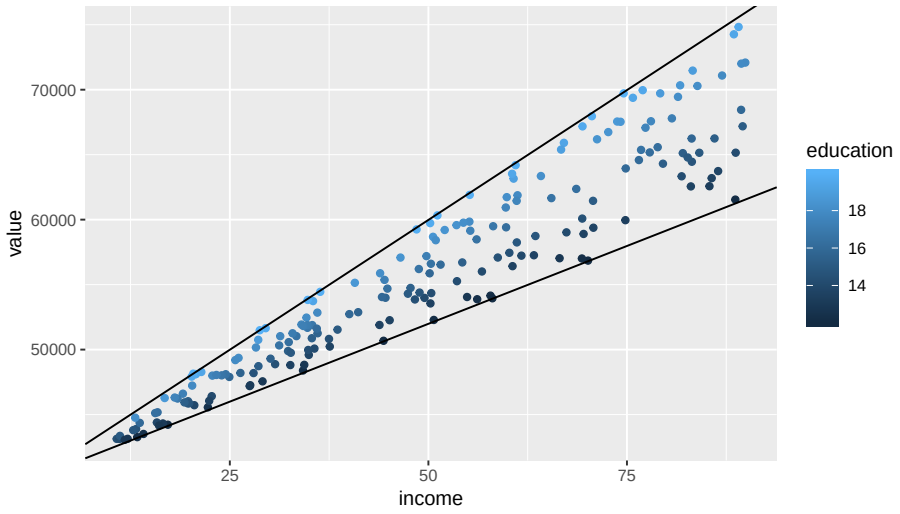

Figure 1.13: Sample data on gross yearly income, number of educational years and home market value.