Exercise 16 - Multiple regression – different types of predictors

Zoltan Kekecs

13 May 2021

Table of Contents

Abstract

This exercise will show you how multiple predictors can be used in the same regression model to achieve better prediction efficiency. It will focus on different types of predictors that can be used in the regression and ANOVA models.

Data management and descriptive statistics

Load data

To explore some of the more advanced predictor types, we will need a new dataset. Let's download the weight_loss dataset.

https://github.com/kekecsz/SIMM32/blob/master/2021/Lab_3/weight_loss_data.csv

This is in a **.csv file format**, so some extra steps might be necessary before you can load it in SPSS. The .csv is a very common data format, so it is good to learn how to open this in spss. On github while you are on the file's page you can hold **alt** and click on the "raw" button on the top right this will download the file in a .txt format. In spss you can open this with the regular open command, but you have to specify, that the file format is .txt. You will see a dialog box where you will have to enter certain information about the file itself, such as which is the row containing variable names (1), what is the decimal symbol (period), whether columns are fixed width or delimited (delimited), data starts in which row (2), and what is the delimiter between variables (coma). In this dialog box you can also specify the

variable types. Click finish when you are done, and make sure that the variable types are correct in the variable view.

This dataset contains simulated data, so it was not collected in an actual research study, instead, it was generated in the program. It is about a study where different types of **interventions were tested to understand their effect on loosing weight**.

Variables:

ID – participant ID

Gender - gender

Age – age

BMI_baseline – Body mass index measured before treatment

BMI_post_treatment – Body mass index measured after treatment

treatment_type – The type of treatment in the group to which the participant was randomized to.

Levels:

no treatment

pill – medication which lowers apetite

psychotherapy – cognitive behavioral therapy

combined – a combination of pill and psychotherapy treatment

lowcaldiet – whether the person is on a low calorie diet (1) or not (0)

exercise – whether the person is doing regular exercise (1) or not (0)

motivation – self report motivation to lose weight (on a 0-10 scale from extremely low motivation to extremely high motivation)

body_acceptance – how much the person feels that he or she is satisfied with his or her body. (on a scale of -7 to +7 from very unsatisfied to very satisfied).

In this exercise we would like to understand the effect of the different treatment types on BMI.

Check the dataset

Analyze > Descriptive Statistics tab > Frequencies

Analyze > Descriptive Statistics tab > Descriptives

Analyze > Descriptive Statistics tab > Explore

Remember to check the variables with the descriptive statistics and with plots for visualization.

Lets build a model to predict post-treatment BMI.

Different types of predictors

Categorical predictor

Categorical variables can be included in linear models, as long as the predicted outcome is continuous.

Our treatment type variable is a string variable (text), but SPSS does not allow string variables to be used in most linear models, so we need to recode our variable to be numeric.

Dummy coding

We will have to dummy-code our categorical variables if we want to enter them as predictors in the linear regression model. We can do this for example by using the **Transform > Recode into different variable** tab.

Lets say for example, that we would like to dummy code a categorical variable with two levels, such as gender in our dataset. We select gender **Transform > Recode into different variable** tab and put it in the center box, and on the right we specify the name of the new variable into which we would like to recode this variable (e.g. male). After clicking OK we can specify which value would we like to recode into which new value. I suggest recoding male as 1 and female as 0 (because we expect that females would have lower BMI, see also below). This way, if the male variable is 1, it means that that particular participant is a male.

```
RECODE gender ('male'=1) (ELSE=0) INTO male.
EXECUTE.
```

In case of a variable with multiple levels, we need to create multiple variables to be able to enter this as a predictor into the regression model. The number of new (dummy) variables is always 1 less than the number of levels of the original categorical variable. For example in the case of treatment type, where we have 4 levels: no treatment, pill, psychotherapy, and treatment 3, we should create 3 dummy variables.

The process should be:

1. **select a "baseline" level**, to which every other level is compared. For example in the case of treatment type this could be "no_treatment". It does not matter in terms of building accurate models, which level we select as the baseline level, but **it will ease the interpretation of the model coefficients if we select a baseline level which intuitively represents something like "no" or "zero"**, or at which the dependent variable is expected to take a lower value. (That is why I chose to create a dummy for male and not female above, because I think BMI would be lower for females, so I treat that as the baseline level, to which males will be compared.)

2. Using **Transform > Recode into different variable,** create new variables, where one of the non-baseline levels of the variable is recoded as 1, while every other value is recoded as 0. Repeat this until you have created **one dummy variable for every level NOT chosen as the baseline**. This means that we will have one less dummy variables than category/factor levels.
In the case of treatment type, we will use no_treatment as the baseline, and we create 3 dummy variables, called treatment_pill, treatment_psychotherapy, and treatment_combined respectively. For example, a value 1 in the dummy variable treatment_psychotherapy means that the given participant was in the psychotherapy group, while a value 0 would mean that the given participant was in one of the other groups, not the psychotherapy group.

```
RECODE treatment_type ('pill'=1) (ELSE=0) INTO treatment_pill.
EXECUTE.
RECODE treatment_type ('psychotherapy'=1) (ELSE=0) INTO
treatment_psychotherapy.
EXECUTE.
RECODE treatment_type ('combined'=1) (ELSE=0) INTO treatment_combined.
EXECUTE.
```

Building a regression model with the dummy variables

Now we can build a regression model, with the post-treatment BMI as a dependent variable, and treatment_pill, treatment_psychtherapy, and treatment_combined as predictors. As usual, we should ask for the confidence intervals in the statistics button.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT BMI_post_treatment
  /METHOD=ENTER treatment_pill treatment_psychotherapy
treatment_combined.
```

The full-model F test indicates that the model is significantly better than the null model ($F(3, 236) = 27.72$, $p < 0.001$).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1024,246	3	341,415	27,720	,000[b]
	Residual	2906,750	236	12,317		
	Total	3930,996	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), treatment_combined, treatment_psychotherapy, treatment_pill

Interpreting the coefficients

The interpretation of the linear regression coefficients are always to same, as we have learned in the exercise about multiple regression.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	34,100	,453		75,263	,000	33,207	34,993
	treatment_pill	-1,717	,641	-,184	-2,679	,008	-2,979	-,454
	treatment_psychotherapy	-1,183	,641	-,127	-1,847	,066	-2,446	,079
	treatment_combined	-5,517	,641	-,590	-8,610	,000	-6,779	-4,254

a. Dependent Variable: BMI_post_treatment

The coefficient of the constant (**the intercept**) is the estimated value of the outcome variable if all predictor's value is 0. In this particular case this means **the expected value of the outcome variable at the baseline level** (or default level) of the categorical variable (in our case, no_treatment). In our case, the value of the intercept is 34.1, meaning that if the person gets no treatment, we their estimated BMI will be 34.1.

Likewise, the interpretation of the coefficients of the predictors is the same as before: with every other variable held constant, moving one-step up on the value of the predictor variable would produce this expected change/difference in the dependent variable. So when we have the dummy variables as predictors, the coefficients of these variables tell us the expected effect of the given category level compared to the baseline level, or in other words the difference between the mean of the outcome variable on the given level of the categorical variable and the baseline level. For example in our case the coefficient -1.72 associated with the predictor treatment_pill means that if a person gets the pill treatment they are expected to have 1.72 lower BMI after the treatment compared to people who got no treatment.

We can include other predictors as well, even continuous predictors. For example, we could include motivation as a predictor in this model, to account for the effect of the person's drive to lose weight.

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT BMI_post_treatment
 /METHOD=ENTER treatment_pill treatment_psychotherapy
treatment_combined motivation.
```

In this case, the interpretation of the intercept's coefficient would be: the expected value of the dependent variable at motivation = 0 and at the baseline/default level of the treatment: no_treatment. The interpretation of the other coefficents stays the same as well as before: that is, if a person reported 3 instead of 2 on the motivation scale, we expect that their BMI would be 0.22 lower than that of a person who reported 2 on the motivation scale. The same goes for the coefficient of the dummy variables, with the difference that their coefficient is always interpreted compared to the default level/baseline level that we chose.

One-way ANOVA

Another way to look at the effect of treatment type on post-treatment BMI is to run a one-way ANOVA. *This will return equivalent results to the linear regression, so you can skip this part if you are happy with doing regression for this analysis and continue with the "Higher order terms" section.* Read on if you are interested in how to do the above mentioned analysis with ANOVA and how to interpret the results.

We can use the auto-recode feature to recode the treatment_type string variable in **Transform > Auto-recode**, which will automatically choose numbers with which it will replace our category names. If we want more control over which numbers will replace which values, we can use **Transform > Recode into different variable.** I used the auto-recode here and specified the name of the new numeric variable as treatment_type_num.

```
AUTORECODE VARIABLES=treatment_type
 /INTO treatment_type_num
 /PRINT.
```

We can build a simple one-way ANOVA model at **Analyze > Compare means > One-way ANOVA**. The dependent is post-treatment BMI, while the factor is treatment_type_num.

We should ask for descriptives, mean plot, and a post-hoc test with Bonferroni correction to better understand the effect if we find one, and a homogeneity test to be able to assess whether the assumption of ANOVA hold true.

```
ONEWAY BMI_post_treatment BY treatment_type_num
 /STATISTICS DESCRIPTIVES HOMOGENEITY
 /PLOT MEANS
 /MISSING ANALYSIS
 /POSTHOC=BONFERRONI ALPHA(0.05).
```

The output tells us that a significant portion of the variability between individuals is explained by taking into consideration the treatment type ($F(3, 236) = 27.72$, $p < 0.001$), meaning that at least one of the groups showed significantly different mean than at least one of the other groups.

ANOVA

BMI_post_treatment

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1024,246	3	341,415	27,720	,000
Within Groups	2906,750	236	12,317		
Total	3930,996	239			

The descriptive table and the mean plot shows us that treatment 3 was the best treatment in terms of producing the lowest post-treatment BMIs.

Multiple Comparisons

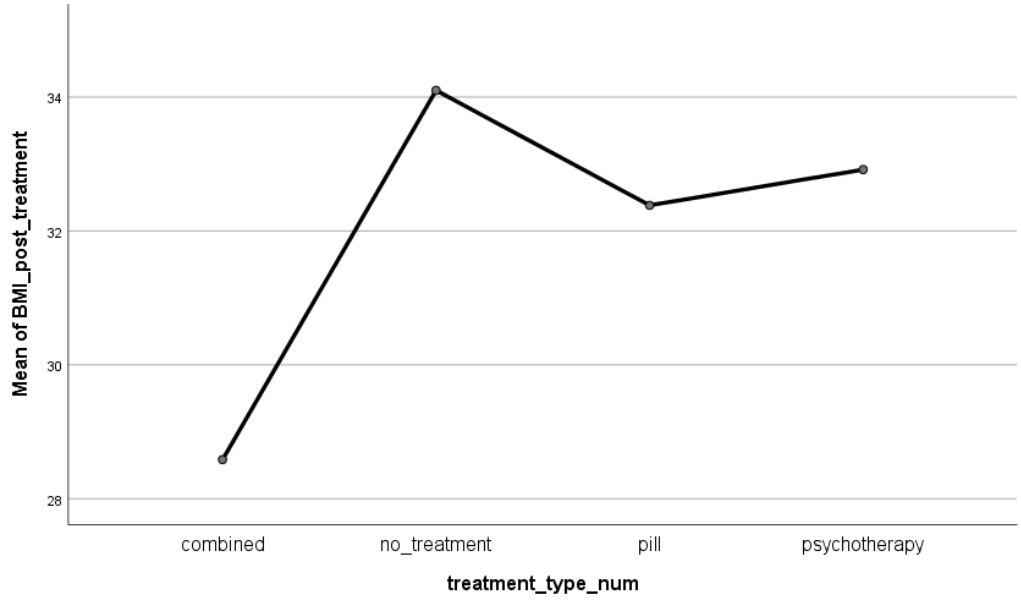
Dependent Variable: BMI_post_treatment

Bonferroni

(I) treatment_type_num	(J) treatment_type_num	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval Lower Bound	95% Confidence Interval Upper Bound
combined	no_treatment	-5,517*	,641	,000	-7,22	-3,81
	pill	-3,800*	,641	,000	-5,50	-2,10
	psychotherapy	-4,333*	,641	,000	-6,04	-2,63
no_treatment	combined	5,517*	,641	,000	3,81	7,22
	pill	1,717*	,641	,047	,01	3,42

	psychotherapy	1,183	,641	,396	-,52	2,89
pill	combined	3,800*	,641	,000	2,10	5,50
	no_treatment	-1,717*	,641	,047	-3,42	-,01
	psychotherapy	-,533	,641	1,000	-2,24	1,17
psychotherapy	combined	4,333*	,641	,000	2,63	6,04
	no_treatment	-1,183	,641	,396	-2,89	,52
	pill	,533	,641	1,000	-1,17	2,24

*. The mean difference is significant at the 0.05 level.



The post-hoc analysis compares all groups pair-by-pair to see which of the groups are significantly different from each-other. To make a statistical inference, we cannot use our original p-value threshold, since we are making multiple comparisons here which inflates the type 1 error rate.

So we use Bonferroni correction, dividing the p-value threshold by as much as many comparisons we make. If we make 10 pairwise comparisons, the p-value threshold below which we can say that the difference is significant would be 0.05/10 = 0.005. SPSS uses an equivalent of this, where the p-value threshold remains unchanged (0.05), but the p-value of the test itself is multiplied by the number of comparisons made. This results in similar decisions (although this way the adjusted p-value can exceed 1, which is not possible statistically, so in these cases SPSS reports 1 as the p-value).

The post-hoc comparisons show us that except for pill vs. psychotherapy and no treatment vs. psychotherapy, every other comparison was significant, so pill is better than no treatment, but less good than combined treatment, and combined treatment is better than

psychotherapy, and we can't really tell whether psychotherapy or pill is better, or whether psychotherapy or no treatment are better.

Multiple Comparisons

Dependent Variable: BMI_post_treatment

Bonferroni

(I) treatment_type_num	(J) treatment_type_num	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
combined	no_treatment	-5,517*	,641	,000	-7,22	-3,81
	pill	-3,800*	,641	,000	-5,50	-2,10
	psychotherapy	-4,333*	,641	,000	-6,04	-2,63
no_treatment	combined	5,517*	,641	,000	3,81	7,22
	pill	1,717*	,641	,047	,01	3,42
	psychotherapy	1,183	,641	,396	-,52	2,89
pill	combined	3,800*	,641	,000	2,10	5,50
	no_treatment	-1,717*	,641	,047	-3,42	-,01
	psychotherapy	-,533	,641	1,000	-2,24	1,17
psychotherapy	combined	4,333*	,641	,000	2,63	6,04
	no_treatment	-1,183	,641	,396	-2,89	,52
	pill	,533	,641	1,000	-1,17	2,24

*. The mean difference is significant at the 0.05 level.

ANOVA is a special case of linear regression

Here, it is important to note that ANOVA is a special case of linear regression, and that we can actually get the same numbers as with one-way ANOVA using the previously discussed linear regression. However, we need to further recode our treatment type variable to be able to build that linear model. This is because SPSS's linear regression is not optimized to include categorical predictors. String variables are not accepted by regression models in SPSS, and if we used the treatment_type_num as a predictor, SPSS would use it as a scale variable (instead of a nominal variable), and would think that the numbers that we use as codes are meaningful, and represent the difference between the levels. Instead, to get the correct interpretation, we need to use dummy coding.

Higher order terms

Let's build a linear regression model with body_acceptance as a predictor of post-treatment BMI the usual way.

The coefficient table tells us that with every step up the first order term of body_acceptance, we can expect 0.8 lower BMI post treatment (so the more satisfied the person is with their body at baseline, the lower their BMI will be at the end of the study, but be mindful that post-treatment BMI is probably also related to pre-treatment BMI, which might be already lower for people with higher body acceptance).

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT BMI_post_treatment
 /METHOD=ENTER body_acceptance.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,263[a]	,069	,065	3,921

a. Predictors: (Constant), body_acceptance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	271,963	1	271,963	17,690	,000[b]
	Residual	3659,033	238	15,374		
	Total	3930,996	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), body_acceptance

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	30,491	,438		69,575	,000	29,628	31,354
	body_acceptance	-,803	,191	-,263	-4,206	,000	-1,178	-,427

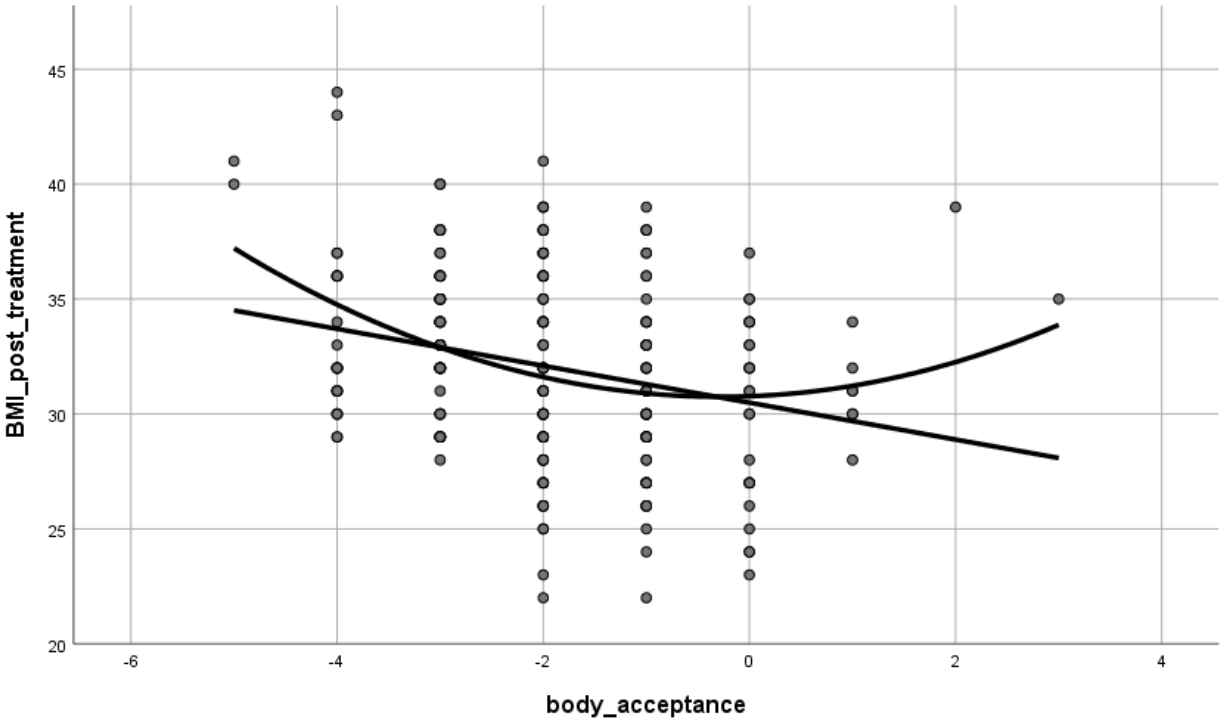
a. Dependent Variable: BMI_post_treatment

The output tells us that this model is significantly better than a null model ($F(1, 238) = 17.69$, $p < 0.001$), and that taking into account body acceptance adds significant predictive power to the model (this being the only predictor)(b = -0.80, beta = -0,263, $p < 0.001$). However, the variance explained by this model is mediocre, explaining only 7% of the variance (adj.R^2 = 0.065).

Let's explore this relationship with a scatterplot.

For example, here we can see that the relationship of post-treatment BMI and body_acceptance may not be entirely linear.

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=body_acceptance
BMI_post_treatment MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: body_acceptance=col(source(s), name("body_acceptance"))
  DATA: BMI_post_treatment=col(source(s), name("BMI_post_treatment"))
  GUIDE: axis(dim(1), label("body_acceptance"))
  GUIDE: axis(dim(2), label("BMI_post_treatment"))
  ELEMENT: point(position(body_acceptance*BMI_post_treatment))
  ELEMENT:
line(position(smooth.quadratic(body_acceptance*BMI_post_treatment)))
  ELEMENT: line(position(smooth.linear(body_acceptance*BMI_post_treatment)))
END GPL.
```



If you suspect that there is non-linear relationship between the outcome and some predictor, you can try to include a second or third order term.

So we build a model including the second order term of body acceptance, to account for this U-shaped relationship (quadratic relationship).

SPSS does not directly support entering higher order terms into regression models, so we will need to create a new variable to be able to build this model. The quadratic term is basically the sqared values of the original variable, so we can easily create such a variable in the **Transform > Compute** variable tab. The following formula will create the square of body_acceptance:

body_acceptance*body_acceptance

However, as we have seen in the model diagnostics exercise, entering such derived variables into regression models can create issues with multicollinearity, so we need to center body_acceptance first by subtracting the mean from this variable.

I named the new variables body_acceptance_centered and the quadratic term body_acceptance_quadratic_centered.

```
DESCRIPTIVES VARIABLES=body_acceptance
  /STATISTICS=MEAN STDDEV MIN MAX.


COMPUTE body_acceptance_centered=body_acceptance - -1.88.
EXECUTE.


COMPUTE
body_acceptance_centered_quadratic=body_acceptance_centered*body_accept
ance_centered.
EXECUTE.
```

Now let's build the linear regression model, where we use body_acceptance_centered AND its quadratic term as a predictor. (Unless you know what you are doing, **always add all the lower order terms in the model as well**.)

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT BMI_post_treatment
 /METHOD=ENTER body_acceptance body_acceptance_quadratic.
```

The coefficient table tells us that with every step up the first order term of body_acceptance, we can expect 0.93 lower BMI post treatment, so the more satisfied the person is with their body at baseline, the lower BMI we can expect, just like before. ($b = -0.93$, $beta = 0.30$, $p < 0.001$). We can also see from the coefficient of the quadratic value of body acceptance that the higher the quadratic term, the bigger the BMI, so the more extreme score the person has on body acceptance, the higher the BMI will be, and this predictor also has a significant unique predictive value in the model ($b = 0.29$, $beta = 0{,}187$, $p = 0.003$). It is not a big surprise that those who are extremely unsatisfied with their body have a higher BMI at post-treatment, because probably they also had higher BMI to begin with, but it is somewhat surprising that those who are extremely satisfied will also have higher BMI. Maybe this is because these people don't care that much about the therapy since they are already OK with their body, so they did not follow the therapeutic regimen so rigorously.

Also, the model now explains more variance than earlier ($F\ (2, 237) = 13.538$, $p < 0.001$, $adj.R^2 = 0.1$). (See the exercise on model selection to be able to tell whether this increase in the predictive effectiveness of the model is statistically significant or not.)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,320[a]	,103	,095	3,858

a. Predictors: (Constant), body_acceptance_centered_quadratic, body_acceptance_centered

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	403,043	2	201,522	13,538	,000[b]
	Residual	3527,953	237	14,886		
	Total	3930,996	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), body_acceptance_centered_quadratic, body_acceptance_centered

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	31,490	,303		104,070	,000	30,894	32,086		
	body_acceptance_centered	-,927	,192	-,304	-4,819	,000	-1,306	-,548	,952	1,050
	body_acceptance_centered_quadratic	,290	,098	,187	2,967	,003	,098	,483	,952	1,050

a. Dependent Variable: BMI_post_treatment

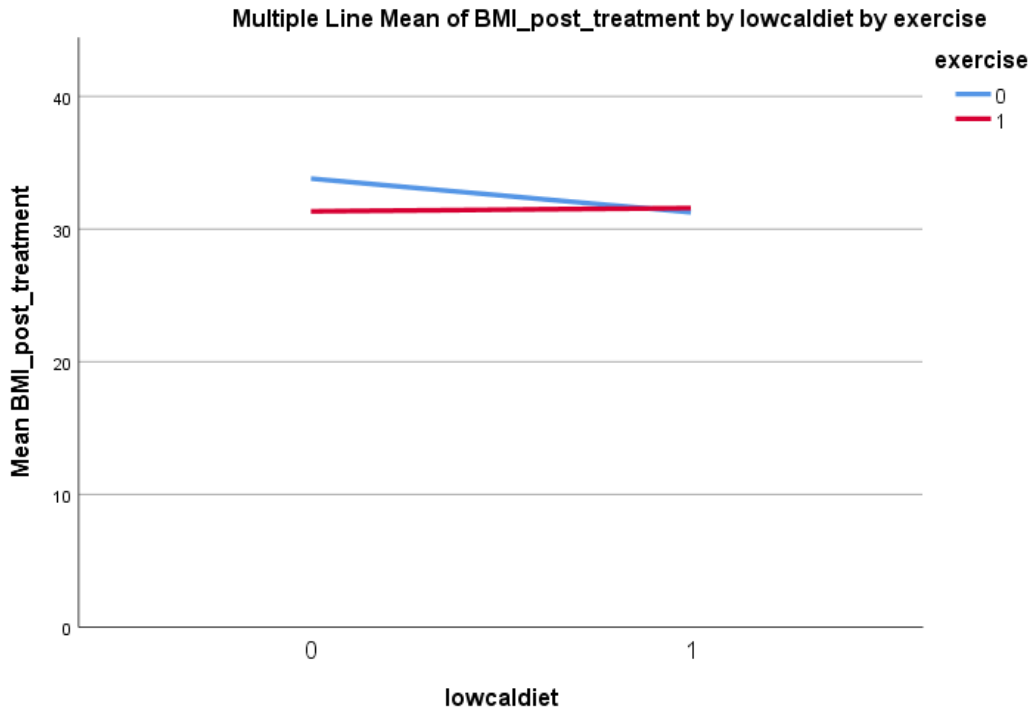
Interactions

A relationship of different predictors can also be modelled, if you suspect that the size or direction of the effect of one of the predictors depends on the other predictor's value.

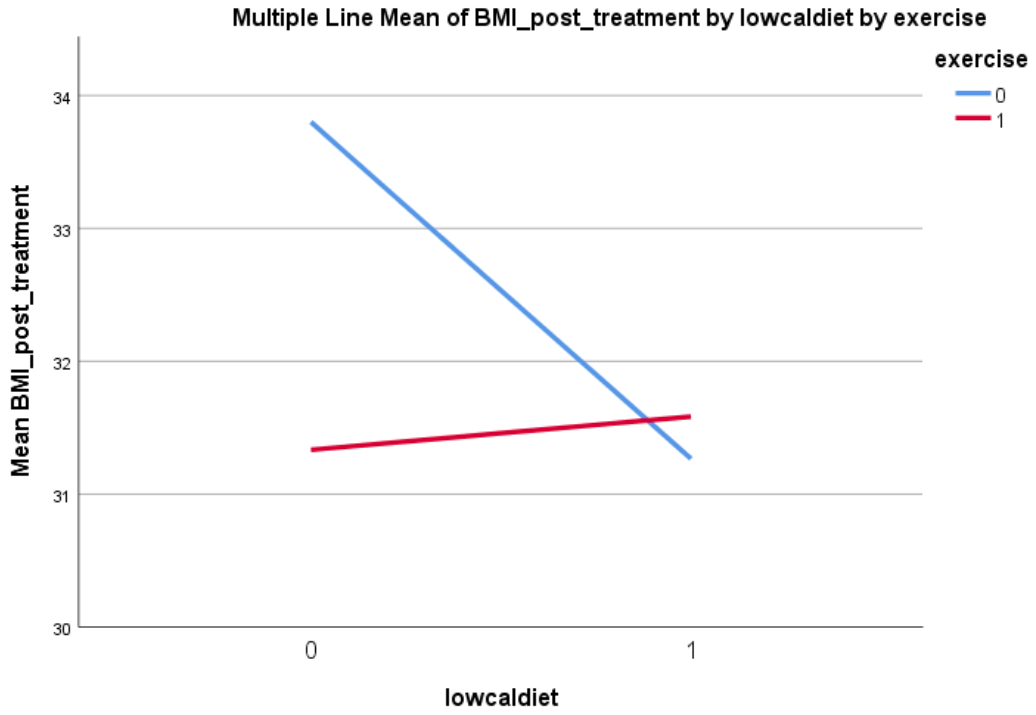
We will study the effect of a low calorie diet and exercise on post treatment BMI, and we will also look at the interaction of these two variables, whether they influence each other's effect on BMI.

We can explore the relationship of the effect of being on a low calorie diet and of doing exercise by plotting the means in all four possible groups (no diet or exercise, diet but no exercise, exercise but no diet, and both diet and exercise) in a line graph in the Chart builder, where we choose a multi-line linechart, enter BMI_post_treatment on the y axis, one of the factors on the x axis and the other factor as the determinant of the colors of the lines.

```
GGRAPH
  /GRAPHDATASET NAME="graphdataset" VARIABLES=received_pill
    MEAN(BMI_post_treatment)[name="MEAN_BMI_post_treatment"]
received_psychotherapy MISSING=LISTWISE
    REPORTMISSING=NO
  /GRAPHSPEC SOURCE=INLINE.
BEGIN GPL
  SOURCE: s=userSource(id("graphdataset"))
  DATA: received_pill=col(source(s), name("received_pill"), unit.category())
  DATA: MEAN_BMI_post_treatment=col(source(s),
name("MEAN_BMI_post_treatment"))
  DATA: received_psychotherapy=col(source(s), name("received_psychotherapy"),
unit.category())
  GUIDE: axis(dim(1), label("received_pill"))
  GUIDE: axis(dim(2), label("Mean BMI_post_treatment"))
  GUIDE: legend(aesthetic(aesthetic.color.interior),
label("received_psychotherapy"))
  SCALE: linear(dim(2), include(30))
  ELEMENT: line(position(received_pill*MEAN_BMI_post_treatment),
    color.interior(received_psychotherapy), missing.wings())
END GPL.
```



Here is the same graph zooming in on the relevant section of the graph:



It seems that the lines are **not parallel** to each other, which might **indicate an interaction** effect. The effect of diet seems obvious in the group who does not do exercise, those on the low calorie diet had lower BMI if they did not exercise. But those who exercised seem to have a BMI that is not affected by the low calorie diet. Maybe this is because of the muscle

mass gained due to the exercise, or because they supplement their diet in another way, or they are not able to keep a rigorous diet even though they claim to do it. So the effect of diet seems to depend on whether the person is doing regular exercise. It is also worth mentioning that those who do exercise already seem to be at a lower BMI without the diet.

Incorporating interaction effects in linear regression

The linear regression model is not optimized to include interaction terms, so we need to create a separate variable representing the interaction term. This is computed as the product of the values in the two predictor variables. In our case we will compute the product of lowcaldiet*exercise so we can use the **Transform > Compute** button to this product and save it into a new variable named INT_lowcaldiet_x_exercise.

```
COMPUTE INT_lowcaldiet_x_exercise=lowcaldiet*exercise.
EXECUTE.
```

Now we can build a linear regression model. As usual, go to **Analyze > Regression > Linear**, specify post-treatment BMI as the dependent, and lowcaldiet*exercise, and the newly created interaction variable, INT_lowcaldiet_x_exercise, as predictors.

```
REGRESSION
 /MISSING LISTWISE
 /STATISTICS COEFF OUTS CI(95) R ANOVA
 /CRITERIA=PIN(.05) POUT(.10)
 /NOORIGIN
 /DEPENDENT BMI_post_treatment
 /METHOD=ENTER lowcaldiet exercise INT_lowcaldiet_x_exercise.
```

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,259[a]	,067	,055	3,942

a. Predictors: (Constant), INT_lowcaldiet_x_exercise, exercise, lowcaldiet

ANOVA[a]

Model	Sum of Squares	df	Mean Square	F	Sig.

1	Regression	263,746	3	87,915	5,658	,001[b]
	Residual	3667,250	236	15,539		
	Total	3930,996	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), INT_lowcaldiet_x_exercise, exercise, lowcaldiet

The model is significantly better than the null model, explaining 6% of the variance of post-treatment BMI (F (3, 236) = 5.66, p = 0.001, adj.R^2 = 0.06).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95,0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	33,800	,509		66,417	,000	32,797	34,803
	lowcaldiet	-2,533	,720	-,313	-3,520	,001	-3,951	-1,115
	exercise	-2,467	,720	-,305	-3,427	,001	-3,885	-1,049
	INT_lowcaldiet_x_exercise	2,783	1,018	,298	2,735	,007	,778	4,788

a. Dependent Variable: BMI_post_treatment

The intercept is easy to interpret now that we know how to interpret the intercept of models with categorical predictors: the intercept of 33.8 indicates that if the value of both lowcaldiet, and exercise is zero (which incidentally also means that the value of lowcaldiet*exercise is zero), we can expect the person to have 33.8 BMI.

The parameter estimates (coefficients) should be interpreted as usual, with one little twist.

The coefficient of -2.53 of lowcaldiet for example means that if the value of lowcaldiet increases by 1 (from 0 to 1), this is the expected change in our estimated BMI. So a person who does not do exercise but does low calorie diet can be expected to have a 33.8 - 2.53 = 31.27 BMI at post treatment.

The coefficient of -2.47 of exercise can be interpreted the same way a person who does exercise but does not do low calorie diet can be expected to have a 33.8 - 2.47 = 31.33 at post treatment.

Similarly, the coefficient of 2.78 corresponding to INT_lowcaldiet_x_exercise is interpreted as usual: with a step up in the value of INT_lowcaldiet_x_exercise we can expect 2.78 increase in BMI. **But here comes the twist:** we have to keep in mind, that the value of INT_lowcaldiet_x_exercise is directly dependent on the values of lowcaldiet and exercise, so the only time the value of INT_lowcaldiet_x_exercise can be 1 is when the value of both lowcaldiet and exercise is 1. So the full interpretation of this coefficient is that with a step

up in the value of INT_lowcaldiet_x_exercise, we can expect 2.78 of change in our estimate of the outcome variable IN ADDITION TO the effects of the other predictors, so to get the expected BMI for a person who is doing both the diet and the exercise we need to compute: 33,800 -2.53 (the effect of diet) – 2.47 (the effect of exercise) + 2.78 (the effect of the interaction of the two). This should not come as a surprise, since this is just the common calculation what we would do if we used the regression equation, but it helps to clarify that the interaction effect is not the full effect of the two predictors when combined, rather, the additional effect or the adjustment to their effect when the two act in unison.

Interpreting complex interactions

The interpretation of the interaction term is always the same: the number you have to add to your equation when the product of the variables in the interaction increases by one.

When you have complex interactions it is always advisable to visualize the data to explore what the interaction really means. In the case of interactions including multiple variables, this usually required multiple charts.

Practice exercise

1 You can try how good you are at interpreting interactions by building another model, in which we predict post-treatment BMI with gender, motivation, and their interaction. (Remember, you might have to dummy code gender and compute the product of gender and motivation to be able to build a linear regression model like this in SPSS.

Components of models including interactions

Generally, you should always include all of the components within an interaction into the model as well. For example, if you are interested in the interaction of treatment x time x therapeutic_aliance, your model should include:

Treatment

Time

therapeutic_aliance

treatment x time

therapeutic_aliance x time

treatment x therapeutic_aliance

treatment x time x therapeutic_aliance

Two-way ANOVA

You can test interactions with a two-way anova as well. *As stated before, the ANOVA is just a special case of the linear regression, so if you are OK with using regression for this analysis, you don't have to learn the ANOVA approach, and all predictions and significance levels will be the same in the output.*

The 2x2 ANOVA test can be found in the **Analyze > General Linear Model > Univariate** tab in SPSS. We should enter post-treatment BMI as the dependent, and the lowcaldiet and exercise variables as fixed factors. In the options menu we should ask for display means of all the components in the model, homogeneity tests, estimates of effect size, and parameter estimates. Also, in the plots menu we can reproduce the same plot as above.

```
UNIANOVA BMI_post_treatment BY exercise lowcaldiet
  /METHOD=SSTYPE(3)
  /INTERCEPT=INCLUDE
  /PLOT=PROFILE(lowcaldiet*exercise)
  /EMMEANS=TABLES(exercise*lowcaldiet)
  /EMMEANS=TABLES(exercise)
  /EMMEANS=TABLES(lowcaldiet)
  /PRINT=ETASQ PARAMETER HOMOGENEITY
  /CRITERIA=ALPHA(.05)
  /DESIGN=exercise lowcaldiet exercise*lowcaldiet.
```

First of all, in the Between Subjects Effects table in the corrected model line we can see that this model is significantly better than the null model in explaining the variability in BMI (F (3, 236) = 5.66, p = 0.001, partial eta squared = 0.067). Notice that the partial eta squared effect size measure is identical to the R^2 in the regression. Below that, we can see in the lowcaldiet and the exercise lines that both of these factors have a main effect, that is a unique predictive value in the model (see table below). In the line corresponding to exercise * lowcaldiet, we see a significant p-value, indicating that have enough evidence to support the existence of an interaction effect (you can find these information in the parameter estimation table). The parameter estimates in the parameter estimation table are different from those of the regression coefficients because SPSS by default uses a different baseline level for the ANOVA than for the regression (don't ask why), but if we were to flip the coding of the predictors (reverse coding them), the parameter estimates would match that of the regression analysis.

Tests of Between-Subjects Effects

Dependent Variable: BMI_post_treatment

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	263,746[a]	3	87,915	5,658	,001	,067

Intercept	245696,004	1	245696,004	15811,373	,000	,985
exercise	69,338	1	69,338	4,462	,036	,019
lowcaldiet	78,204	1	78,204	5,033	,026	,021
exercise * lowcaldiet	116,204	1	116,204	7,478	,007	,031
Error	3667,250	236	15,539			
Total	249627,000	240				
Corrected Total	3930,996	239				

a. R Squared = ,067 (Adjusted R Squared = ,055)