Exercise 17 - Multiple regression – different types of predictors

Zoltan Kekecs

13 May 2020

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 from github.

https://github.com/kekecsz/SIMM32_2019_spring/blob/master/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 (fake) data. IT is about a study where different types of interventions were tested to help overweight people to lose 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

treatment 3 – a third kind of treatment (see below)

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.

The easiest way to look at the effect of treatment type on post-treatment BMI is to run a one-way ANOVA.

Our treatment type variable is a string variable (text), but SPSS does not allow string variables to be used in most statistical test, including one-way ANOVAs. So we need to recode our variable to be numeric.

We can use the auto-recode feature to do this 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.
```

One-way ANOVA

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 ($F(3, 236) = 26.51$, $p < 0.001$), meaning that at least one of the groups showed significantly different mean than at least one of the other groups.

ANOVA

BMI_post_treatment

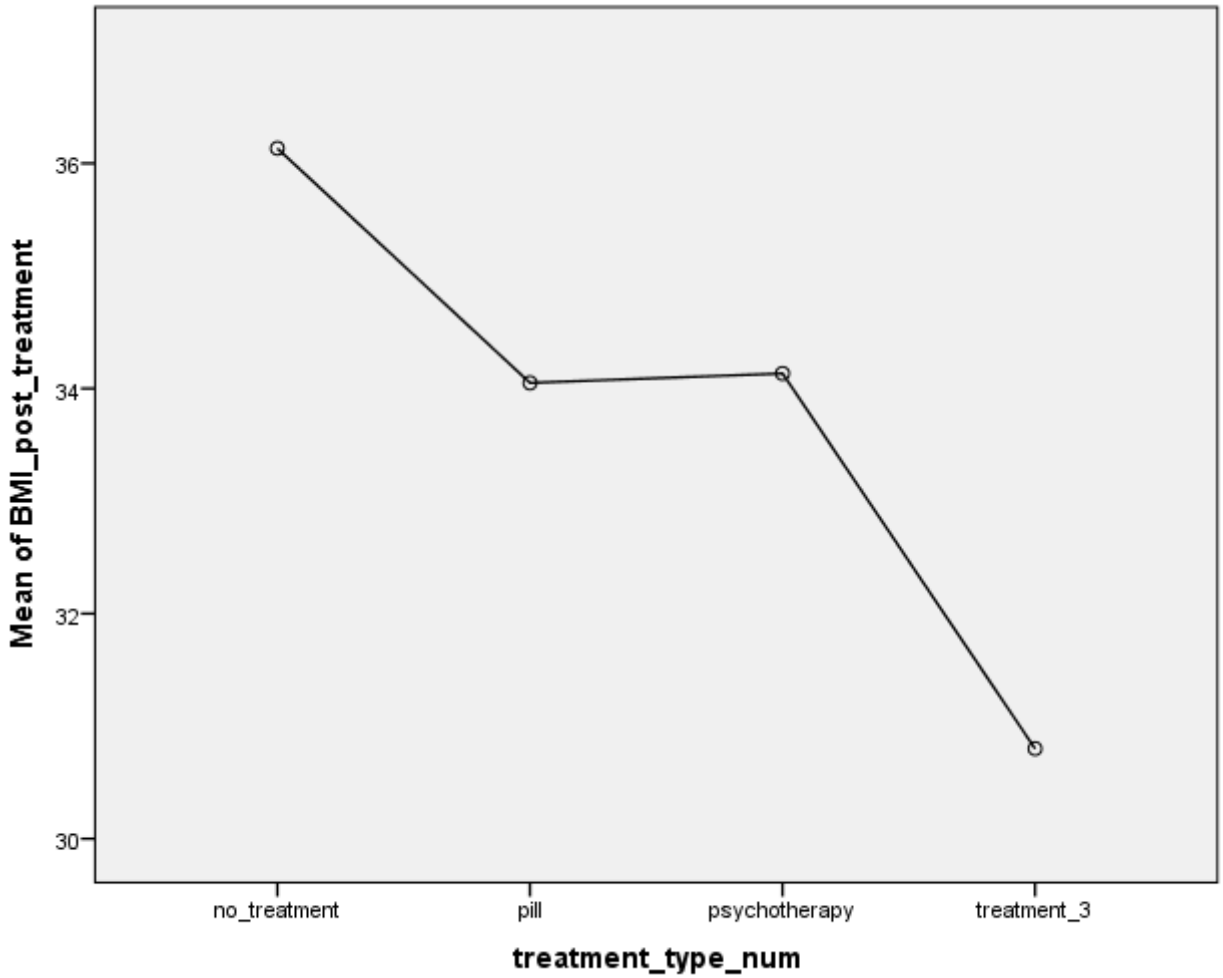
	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	876,979	3	292,326	26,511	,000
Within Groups	2602,317	236	11,027		
Total	3479,296	239			

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.

Descriptives

BMI_post_treatment

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
no_treatment	60	36,13	3,491	,451	35,23	37,04	27	44
pill	60	34,05	2,954	,381	33,29	34,81	25	39
psychotherapy	60	34,13	3,397	,439	33,26	35,01	25	41
treatment_3	60	30,80	3,414	,441	29,92	31,68	22	39
Total	240	33,78	3,815	,246	33,29	34,26	22	44



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 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, every other comparison was significant, so pill is better than no treatment, but less good than treatment 3, and psychotherapy is better than no treatment, but less good than treatment 3, and we can't really tell whether psychotherapy or pill is 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	95% Confidence Interval Upper Bound
no_treatment	pill	2,083*	,606	,004	,47	3,70
	psychotherapy	2,000*	,606	,007	,39	3,61
	treatment_3	5,333*	,606	,000	3,72	6,95
pill	no_treatment	-2,083*	,606	,004	-3,70	-,47
	psychotherapy	-,083	,606	1,000	-1,70	1,53
	treatment_3	3,250*	,606	,000	1,64	4,86
psychotherapy	no_treatment	-2,000*	,606	,007	-3,61	-,39
	pill	,083	,606	1,000	-1,53	1,70
	treatment_3	3,333*	,606	,000	1,72	4,95
treatment_3	no_treatment	-5,333*	,606	,000	-6,95	-3,72
	pill	-3,250*	,606	,000	-4,86	-1,64
	psychotherapy	-3,333*	,606	,000	-4,95	-1,72

*. 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.

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 each levels which are not chosen as the baseline. 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_treatment_3 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 ('treatment_3'=1) (ELSE=0) INTO treatment_treatment_3.

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_treatment_3 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_treatment_3.
```

Now when we compare this anova table in the regression output with the ANOVA table produced by the one-way ANOVA output, we will notice that they contain identical numbers.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	876,979	3	292,326	26,511	,000[b]
	Residual	2602,317	236	11,027		
	Total	3479,296	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), treatment_treatment_3, treatment_psychotherapy, treatment_pill

Also, the numbers in the coefficients table should be familiar from the one-way ANOVA output. The coefficient for the intercept is identical to the mean post-treatment BMI of the no-treatment group (shown in the Descriptives table in the One-way ANOVA output). And the coefficients for the pill, psychotherapy, and treatment_3 predictors are the same as the mean differences shown in the post-hoc comparisons of these groups with the no_treatment group in the One-way ANOVA output.

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)	36,133	,429		84,287	,000	35,289	36,978

	treatment_pill	-2,083	,606	-,237	-3,436	,001	-3,278	-,889
	treatment_psychotherapy	-2,000	,606	-,227	-3,299	,001	-3,194	-,806
	treatment_treatment_3	-5,333	,606	-,607	-8,797	,000	-6,528	-4,139

a. Dependent Variable: BMI_post_treatment

Basically this tells us how to interpret the coefficients in the regression output when we use categorical predictors: the coefficient of the constant (the intercept) tells us the expected value of the dependent variable at the baseline level (or default level) of the categorical variable (in our case, no_treatment). The coefficients at the dummy variables tell us the expected effect of the given level compared to the baseline level, or the difference between the given level of the categorical variable and the baseline level. In our case if a person gets the pill treatment they are expected to have 2.08 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_treatment_3 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 are always: with every other variable held constant, moving one-step up on the scale of the predictor variable would produce this expected change/difference in the dependent variable: that is, if a person reported 3 instead of 2 on the motivation scale, we expect that their BMI would be 0.05 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.

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)	36,425	,934		39,001	,000	34,585	38,265
	treatment_pill	-2,081	,607	-,237	-3,426	,001	-3,278	-,884
	treatment_psychotherapy	-1,997	,607	-,227	-3,287	,001	-3,193	-,800
	treatment_treatment_3	-5,312	,610	-,604	-8,701	,000	-6,514	-4,109
	motivation	-,050	,141	-,020	-,352	,725	-,328	,229

a. Dependent Variable: BMI_post_treatment

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 0.37 higher BMI post treatment (so the less satisfied the person is with their body at baseline, the more effective the treatment is expected to be).

```
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	,251[a]	,063	,059	3,701

a. Predictors: (Constant), body_acceptance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	218,756	1	218,756	15,968	,000[b]
	Residual	3260,540	238	13,700		
	Total	3479,296	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)	32,696	,361		90,490	,000	31,984	33,408
	body_acceptance	-,598	,150	-,251	-3,996	,000	-,892	-,303

a. Dependent Variable: BMI_post_treatment

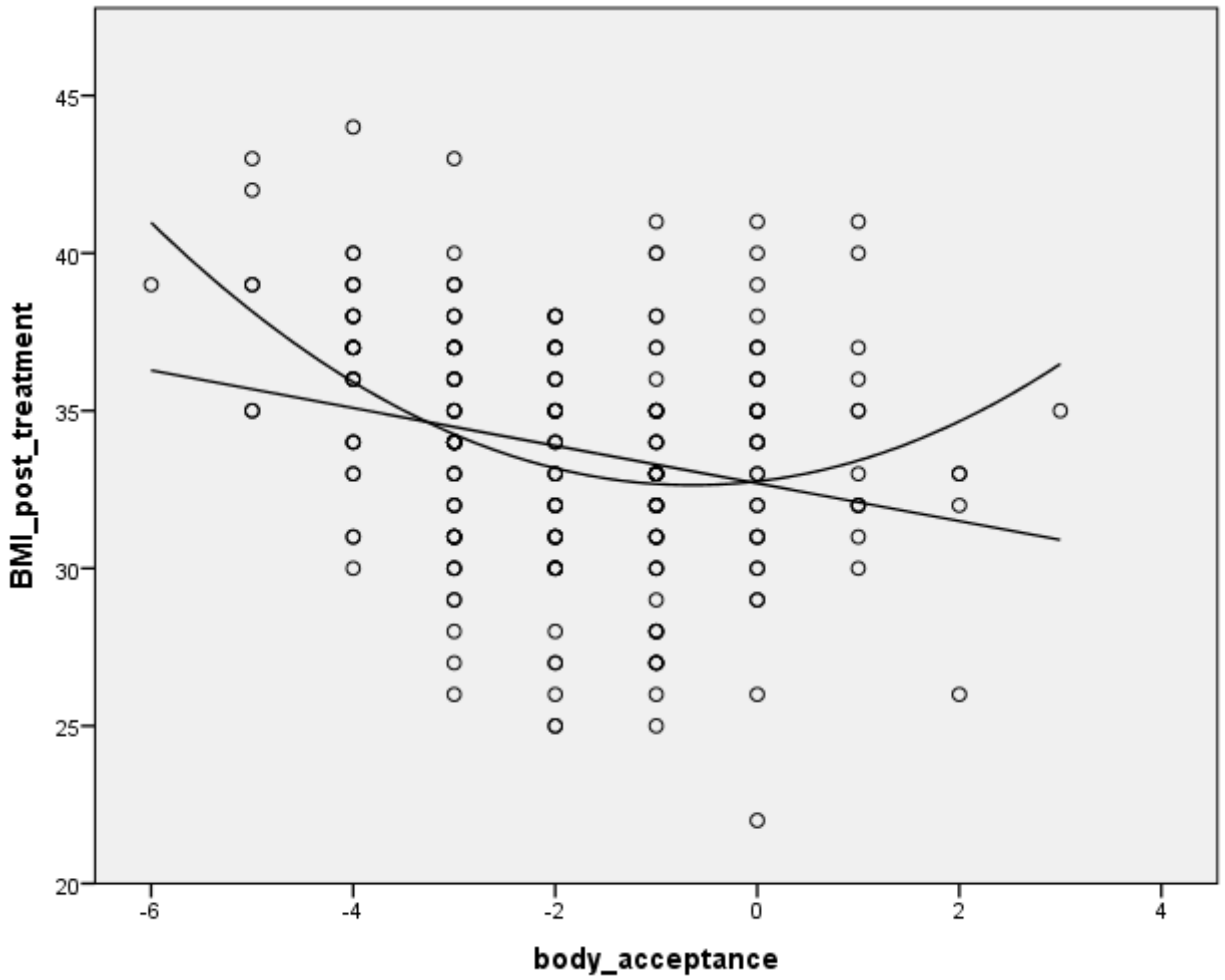
The output tells us that this model is significantly better than a null model (F (1, 238) = 15.97, p < 0.001), and that taking into account body acceptance adds significant predictive power to the model (this being the only predictor)(b = -0.60, beta = -0,25, p < 0.001). Specifically, the lower the body acceptance, the higher the BMI at post-treatment it seems. However, the variance explained by this model is mediocre, explaining only 6% of the variance (adj.R^2 = 0.059).

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 voriginal 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

I named the new variable body_acceptance_quadratic

```
COMPUTE body_acceptance_quadratic=body_acceptance*body_acceptance.
EXECUTE.
```

Now let's build the linear regression model, where we use body_acceptance 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 0.37 higher BMI post treatment (so the less satisfied the person is with their body at baseline, the more effective the treatment is expected to be). Note that the direction of the coefficient actually changed from negative to positive by adding the second order term, and it seem to have lost its unique predictive value in the model ($b = -0.37$, $beta = 0.16$, $p = 0.18$). We can also see from the coefficient of the higher the quadratic term, the bigger the BMI, so the more extreme score the person has on body acceptance, the less effective the treatment seems to be, and this predictor has a significant unique predictive value in the model ($b = 0.29$, $beta = 0,48$, $p < 0.001$). Also, the model now explains more variance than earlier ($F\ (2, 239) = 16.961$, $p < 0.001$, $adj.R^2 = 0.12$). (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	,354[a]	,125	,118	3,584

a. Predictors: (Constant), body_acceptance_quadratic, body_acceptance

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	435,649	2	217,824	16,961	,000[b]
	Residual	3043,647	237	12,842		
	Total	3479,296	239			

a. Dependent Variable: BMI_post_treatment

b. Predictors: (Constant), body_acceptance_quadratic, body_acceptance

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	32,760	,350		93,552	,000	32,070	33,450
	body_acceptance	,372	,277	,156	1,344	,180	-,173	,917
	body_acceptance_quadratic	,290	,071	,477	4,110	,000	,151	,429

a. Dependent Variable: BMI_post_treatment

Interactions

A relationship of different predictors can also be modelled, if you suspect that the effect of one of the predictors depends on the other predictor. We are going to address that in this section.

The mysterious “treatment_3” we worked with in the earlier example is actually the combination of pill and psychopterapy. If we know that, we can now imagine this study as a 2x2 factorial design, where not only the effect of receiving the appetite control pill, and the

effect of receiving the psychotherapy treatment can be tested, but also the combined effect (interaction) of the two treatments.

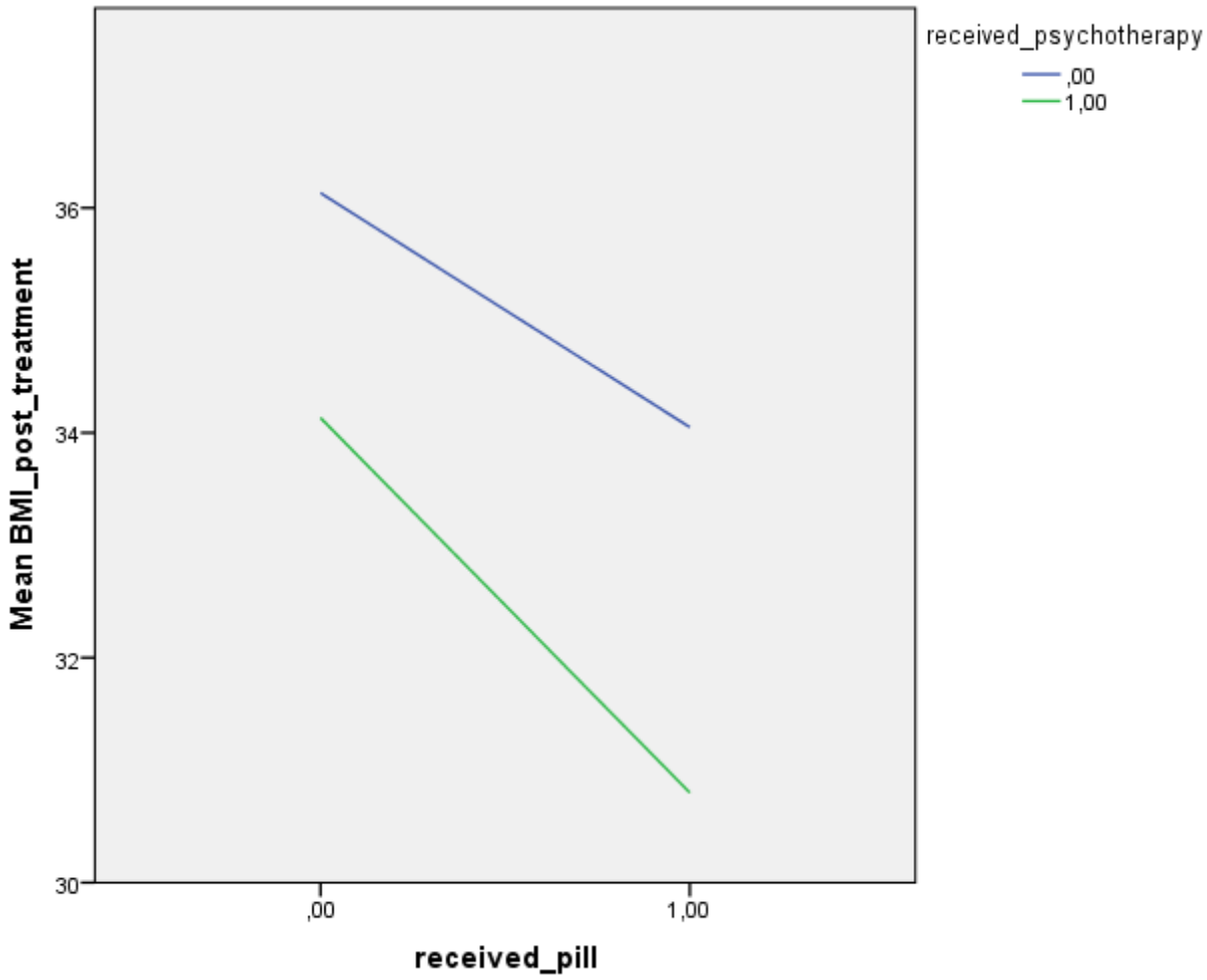
To perform this analysis, we will need to recode out variable treatment_type once again. Now that we know that treatment_3 means, that the person got both the pill and psychotherapy. In **Transform > Recode into different variable** we can create a variable called received_pill by recoding “pill” and “treatment_3” as 1, and every other value as 0 within the variable treatment_type. Similarly, we can create a variable called received_psychotherapy by recoding “psychotherapy” and “treatment_3” as 1, and every other value as 0 within the variable treatment_type.

```
RECODE treatment_type ('pill'=1) ('treatment_3'=1) (ELSE=0) INTO received_pill.

EXECUTE.

RECODE treatment_type ('psychotherapy'=1) ('treatment_3'=1) (ELSE=0) INTO
received_psychotherapy.

EXECUTE.
```

We can explore the relationship of the effect of receiving pill and psychotherapy treatment by plotting the means in all four groups 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.
```



It seems that the lines are not completely parallel to each other, which might indicate an interaction effect: the effects of pill and psychotherapy are not simply additive, but the combined effect seems to be even stronger than the effect of the individual treatments added together.

Two-way ANOVA

The most straight-forward way of testing this is to use a 2x2 ANOVA. This 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 new received_pill and received_ psychotherapy 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 isze, and parameter estimates. Also, in the plots menu we can reproduce the same plot as above.

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

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) = 26.51, p < 0.001, partial eta^2 = 0.25). If this wasn't true, there would be not too much sense in interpreting the individual effects of the predictors in the model. Below that, we can see in the received_pill and the received_psychotherapy 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 received_pill * received_psychotherapy, we see a non.significant p-value, indicating that we don't have enough evidence to support the existence of an interaction effect (b = -1.25 [-2.94, 0.44], t = -1.46, p = 0.146, partial eta^2 = 0.01)(you can find these information in the parameter estimation table).

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	876,979[a]	3	292,326	26,511	,000	,252
Intercept	273847,704	1	273847,704	24834,817	,000	,991
received_pill	440,104	1	440,104	39,912	,000	,145
received_psychotherapy	413,438	1	413,438	37,494	,000	,137
received_pill * received_psychotherapy	23,438	1	23,438	2,126	,146	,009
Error	2602,317	236	11,027			
Total	277327,000	240				

Corrected Total	3479,296	239				

a. R Squared = ,252 (Adjusted R Squared = ,243)

The partial eta squared for the whole model is 0.25, which is interpreted the same way as the R^2, that is, 25% of the variability in BMI is explained by taking into account all the predictors in the model.

The same with regression

As we have seen with the One-Way ANOVA, the 2x2 ANOVA is also just a special case of linear regression, and we can get the exact same information by using a linear regression model.

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 the product of received_pill * received_psychotherapy. so we can use the Transform > Compute button to compute this product and save it into a new variable named INT_pill_x_psychotherapy.

```
COMPUTE INT_pill_x_psychotherapy=received_pill * received_psychotherapy.
EXECUTE.
```

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

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

As before, we can see that similar numbers appear as in the 2x2 ANOVA above:

The R squared is identical to the partial eta^2 (R^2 = 0.25), and the model F-test is the same (F (3, 236) = 26.51, p < 0.001, adj.R^2 = 0.24).

The parameter estimates returned by the ANOVA are slightly different from the model coefficients produced by the regression though. In order to understand why, we need to correctly interpret them:

Coefficients table from Regression

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	36,133	,429		84,287	,000	35,289	36,978
	received_pill	-2,083	,606	-,274	-3,436	,001	-3,278	-,889
	received_psychotherapy	-2,000	,606	-,263	-3,299	,001	-3,194	-,806
	INT_pill_x_psychotherapy	-1,250	,857	-,142	-1,458	,146	-2,939	,439

a. Dependent Variable: BMI_post_treatment

Paramater estimates table from ANOVA

Parameter Estimates

Dependent Variable: BMI_post_treatment

Parameter	B	Std. Error	t	Sig.	95% Confidence Interval		Partial Eta Squared
					Lower Bound	Upper Bound	
Intercept	30,800	,429	71,846	,000	29,955	31,645	,956
[received_pill=,00]	3,333	,606	5,498	,000	2,139	4,528	,114
[received_pill=1,00]	0[a]	.	.	.	.	.	.
[received_psychotherapy=,00]	3,250	,606	5,361	,000	2,056	4,444	,109
[received_psychotherapy=1,00]	0[a]	.	.	.	.	.	.
[received_pill=,00] * [received_psychotherapy=,00]	-1,250	,857	-1,458	,146	-2,939	,439	,009
[received_pill=,00] * [received_psychotherapy=1,00]	0[a]	.	.	.	.	.	.
[received_pill=1,00] * [received_psychotherapy=,00]	0[a]	.	.	.	.	.	.
[received_pill=1,00] * [received_psychotherapy=1,00]	0[a]	.	.	.	.	.	.

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

The only reason for the difference, is that we use a different “default level”, a level of factors to which the effects are compared.

In the regression analysis, the default level (or baseline level) to which everything is compared is the lowest level of each factor, received pill = 0 and received psychotherapy = 0. At this default level, we expect people to have 36.13 BMI on average. The coefficient for received pill means, that if I move one 1 point up on the scale of received pill (from 0 to 1) while to other parameters remain constant (received psychotherapy = 0), we can expect 2.08 points of decrease in BMI, (34.05). The coefficient for psychotherapy tells us the same thing: compared to a person not getting anything, a person who gets psychotherapy, but the other variables kept constant (did not receive pill treatment), we can expect a 2 point lower BMI (34.13). The interaction term’s coefficient shows how to change our expectations if this interaction variable changes from 0 to 1. Because this is the product of the received_pill and received_psychotherapy variables, the interaction term can only be 1, if both the received_pill and received_psychotherapy are also 1. In this case, the coefficient tells us that we should expect an additional 1.25 points of decrease in BMI, on top of the effect of a decrease of 2.08 due to getting pills, and a decrease of 2 due to getting psychotherapy. So in total, for a person who gets both pills and psychotherapy, we would expect that compared to the default level, the BMI would be 36.13 – 2.08 – 2 – 1.25 = 30.8. Notice, that these predictions are exactly matching the predicted marginal means in the first Estimated Marginal Means table in the ANOVA output.

1. received_pill * received_psychotherapy

Dependent Variable: BMI_post_treatment

				95% Confidence Interval	
received_pill	received_psychotherapy	Mean	Std. Error	Lower Bound	Upper Bound
,00	,00	36,133	,429	35,289	36,978
	1,00	34,133	,429	33,289	34,978
1,00	,00	34,050	,429	33,205	34,895
	1,00	30,800	,429	29,955	31,645

So how can be the ANOVA estimates be different than the Regression coefficients, and still return the same final estimations? The reason is the different default level. For some reason, the ANOVA takes 1 on both variables as the default level. This means that we are computing the effects of NOT getting one, or the other or both treatments compared to a default state where people get both treatments. So there, we can expect 3.33 higher BMI if a person does not get pills compared to if the person goth both treatments, and we can expect 3.25 points of higher BMI for a person who does not get psychotherapy compared to if the person goth both treatments. Now if they got neither one of the treatments, first we would estimate the BMI by taking into account the effect of NOT getting pills and NOT getting psychotherapy, and add that to the BMI estimated at the default level. However, it turn out that if a person does not get one of the treatments already, it does not make that

big of a difference if they did not receive the other treatment either, compared to the raw effects of not getting the two treatments added together. So here, the interaction effect actually “decreases” the impact of not getting both treatments, that is why it is negative, just like in the case of the regression coefficient (-1.25).

If we recoded out variables so that the variables represent NOT getting the treatment as 1 and getting the treatment as 0, we would get the exact same parameter estimates (coefficients) in both the regression and ANOVA analyses.

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.

You can try how good you are at interpreting this 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.

If you want to use a two-way ANOVA instead, you will have to enter motivation as a “covariate”, and specify in the Model menu that you want the main effect of the dummy gender variable, motivation, and the interaction of these to variables.

The interpretation remains the same even if you enter multiple variables into the interaction term. For example if we get a coefficient 0.05 for the interaction term of gender, motivation, treatment_pill, and treatment_psychotherapy, this means that when the product of these numerical variables increases by one point, you will have to add 0.05 to the estimated outcome.

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.

Components of models including interactions

Generally, you should always include all of the components within an interaction into the model as well. For example, 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