

Interaction Example

Say we are interested in the effects of publication pressure, and the motivational focus of researchers on the average quality of the papers they produce. With these three variables we will build a model with one dependent variable, two predictor variables, and an interaction between the two predictor variables. Plots are by far the easiest way to interpret interactions, so that is what we will mainly use here.

In this (completely made-up) example, average quality papers is our dependent variable. It is some continuous measure based on things like how reproducible the results in the paper are, the information value of the paper, the readability of the paper, etc. The higher the score from 1 - 100, the higher the average quality of the researcher's papers.

Publication pressure is the amount of pressure the researcher experiences to publish a lot of papers – this could be from demands to get a tenured position, their boss, the requirements to obtain grants, or the demands to get recognition in academia, etc. It is measured with a categorical variable with two levels, 0 low pressure and 1 high pressure. Publication pressure will serve as a predictor variable.

Motivational focus is another categorical predictor variable, with two categories (also clearly oversimplified, but it will have to do for this example). Category 0 is 'truth-finding': A researcher in this category is mainly motivated by 'truth-finding', that is, wanting to figure out how the world truly works. Category 1 is 'status': A researcher in this category is mainly focused on obtaining a high 'status', that is, becoming successful in terms of earning a lot of money, getting recognition, and/or obtaining power.

Here is the regression equation for this model:

$$\text{Av_quality papers} = \text{int} + B1 \cdot \text{pubpress} + B2 \cdot \text{motiv} + B3 \cdot \text{pubpress} \cdot \text{motiv} + \text{error}$$

Imagine that in reality, we have the following parameters for this equation:

$$\text{int} = 80$$

$$B1 = -8$$

$$B2 = -30$$

$$B3 = -20$$

And the errors are normally distributed with a mean of zero, and standard deviation of 5.

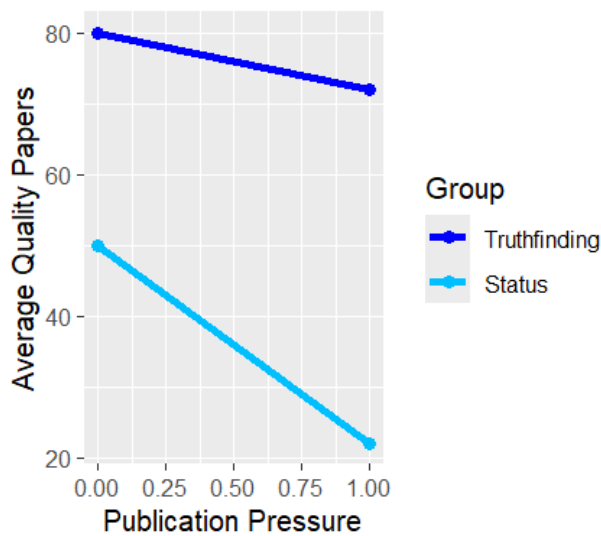
Such that we get:

$$\text{Av_quality papers} = 80 - 8 \cdot \text{pubpress} - 30 \cdot \text{motiv} - 20 \cdot \text{pubpress} \cdot \text{motiv} + \text{error}$$

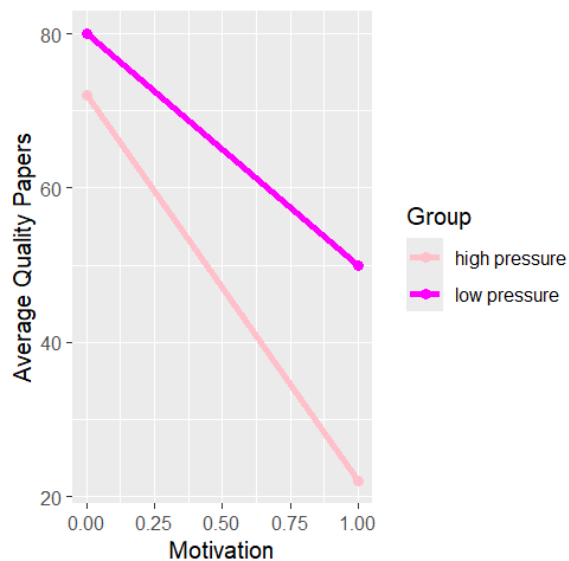
That is in reality, say we have two main effects here on Average quality papers, namely:

- An effect of publication pressure on quality, where the higher pressure group is expected to have a lower quality than the low pressure group.
- An effect of Motivational focus on quality, where we see that overall, the Truth finding group produces on average higher quality papers than the Status group.

There is also an interaction between Publication pressure and Motivational focus. This means that the relationship between publication pressure and average paper quality differs depending on the motivational focus of the researcher, as can be seen from the figure below.



Note: We could also flip this around and say that the relationship between motivational focus and average paper quality differs depending on the level of pressure. We can flip the plot around accordingly:



In both pictures the points are the (true) expected values for the average quality of the papers, for the different levels of publication pressure, and the different motivational foci of the researchers. Note that the numbers plotted are identical (just the x axis and grouping variable are flipped).

By writing out the regression equation for each of the possible groups, and filling in the scores for the predictor variables and the regression coefficients, you get the expected values (predicted values) for each group that you see in the plots I made. This is a lot of work, **but can be helpful to see what particular interaction terms do:**

Group: Low pressure 0, truth finding 0

$$\text{Expected value } Av_qual_papers = int + B1*0 + B2*0 + B3*0*0$$

$$= int$$

That is, 80 is the expected quality for the low pressure truth finding group.

Group: high pressure 1, truthfinding 0

$$\text{Expected value } Av_qual_papers = int + B1*1 + B2*0 + B3*1*0$$

$$= int + B1$$

That is, $80 - 8 = 72$ is the expected value for the high pressure truthfinding group.

Group: Low pressure 0, status 1

$$\text{Expected value } Av_qual_papers = int + B1*0 + B2*1 + B3*0*1$$

$$= int + B2$$

That is, $80 - 30 = 50$ is the expected value for the low pressure status group.

Group: high pressure 1, status 1

$$\text{Expected value } Av_qual_papers = int + B1*1 + B2*1 + B3*1*1$$

$$= int + B1 + B2 + B3$$

That is, $80 - 8 - 30 - 20 = 22$ is the expected value for the high pressure status group.

That is, the interaction effect essentially results in an extra 'quality penalty' of 20 for the high pressure-status group compared to the other three groups. If in reality there was no interaction effect the EV for this group would be 42; and the EVs for the other groups would be stay as they were above.

True (population) parameters versus estimated regression coefficients

Note, in this conceptual example of a 2-way interaction, we knew we had a 2-way interaction in reality, and had access to the population model parameters. This makes things simple, because we didn't have to do any estimation, or worry about significance. Of course, in practice we do not have the population parameters (otherwise we wouldn't need to do research), instead we would collect data and analyze the data to estimate the population parameters. Based on the estimated parameters and the data, we would make plots similar to those above. Such plots would still be the main way to interpret (significant) interaction effects, but there would be some uncertainty about the expected values.

For example, say I collected data from about 500 participants, divided among the different groups. Here are the results, the estimated coefficients, standard errors, t-values, and p-values, for a random sample of N=500 (estimated in R):

```
> twowayinteractions=lm(Av_qual_papers ~ pubpress + motiv + pubpress*motiv,data=dataf)
```

```
> summary(twowayinteractions)
```

Call:

```
lm(formula = Av_qual_papers ~ pubpress + motiv + pubpress * motiv,
    data = dataf)
```

Residuals:

```
      Min       1Q   Median       3Q      Max
-16.2982 -3.3231 -0.0874  3.3060 16.5151
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   80.2995    0.5144  156.12 <2e-16 ***
pubpress      -7.8024    0.6672  -11.69 <2e-16 ***
motiv        -30.4330    0.7238  -42.05 <2e-16 ***
pubpress:motiv -19.6506    0.9346  -21.02 <2e-16 ***
---
```

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.118 on 496 degrees of freedom

Multiple R-squared: 0.9548, Adjusted R-squared: 0.9545

F-statistic: 3493 on 3 and 496 DF, p-value: < 2.2e-16

Model results:

Residual standard error: 4.914 on 488 degrees of freedom

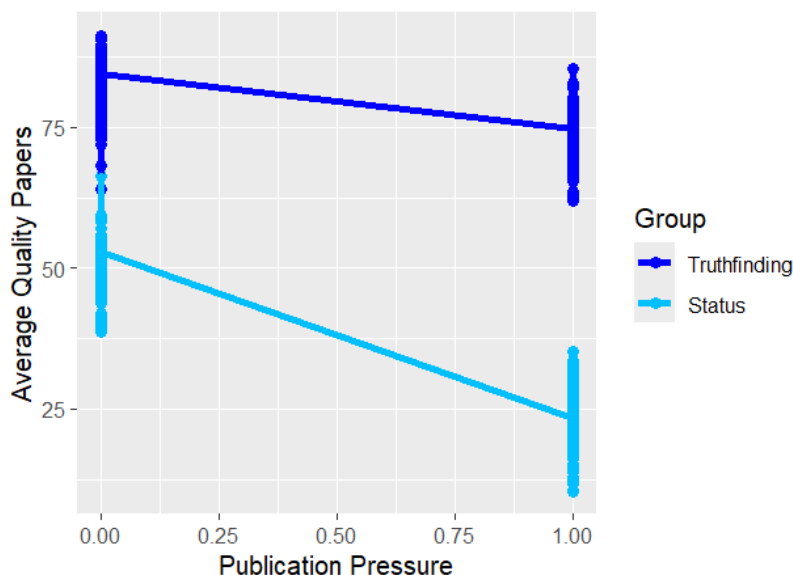
Multiple R-squared: 0.9651, Adjusted R-squared: 0.9643

F-statistic: 1226 on 11 and 488 DF, p-value: $< 2.2e-16$

Note that the p-values are almost all very small. This is because it is a large dataset and the effects I made up for the population are quite large. The population parameters are estimated back quite well because of this too (compare the estimated coefficients to the true coefficients I presented earlier).

To help interpret the results you could make plots of the predicted values for each group, as I did earlier for the population values. For example by saving the predicted scores in SPSS based on our regression analysis, and then making plots out of those. You could of course also make plots of the observed data instead of the predicted values.

Below I have plotted the simulated data (n=500) for this example:



What about interactions with continuous variables

If you have continuous variables, the upside is that you do not have to mess around with dummy variables. The downside is that it is a bit harder to make plots like the above (or do the equations like we did), where we made nice separate lines for different groups to highlight the effects of the moderating variables. Typically, however, people do make exactly those kind of plots for continuous variables (and work with those equations in the same way!). This is sometimes called 'simple slope analysis'. Basically, what you do is for your continuous predictors, you pick a bunch of values, and for those specific values you make plots like the 2d plots we made before (or do the equations). Imagine for example that publication pressure was a continuous variable ranging from 0 to 10. What people do is then pick two or more values out of that range, and make plots for those values. So, we could pick a score of 2 and of 8 to represent 'little pressure' and 'lots of pressure' and make exactly the kinds of 2d plots you saw before (or do the equations for scores of 2 and 8).