# What is multivariate regression analysis and how is it used in Mplus data analysis?

Authored by

**stats writer**

June 29, 2024

Multivariate regression analysis is a statistical technique used to examine the relationship between multiple independent variables and a single dependent variable. In Mplus data analysis, this method is used to determine the extent to which a set of independent variables predicts or influences a particular outcome. It allows researchers to understand the complex relationships between variables and identify which factors have the strongest impact on the dependent variable. This technique is particularly useful when analyzing large datasets with multiple variables, as it can provide more comprehensive insights and help identify important predictors. The results of multivariate regression analysis are typically presented in the form of regression coefficients, which indicate the strength and direction of the relationship between each independent variable and the dependent variable. Overall, multivariate regression analysis is a valuable tool in Mplus data analysis for understanding the complex relationships between variables and making informed decisions based on statistical evidence.

## Multivariate Regression Analysis | Mplus Data Analysis Examples

## Note: This example was done using Mplus version 5.2. The syntax may not work, or may function differently, with other versions of Mplus.

## As the name implies, multivariate regression is a technique that estimates a single regression model with more than one outcome variable. When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.

**Please note: The purpose of this page is to show how to use various data analysis commands.**
**It does not cover all aspects of the research process which researchers are expected to do. In**
**particular, it does not cover data cleaning and checking, verification of assumptions, model**
**diagnostics and potential follow-up analyses.**

## Examples of multivariate regression analysis

**Example 1. A researcher has collected data on three psychological variables,**
**four academic variables (standardized test scores), and the type of educational**
**program the student is in for 600 high school students. She is interested in how**
**the set of psychological variables is related to the academic variables and the**
**type of program the student is in.**

**Example 2. A doctor has collected data on cholesterol, blood pressure, and**
**weight. She also collected data on the eating habits of the subjects**
**(e.g., how many ounces of red meat, fish, dairy**

products, and chocolate consumed
per week). She wants to investigate the relationship between the three
measures of health and eating habits.

Example 3. A researcher is interested in determining what factors influence
the health African Violet plants. She collects data on the average leaf
diameter, the mass of the root ball, and the average diameter of the blooms, as
well as how long the plant has been in its current container. For predictor variables,
she measures several elements in the soil, as well as the amount of light
and water each plant receives.

### Description of the data

Let's pursue Example 1 from above.

We have a hypothetical dataset with 600 observations on seven variables which
can be obtained by clicking on mvreg.dat.
The psychological variables are locus of control (locus),

self-concept (self), and
motivation (motiv). The academic variables are standardized tests scores in
reading (read), writing (write), and science (science), as well as a categorical
variable (prog) giving the type of program the student is in; general (prog=1), academic
(prog=2), or
vocational (prog=3). In addition to the three-category variable prog, the dataset
contains a dummy variable for each level of prog (prog1, prog2,
and prog3), for example, prog1 is equal to 1 when prog=1
(general),
and 0 otherwise. You can store the data file anywhere you like, but our examples will
assume it has been stored in c:data. (Note that the names of
variables should NOT be included at the top of the data file. Instead, the
variables are named as part of the variable command.) You may want to do your
descriptive statistics in a general use statistics

package, such as SAS, Stata
or SPSS, because the options for obtaining descriptive statistics are limited in
Mplus. Even if you chose to run descriptive statistics in another package, it is
a good idea to run a model with type=basic before you do anything else,
just to make sure the dataset is being read correctly. The input file below
shows such a model.

```
Data:
File is c:datamvreg.dat ;
Variable:
Names are locus self motiv read write science prog prog1 prog2 prog3;
Missing are all (-9999) ;
analysis:
type = basic;
```

As we mentioned above, you will want to look at the output from this command carefully to be sure that
the dataset was read into Mplus correctly. You will want to make sure that

**you have the correct number of observations, and that the variables all have**
**means that are close to those from the descriptive statistics generated in a**
**general purpose statistical package. If there are missing values for some or all**
**of the variables, the descriptive statistics generated by Mplus will not match**
**those from a general purpose statistical package exactly, because by default, Mplus versions**
**5.0 and later use maximum likelihood based procedures for handling missing**
**values.**

**<output omitted>**

**SUMMARY OF ANALYSIS**

**Number of groups 1**
**Number of observations 600**

**<output omitted>**

**SAMPLE STATISTICS**

**Means**

LOCUS SELF MOTIV READ WRITE

_____ _____ _____ _____ _____

1 0.097 0.005 0.004 51.902 52.385

Means
SCIENCE PROG PROG1 PROG2 PROG3

_____ _____ _____ _____ _____

1 51.763 2.088 0.230 0.452 0.318

### Analysis methods you might consider

Below is a list of some analysis methods you may have encountered.
Some of the methods listed are quite reasonable while others have either
fallen out of favor or have limitations.

### Multivariate regression analysis

The input file for our multivariate regression in Mplus is shown below. In the variable command, the usevariables option is used because only some of the variables in our dataset are used in the model.
In the model command, each of the outcome variables (i.e., locus, self, and
motiv) is predicted by the four predictor variables using

**the keyword**
**on. In the output command we have requested fully standardized output**
**(in addition to the unstandardized coefficients) using the stdyx option, this will produce standardized estimates of the**
**coefficients, which you may find useful, but it also requests that Mplus produce**
**the R-square statistic for each of the outcome variables.**

**data:**
**file is C:datamvreg.dat ;**
**variable:**
**names are locus self motiv read write science prog prog1 prog2 prog3;**
**missing are all (-9999) ;**
**usevariables are locus self motiv read write science prog2 prog3;**
**model:**
**locus on read write science prog2 prog3;**
**self on read write science prog2 prog3;**
**motiv on read write science prog2 prog3;**
**output:**
**stdyx;**

## SUMMARY OF ANALYSIS

Number of groups 1
Number of observations 600

Number of dependent variables 3
Number of independent variables 5
Number of continuous latent variables 0

Observed dependent variables

Continuous
LOCUS SELF MOTIV

Observed independent variables
READ WRITE SCIENCE PROG2 PROG3

Estimator ML
Information matrix OBSERVED
Maximum number of iterations 1000
Convergence criterion 0.500D-04
Maximum number of steepest descent iterations 20
Maximum number of iterations for H1 2000
Convergence criterion for H1 0.100D-03

<output omitted>

# THE MODEL ESTIMATION TERMINATED NORMALLY

# TESTS OF MODEL FIT

## Chi-Square Test of Model Fit

**Value 0.000**
**Degrees of Freedom 0**
**P-Value 0.0000**

## Chi-Square Test of Model Fit for the Baseline Model

**Value 311.076**
**Degrees of Freedom 18**
**P-Value 0.0000**

## CFI/TLI

**CFI 1.000**
**TLI 1.000**

## Loglikelihood

**H0 Value -8807.819**
**H1 Value -8807.819**

## Information Criteria

**Number of Free Parameters 24**

**Akaike (AIC) 17663.637**

**Bayesian (BIC) 17769.164**

**Sample-Size Adjusted BIC 17692.970**

**(n* = (n + 2) / 24)**

**RMSEA (Root Mean Square Error Of Approximation)**

**Estimate 0.000**

**90 Percent C.I. 0.000 0.000**

**Probability RMSEA**

**MODEL RESULTS**

**Two-Tailed**

**Estimate S.E. Est./S.E. P-Value**

**LOCUS ON**

**READ 0.013 0.004 3.380 0.001**

**WRITE 0.012 0.003 3.599 0.000**

**SCIENCE 0.006 0.004 1.590 0.112**

**PROG2 0.128 0.064 2.008 0.045**

**PROG3 0.252 0.068 3.694 0.000**

**SELF ON**

READ 0.001 0.004 0.311 0.755

WRITE -0.004 0.004 -1.121 0.262

SCIENCE 0.005 0.004 1.290 0.197

PROG2 0.276 0.072 3.827 0.000

PROG3 0.423 0.077 5.474 0.000

MOTIV ON

READ 0.010 0.005 2.085 0.037

WRITE 0.018 0.004 4.143 0.000

SCIENCE -0.009 0.005 -1.981 0.048

PROG2 0.360 0.080 4.514 0.000

PROG3 0.620 0.085 7.252 0.000

SELF WITH

LOCUS 0.057 0.017 3.335 0.001

MOTIV WITH

LOCUS 0.060 0.019 3.179 0.001

SELF 0.130 0.022 5.935 0.000

Intercepts

LOCUS -1.625 0.156 -10.401 0.000

SELF -0.372 0.177 -2.100 0.036

MOTIV -1.311 0.196 -6.689 0.000

**Residual Variances**
**LOCUS 0.365 0.021 17.321 0.000**
**SELF 0.470 0.027 17.320 0.000**
**MOTIV 0.574 0.033 17.321 0.000**

**Because we used the stdyx option of the output command the**
**output includes standardized coefficients. We did this primarily to obtain the**
**R-square values for the outcome variables, so we have omitted the standardized**
**output to save space.**

**<output omitted>**

**R-SQUARE**

**Observed Two-Tailed**
**Variable Estimate S.E. Est./S.E. P-Value**

**LOCUS 0.187 0.029 6.508 0.000**
**SELF 0.054 0.018 3.010 0.003**
**MOTIV 0.150 0.027 5.580 0.000**

**If you ran a separate OLS regression**

for each outcome variable, you would get exactly the same coefficients and
standard errors as shown above. So why conduct a multivariate regression? One advantage of estimating the series of
equations as a single model is that you can conduct tests of the coefficients
across the different outcome variables. For example, the input file below uses
the model test command to test
the null hypothesis that the coefficients for the variable read are equal to
0 in all three equations. Notice that in the model command each of the
terms we wish to test (i.e., each instance of read) is followed by a label
in parentheses (e.g., "(r1)"). These parameter labels are then used to refer to
the associated coefficients in the model test command. There are a few important
things to note about parameter labels. First, the labels must always appear at
the end of a line (but not necessarily the end of the command). Second, the

**labels apply to all**
**parameters listed on the line (meaning all of the parameters on the line are**
**constrained to equality). This is why read is the only predictor variable**
**on the line with the label on it. In the model test command, we give the null**
**hypotheses we wish to test together, in this case, that each of the parameters**
**for read (identified as r1, r2, and r3) are simultaneously equal to zero.**

**data:**
**file is C:datamvreg.dat ;**
**variable:**
**names are locus self motiv read write science prog prog1 prog2 prog3;**
**missing are all (-9999) ;**
**usevariables are locus self motiv read write science prog2 prog3;**
**model:**
**locus on read (r1)**
**write science prog2 prog3;**
**self on read (r2)**

**write science prog2 prog3;**
**motiv on read (r3)**
**write science prog2 prog3;**
**model test:**
**r1 = 0;**
**r2 = 0;**
**r3 = 0;**
**output:**
**stdyx;**

**The output generated by this syntax will be identical to the output shown above, except that**
**it will include the additional output generated by the model test**
**command, the additional**
**output is shown below (all other output is omitted).**

**Wald Test of Parameter Constraints**

**Value 14.486**
**Degrees of Freedom 3**
**P-Value 0.0023**

**The Wald test statistic of 14.486 with 3 degrees of**

freedom has an associated
p-value of 0.0023. These results reject the null hypothesis that the coefficients for
read
across the three equations are simultaneously equal to 0, in other words, the coefficients
for read, taken for all three outcomes together, are statistically significant.

We can also test the null hypothesis that the coefficients for prog=2 (prog2) and prog=3
(prog3) are simultaneously equal to 0 in the equation for locus_of_control.
When used to test the coefficients for dummy variables that form a single
categorical predictor, this type of test is sometimes called an overall
test for the effect of the categorical predictor (i.e., prog).

```
data:
file is C:datamvreg.dat ;
variable:
names are locus self motiv read write science prog
prog1 prog2 prog3;
```

**missing are all (-9999) ;**
**usevariables are locus self motiv read write science prog2 prog3;**
**model:**
**locus on read write science**
**prog2 (p1)**
**prog3 (p2);**
**self on read write science prog2 prog3;**
**motiv on read write science prog2 prog3;**
**model test:**
**p1 = 0;**
**p2 = 0;**
**output:**
**stdyx;**

**Wald Test of Parameter Constraints**

**Value 13.788**
**Degrees of Freedom 2**
**P-Value 0.0010**

**The results of the above test indicate that the two coefficients**
**together are significantly different from 0, in other**

words, the overall
effect of prog on locus_of_control is statistically significant.

The next example tests the null hypothesis that the coefficient for the variable
write in the equation with locus_of_control as the outcome is equal to the coefficient
for write in the equation with self_concept as the outcome. Another way
of stating this null hypothesis is that the effect of write on locus_of_control is equal to the effect of
write on self_concept.

Data:
File is mvreg.dat ;
Variable:
Names are locus self motiv read write science prog prog1 prog2 prog3;
Missing are all (-9999) ;
usevariables are locus self motiv read write science prog2 prog3;
model:
locus on read

**write (wl)**
**science prog2 prog3;**
**self on read**
**write (ws)**
**science prog2 prog3;**
**motiv on read write science prog2 prog3;**
**model test:**
**wl = ws;**

**Wald Test of Parameter Constraints**

**Value 12.006**
**Degrees of Freedom 1**
**P-Value 0.0005**

**The results of this test indicate that the coefficients for write with locus_of_control and self_concept as the outcome are significantly different.**

**Below we test the null hypothesis that the coefficient of science in the equation for locus_of_control is equal to the coefficient for science in the equation for self_concept, and that the coefficient for**

**the variable**
**write in the equation with the outcome variable locus_of_control equals the coefficient for write in the equation with the outcome variable self_concept.**

**data:**
**file is mvreg.dat ;**
**variable:**
**names are locus self motiv read write science prog prog1 prog2 prog3;**
**missing are all (-9999) ;**
**usevariables are locus self motiv read write science prog2 prog3;**
**model:**
**locus on read**
**write (w1)**
**science (s1)**
**prog2 prog3;**
**self on read**
**write (w2)**
**science (s2)**
**prog2 prog3;**
**motiv on read write science prog2 prog3;**
**model test:**

**w1 = w2;**

**s1 = s2;**

**output:**

**stdyx;**

**Wald Test of Parameter Constraints**

**Value 12.902**

**Degrees of Freedom 2**

**P-Value 0.0016**

**The results of the above test indicate that taken together the two sets of coefficients are significantly different. Note that the degrees of freedom is now 2, reflecting the fact that we are comparing two sets of coefficients, rather than 1.**

**Unlike some other packages, Mplus does not automatically provide a test for the overall model. However, we can produce an equivalent test by constraining the regression coefficients to 0 in our model and comparing**

**the fit of that model to**
**the current saturated model. To constrain all of the regression coefficients to**
**0, we first constrain all of the coefficients by giving them the label n**
**(recall from above that the label applies to all coefficients on the line. Below**
**that, we use the model constraint command to fix n to 0.**

**Data:**
**File is mvreg.dat ;**
**Variable:**
**Names are locus self motiv read write science prog prog1 prog2 prog3;**
**Missing are all (-9999) ;**
**usevariables are locus self motiv read write science prog1 prog2 ;**
**model:**
**locus on read write science prog1 prog2 (n) ;**
**self on read write science prog1 prog2 (n);**
**motiv on read write science prog1 prog2 (n);**
**model constraint:**
**n = 0;**

**Because this isn't the model we want to interpret, we have omitted most of the output.**
**Shown below are the chi-square test of model fit (which provides the overall test) and the**
**MODEL RESULTS section so that we can check to see we have estimated the desired model.**

## TESTS OF MODEL FIT

### Chi-Square Test of Model Fit

**Value 214.658**
**Degrees of Freedom 15**
**P-Value 0.0000**

**The chi-square test of model fit compares the fit of the current model to a saturated model. In the**
**models we estimated above (i.e., the unconstrained or saturated models), this value was 0, because the model was saturated (i.e., has 0 degrees of freedom). By adding**
**constraints to the model we have freed up 15 parameters, so now we get a**
**positive**
**value. The chi-square value of 214.658 with 15 degrees**

of freedom with an associated
p-value of less than 0.0001, indicates that the constrained model fits
significantly worse than the saturated model. In other words, the saturated
model shown above fits significantly better than the model with the regression
coefficients constrained to 0.

The MODEL RESULTS are shown below. It can be a good idea to check this section
to make sure the model estimated was the desired model. Note that all of the regression
coefficients (denoted ON) are constrained to 0, while the residual covariances
(denoted WITH) and variances, as well as the intercepts have been estimated.

MODEL RESULTS

Two-Tailed
Estimate S.E. Est./S.E. P-Value

LOCUS ON
READ 0.000 0.000 999.000 999.000

WRITE 0.000 0.000 999.000 999.000
SCIENCE 0.000 0.000 999.000 999.000
PROG1 0.000 0.000 999.000 999.000
PROG2 0.000 0.000 999.000 999.000

SELF ON
READ 0.000 0.000 999.000 999.000
WRITE 0.000 0.000 999.000 999.000
SCIENCE 0.000 0.000 999.000 999.000
PROG1 0.000 0.000 999.000 999.000
PROG2 0.000 0.000 999.000 999.000

MOTIV ON
READ 0.000 0.000 999.000 999.000
WRITE 0.000 0.000 999.000 999.000
SCIENCE 0.000 0.000 999.000 999.000
PROG1 0.000 0.000 999.000 999.000
PROG2 0.000 0.000 999.000 999.000

SELF WITH
LOCUS 0.081 0.020 4.133 0.000

MOTIV WITH
LOCUS 0.135 0.023 5.832 0.000
SELF 0.167 0.025 6.791 0.000

## Intercepts

**LOCUS 0.097 0.027 3.531 0.000**

**SELF 0.005 0.029 0.171 0.864**

**MOTIV 0.004 0.034 0.116 0.907**

## Residual Variances

**LOCUS 0.449 0.026 17.321 0.000**

**SELF 0.497 0.029 17.321 0.000**

**MOTIV 0.675 0.039 17.321 0.000**

### Things to consider

### References

**Afifi, A., Clark, V. and May, S. 2004. Computer-Aided Multivariate Analysis. 4th ed. Boca Raton, Fl: Chapman & Hall/CRC.**