What is Multivariate Regression Analysis and how can it be performed using SAS Data Analysis?

Authored by

stats writer

June 29, 2024

Multivariate Regression Analysis is a statistical method used to analyze the relationship between multiple independent variables and a single dependent variable. It allows for the identification and quantification of the impact of each independent variable on the dependent variable, while controlling for the effects of other variables. This type of analysis is commonly used in fields such as economics, finance, and social sciences to understand the factors that influence a particular outcome.

Using SAS Data Analysis, Multivariate Regression Analysis can be performed by first importing the data into the SAS software. Then, the analyst can specify the dependent and independent variables and run the regression model. The results will include the coefficients of each independent variable, their significance levels, and the overall model fit. SAS also offers various tools for data visualization and model diagnostics, which can aid in the interpretation and evaluation of the results. This allows for a comprehensive and efficient analysis of complex relationships between variables. Overall, SAS Data Analysis provides a reliable and user-friendly platform for conducting Multivariate Regression Analysis.

Multivariate Regression Analysis | SAS Data Analysis Examples

As the name implies, multivariate regression is a technique that estimates a single regression model with multiple outcome variables and one or more predictor variables.

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 relate to the academic variables and gender. In
particular, the researcher is interested in how many dimensions are necessary to understand
the association between the two sets of variables.

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 the current container. For predictor variables,
she measures several elements in the soil, in addition to 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, https://stats.idre.ucla.edu/wp-content/uploads/2016/02/mvreg.sas7bdat, with 600 observations on seven variables.
The psychological variables are locus of control, self-

concept and
motivation. The academic variables are standardized tests scores in
reading, writing, and science, as well as a categorical variable giving the type of program the student is in (general, academic, or
vocational). In our example the dataset https://stats.idre.ucla.edu/wp-content/uploads/2016/02/mvreg.sas7bdat is saved in a
library called data.

Let's look at the data (note that there are no missing values in this data set).

```
proc means data = data.mvreg;
vars locus_of_control self_concept motivation read write science;
run;
```

The MEANS Procedure

Variable Label N Mean Std Dev Minimum Maximum

LOCUS_OF_CONTROL 600 0.0965333 0.6702799

-1.9959567 2.2055113
SELF_CONCEPT 600 0.0049167 0.7055125 -2.5327499 2.0935633
MOTIVATION 600 0.0038979 0.8224000 -2.7466691 2.5837522
READ 600 51.9018333 10.1029831 24.6200066 80.5864944
WRITE 600 52.3848332 9.7264550 20.0688801 83.9348221
SCIENCE 600 51.7633331 9.7061791 21.9895325 80.3694153

---

proc freq data = data.mvreg;
table prog;
run;

The FREQ Procedure

program type

Cumulative Cumulative
PROG Frequency Percent Frequency Percent

---

1 138 23.00 138 23.00

2 271 45.17 409 68.17

3 191 31.83 600 100.00

```
proc corr data = data.mvreg nosimple;
var locus_of_control self_concept motivation;
run;
```

The CORR Procedure

3 Variables: LOCUS_OF_CONTROL SELF_CONCEPT MOTIVATION

Pearson Correlation Coefficients, N = 600

Prob > |r| under H0: Rho=0

	LOCUS_OF_ CONTROL	SELF_ CONCEPT	MOTIVATION
LOCUS_OF_CONTROL	1.00000	0.17119	0.24513

```
proc corr data = data.mvreg nosimple;
var read write science;
run;
```

The CORR Procedure

3 Variables: READ WRITE SCIENCE

Pearson Correlation Coefficients, N = 600
Prob > |r| under H0: Rho=0

READ WRITE SCIENCE

READ 1.00000 0.62859 0.69069

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

Technically speaking, we will be conducting a multivariate multiple
regression. This regression is "multivariate" because there is more than
one outcome variable. It is a "multiple" regression because there is more
than one predictor variable. Of course, you can conduct

a multivariate
regression with only one predictor variable, although that is rare in practice.

To conduct a multivariate regression in SAS, you can use proc glm,
which is the same procedure that is often used to perform ANOVA or OLS
regression. The syntax for estimating a multivariate regression is similar
to running a model with a single outcome, the primary difference is the use of
the manova statement so that the output includes the multivariate statistics. The f- and p-values for four multivariate criterion are given, including Wilks' lambda, Lawley-Hotelling
trace, Pillai's trace, and Roy's largest root. By specifying h=_ALL_ on
the manova statement, we indicate that we would like multivariate
statistics for all of the predictor variables in the model, if we were only
interested in the multivariate statistics for some variables, we could replace

ALL with the name of a variable (e.g. h=read).

```
proc glm data = data.mvreg;
class prog;
model locus_of_control self_concept motivation
= read write science prog / solution ss3;
manova h=_ALL_;
run;
quit;
```

The SAS output for multivariate regression can be very long, especially if the model has many outcome variables. The output from our example has four parts: one for each of the three outcome variables, and the fourth from the manova statement. Below we will discuss the output in sections.

The GLM Procedure

Class Level Information

Class Levels Values

PROG 3 1 2 3

Number of Observations Read 600
Number of Observations Used 600

Above we see that the class variable prog has three levels. Just below the class level information, we see the number of observations read form the data and the number of observations used in the analysis. If the variables used in the analysis contained missing values the number of observations used would be smaller than the number of observations read.

Dependent Variable: LOCUS_OF_CONTROL

Sum of
Source DF Squares Mean Square F Value Pr > F

Model 5 50.2595509 10.0519102 27.28 F

READ 1 4.16815963 4.16815963 11.31 0.0008
WRITE 1 4.72524304 4.72524304 12.82 0.0004
SCIENCE 1 0.92248638 0.92248638 2.50 0.1141
PROG 2 5.02961991 2.51480995 6.83 0.0012

Standard
Parameter Estimate Error t Value Pr > |t|

Intercept -1.373094234 B 0.16259260 -8.44

The output for the first outcome variable (locus_of_control) is
followed by similar output for each additional outcome (self_concept and
motivation). This output is shown below, but we will not discuss it
further, instead we will move on to the multivariate output.

The GLM Procedure

Dependent Variable: SELF_CONCEPT

Sum of
Source DF Squares Mean Square F Value Pr > F

Model 5 16.1107053 3.2221411 6.79 F

READ 1 0.04557875 0.04557875 0.10 0.7568
WRITE 1 0.59051932 0.59051932 1.24 0.2652
SCIENCE 1 0.78237876 0.78237876 1.65 0.1998

PROG 2 14.21838537 7.10919268 14.97 |t|

Intercept 0.0510179965 B 0.18457670 0.28 0.7823
READ 0.0013076138 0.00422047 0.31 0.7568
WRITE -.0042934282 0.00384990 -1.12 0.2652
SCIENCE 0.0053059405 0.00413348 1.28 0.1998
PROG 1 -.4233591913 B 0.07772768 -5.45 F

Model 5 60.7672827 12.1534565 20.96 F

READ 1 2.49445035 2.49445035 4.30 0.0385
WRITE 1 9.85052717 9.85052717 16.99 |t|

Intercept -.6911458885 B 0.20395228 -3.39 0.0007
READ 0.0096735465 0.00466350 2.07 0.0385
WRITE 0.0175354486 0.00425404 4.12

The final section of output for our model is output for the multivariate
tests of the model.

The GLM Procedure
Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where

H = Type III SSCP Matrix for READ

E = Error SSCP Matrix

Characteristic Characteristic Vector V'EV=1

Root Percent LOCUS_OF_CONTROL SELF_CONCEPT MOTIVATION

0.02414400 100.00 0.05725523 -0.00912678 0.02560444

0.00000000 0.00 -0.00704393 0.05979895 0.00102214

0.00000000 0.00 -0.03710958 -0.01295454 0.04972124

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall READ Effect

H = Type III SSCP Matrix for READ

E = Error SSCP Matrix

S=1 M=0.5 N=295

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.97642519 4.76 3 592 0.0027

Pillai's Trace 0.02357481 4.76 3 592 0.0027

Hotelling-Lawley Trace 0.02414400 4.76 3 592 0.0027

Roy's Greatest Root 0.02414400 4.76 3 592 0.0027

SAS prints similar output for each of the

predictor variables in the model (in this case write, science,
and prog), this output is shown below, but we will not discuss it
further. Instead we will move on to additional tests.

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for WRITE
E = Error SSCP Matrix

Characteristic Characteristic Vector V'EV=1
Root Percent LOCUS_OF_CONTROL SELF_CONCEPT MOTIVATION

0.05552705 100.00 0.03976623 -0.02762931 0.04077279
0.00000000 0.00 0.00235865 0.05460081 0.01173502
0.00000000 0.00 0.05583890 0.00907776 -0.03645138

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WRITE Effect
H = Type III SSCP Matrix for WRITE
E = Error SSCP Matrix

S=1 M=0.5 N=295

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.94739400 10.96 3 592 F

Wilks' Lambda 0.98340548 3.33 3 592 0.0193
Pillai's Trace 0.01659452 3.33 3 592 0.0193
Hotelling-Lawley Trace 0.01687455 3.33 3 592 0.0193
Roy's Greatest Root 0.01687455 3.33 3 592 0.0193

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for PROG
E = Error SSCP Matrix

Characteristic Characteristic Vector V'EV=1
Root Percent LOCUS_OF_CONTROL SELF_CONCEPT MOTIVATION

0.12087752 99.34 0.01903925 0.02549291 0.03813193
0.00080748 0.66 0.04668032 -0.04866125 0.01435613
0.00000000 0.00 0.04651187 0.02844692 -0.03832351

Multivariate Analysis of Variance

MANOVA Test Criteria and F Approximations for the Hypothesis of No Overall PROG Effect

H = Type III SSCP Matrix for PROG

E = Error SSCP Matrix

S=2 M=0 N=295

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.89143832 11.67 6 1184

As mentioned above, if you ran a separate regression for each outcome variable, you would get exactly the same coefficients, standard errors, t- and p-values, and confidence intervals as shown above. So why conduct a multivariate regression? One of the advantages is that you can conduct tests of the coefficients across the different models. Below we show a few of the hypothesis tests you can perform.

For the first test, the null hypothesis is that the coefficient for prog=1 is equal to the coefficient for prog=2 for each dependent variable separately. An alternative way to state this hypothesis

is that the difference
between the two coefficients (i.e., prog=1 - prog=2) is equal to 0.
The estimate statement can be used to perform this test. The text between the apostrophes (i.e.,
') is a label for the output. Next we list the variable name (prog)
followed by a series of numbers, one for each level of prog in order,
these are the values by which the coefficients will be multiplied to perform the
test. To estimate the difference between the coefficient for prog=1
and prog=2 we multiply the coefficient for prog=1 by 1, and the
coefficient for prog=2 by -1, prog=3 is not involved in this test,
so we multiply it by 0.

```
proc glm data = data.mvreg;
class prog;
model locus_of_control self_concept motivation
= read write science prog / solution ss3;
manova h= _ALL_ ;
```

```
estimate 'prog 1 vs. prog 2' prog 1 -1 0;
run;
quit;
```

The output produced by this model is similar to the output for the previous
model, except that it contains additional output associated with the use of the
estimate statement. To save space, we will only show the additional
output.

Dependent Variable: LOCUS_OF_CONTROL

Parameter	Estimate	Standard Error	t Value	Pr > |t|
prog 1 vs. prog 2	-0.12779508	0.06395501	-2.00	0.0462

Dependent Variable: SELF_CONCEPT

Parameter	Estimate	Standard Error	t Value	Pr > |t|
prog 1 vs. prog 2	-0.27648339	0.07260235	-3.81	0.0002

Dependent Variable: MOTIVATION

Standard
Parameter Estimate Error t Value Pr > |t|

prog 1 vs. prog 2 -0.36032939 0.08022363 -4.49

There is separate output for each of the outcome variables. Each of the tables in the output gives the estimate (in this case the difference between the coefficients), the standard error of this estimate, the t-value and associated p-value. The output indicates that the coefficient for prog=1 is significantly different from the coefficient for prog=2 for each of the outcomes.

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. We request this test by adding a second manova statement, where h gives the predictor variable

or variables
to be tested (i.e., h=write) and m gives the combination of outcome variables to test
(i.e., m=locus_of_control - self_concept).

```
proc glm data = data.mvreg;
class prog ;
model locus_of_control self_concept motivation
= read write science prog / solution ss3;
manova h= _ALL_ ;
manova h=write m=locus_of_control - self_concept;
run;
quit;
```

Again, we will only show the portion of the output associated with the new manova statement.
The first table (shown below) gives the matrix for the outcome variables.
In
this case, we want to subtract the coefficients for self_concept (multiplied by
-1) from the values of the coefficients for locus_of_control (multiplied by 1).
Because motivation isn't involved in the test, it is

multiplied by zero.

M Matrix Describing Transformed Variables

LOCUS_OF_
CONTROL SELF_CONCEPT MOTIVATION

MVAR1 1 -1 0

The GLM Procedure
Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Type III SSCP Matrix for WRITE
E = Error SSCP Matrix

Variables have been transformed by the M Matrix

Characteristic Characteristic Vector V'EV=1
Root Percent MVAR1

0.02001074 100.00 0.04807919

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall WRITE Effect
on the Variables Defined by the M Matrix

Transformation

H = Type III SSCP Matrix for WRITE

E = Error SSCP Matrix

S=1 M=-0.5 N=296

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.98038183 11.89 1 594 0.0006
Pillai's Trace 0.01961817 11.89 1 594 0.0006
Hotelling-Lawley Trace 0.02001074 11.89 1 594 0.0006
Roy's Greatest Root 0.02001074 11.89 1 594 0.0006

The last table in the output shows that regardless of which multivariate statistic is used,
the coefficient for write with locus_of_control as the outcome
and the coefficient for write with self_concept as the outcome
are significantly different.

For the final example, we test the null hypothesis that the coefficient for
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 for locus_of_control is equal to the coefficient for write in the
equation for self_concept. To perform this test we need to use both the
contrast statement and the manova statement. In the contrast
statement, we specify the predictor variables we wish to test, in this case, we
want to multiply the coefficients for write and science by 1. In
the manova statement, we specify the portions of the test specific to the
outcome variables, in this case, we want to compare the coefficients for
locus_of_control and self_concept, by subtracting one set of
coefficients from the other.

```
proc glm data = data.mvreg;
class prog;
model locus_of_control self_concept motivation
```

= read write science prog / solution ss3;
contrast 'write & science' write 1,
science 1 /e;
manova m=locus_of_control - self_concept;
run;
quit;

As before, we will only show the portions of output associated with the test we are performing.
Towards the beginning of the output (just after the class level information section) we see the table of contrasts for the coefficients.
The matrix has two columns, one for each of the effects we wish to test.

Coefficients for Contrast write & science

Row 1 Row 2

Intercept 0 0

READ 0 0

WRITE 1 0

SCIENCE 0 1

PROG 1 0 0

PROG 2 0 0

PROG 3 0 0

The output shown below is generated by the manova statement, and as before
it appears towards the end of the output.

M Matrix Describing Transformed Variables

LOCUS_OF_
CONTROL SELF_CONCEPT MOTIVATION

MVAR1 1 -1 0

Multivariate Analysis of Variance

Characteristic Roots and Vectors of: E Inverse * H, where
H = Contrast SSCP Matrix for write & science
E = Error SSCP Matrix

Variables have been transformed by the M Matrix

Characteristic Characteristic Vector V'EV=1

Root Percent MVAR1

0.02150343 100.00 0.04807919

MANOVA Test Criteria and Exact F Statistics for the Hypothesis of No Overall write & science Effect on the Variables Defined by the M Matrix Transformation

H = Contrast SSCP Matrix for write & science

E = Error SSCP Matrix

S=1 M=0 N=296

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.97894924 6.39 2 594 0.0018

Pillai's Trace 0.02105076 6.39 2 594 0.0018

Hotelling-Lawley Trace 0.02150343 6.39 2 594 0.0018

Roy's Greatest Root 0.02150343 6.39 2 594 0.0018

The last table in the above output shows that regardless of which multivariate statistic is used, taken together, the two sets of coefficients are significantly different.

Things to consider

References

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

See also

SAS Library: Multivariate regression in SAS