Linear Regression in Mplus

Patricio Troncoso & Margarita Panayiotou

Some background information

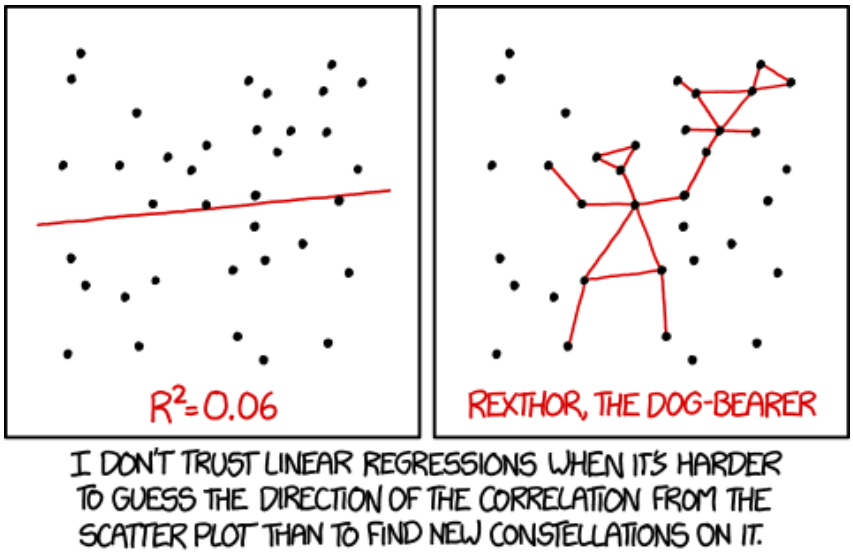
For this practical, we'll be using open data from the following source: Wright, D. and London, K. (2011). Modern Regression Techniques Using R, London: Sage. (Chapter 8). This is available *here*.

This dataset contains information about a randomised controlled trial, in which primary school pupils were asked about the time they spent doing exercises (weekly). Then, their classes (within schools) were divided into 4 different groups and pupils within those classes were given different treatments:

1. Control (no treatment at all)
2. Leaflet (information about exercising)
3. Leaflet+plan (the leaflet and a plan of exercise)
4. Leaflet+plan+quiz (leaflet, plan and a quiz to do about exercising)

Pupils were asked again about the time they spent doing exercises (weekly) some time after the treatments were given.

The variable **sqw2** is the outcome (time spent doing exercises after treatment). The variable **sqw1** is the measure before the treatment. The variable **wcond** is the group or condition to which the class of the pupils was assigned.



Create the input

a) Read in data

Make sure to save the dataset in the same folder as your input file. Also don't forget to mention the variable names in the correct order:

w1sn02 w2sn02 w1int w2int w1att w2att z1pbc z2pbc sqw1 sqw2;

prac4_linearRegression

```
TITLE: simple linear regression

DATA:

file = 'RegressionMplus.dat';
!this is a free format file

VARIABLE:

names = wcond class w1sn02 w2sn02 w1int w2int w1att
        w2att z1pbc z2pbc sqw1 sqw2;

usevariables = ; |

missing = ; !no missing data
```

b) Prepare the analysis and output sections

The MODEL command will be used later on to run our model. For now we can leave it empty.

```
ANALYSIS:

estimator = ml; !maximum likelihood

MODEL:


OUTPUT: stand res sampstat;
```

Task 1: Correlation

Before we start any data analysis process:
What is the correlation between exercise before (**sqw1**) and after (**sqw1**) the intervention?

There are two ways in which one can run correlations in Mplus, but first make sure you add the names of the two variables in the USEVARIABLES:

USEVARIABLES = sqw1 sqw2;

Option 1: The "sampstat" in the OUTPUT: always gives you the correlations between variables. With sampstat and the MODEL: command empty you get this in the Output file:

prac4_linearregression

```
     SAMPLE STATISTICS


          Means
             SQW1          SQW2
             ________      ________
               1.681         1.789


          Covariances
             SQW1          SQW2
             ________      ________
SQW1           0.363
SQW2           0.259         0.317


          Correlations
             SQW1          SQW2
             ________      ________
SQW1           1.000
SQW2           0.765         1.000
```

Option 2: Define the correlation in the MODEL: command.

MODEL:
sqw1 with sqw2; !WITH defines a correlation

You will find the results in the STDYX Standardization section of the output:

```
STANDARDIZED MODEL RESULTS

STDYX Standardization

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

 SQW1     WITH
    SQW2               0.765      0.019     41.337      0.000
```

Correlations go from -1 to 1, so we can safely say that these two variables are highly correlated. This is not surprising at all, since what we're measuring is behaviour (time spent doing exercises) and you know what they say "the best predictor of future behaviour is past behaviour".

In any case, this is overly simplistic, so we'll proceed to run a series of linear regression models. The correlation between the two variables is high, positive, and statistically significant. Now let's run a simple linear regression.

Task 2: Simple linear regression

Simple linear regression has the following algebraic form:

$$Y_i = \beta_0 + \beta_1 \chi_{i1} + \varepsilon_i Y_i = \beta_0 + \beta_1 \chi_{i1} + \varepsilon_i$$

Where:

Y_i is the dependent variable (outcome of interest)
β_0 is the intercept (overall mean)
β_1 is the change in the dependent variable when the independent variable changes in one unit
χ_{i1} is the independent variable (aka explanatory variable, predictor or covariate)
ε_i is the residual (aka error)

In Mplus y (dependent variable) always appears to the left of the command, whereas x (independent variable) appears to the right of the command:

MODEL:
y ON x; !ON defines a regression model. X predicts Y

Question time!

Q1. What is the effect of sqw1 (independent) on sqw2 (dependent)?

Q2. How much variance is explained?

You can find the answers in the below sections of your Mplus output:

MODEL RESULTS (unstandardized findings)

STDYX Standardization (standardized findings)

prac4_linearregression

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SQW2 ON				
SQW1	0.714	0.027	26.628	0.000
Intercepts				
SQW2	0.589	0.048	12.288	0.000
Residual Variances				
SQW2	0.131	0.008	15.859	0.000

STANDARDIZED MODEL RESULTS

STDYX Standardization

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SQW2 ON				
SQW1	0.765	0.019	41.333	0.000
Intercepts				
SQW2	1.046	0.110	9.542	0.000
Residual Variances				
SQW2	0.415	0.028	14.661	0.000

R-Square (variance explained)

R-SQUARE

Observed Variable	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
SQW2	0.585	0.028	20.667	0.000

Task 3: Multiple linear regression

Normally in social sciences, you wouldn't expect phenomena to be explained by only one variable. In the example we're working on

$$Y_i = \beta_0 + \beta_1\chi_{i1} + \beta_2\chi_{i2} + \cdots + \beta_n\chi_{in} + \epsilon_i Y_i = \beta_0 + \beta_1\chi_{i1} + \beta_2\chi_{i2} + \cdots + \beta_n\chi_{in} + \varepsilon_i$$

All the terms mean the same as before, so we'll focus on the new information:

β_2 is the expected increase in the dependent variable conditional on the rest of the variables remaining constant

χ_{i2} is a second independent variable

$\beta_n\chi_{in}$ is used to indicate that there can be multiple independent variables

Task: Now let's run a multiple linear regression model by adding the experimental condition variable.

Tip: The experimental condition is a categorical (nominal) variable with 4 categories. In Mplus nominal variables need to be transformed into dummy variables first.

Task 3a: Create 3 dummy variables called wcond2 wcond3 wcond4

Tip: Add the below before ANALYSIS:

DEFINE:

```
wcond2 = 0;
if (wcond ==2) then wcond2 = 1;
wcond3 = 0;
if (wcond == 3) then wcond3 = 1;
wcond4 = 0;
if (wcond == 4) then wcond4 = 1;
```

Tip: Make sure to add the three new variables into the USEVARIABLES list.

Task 3a: Run the multiple regression with sqw1 and the 3 conditions as predictors

Question time!

Q3. What is the effect of condition 2 (wcond2) in relation to the control group?

Q4. What is the effect of condition 3 (wcond3) in relation to the control group?

Q5. What is the effect of condition 4 (wcond4) in relation to the control group?

Tip: Remember, dependent variables are mentioned before ON and independent variables are mentioned after ON, like below.

MODEL:

!Multiple regression

!y is predicted by x1 x2 x3 x4

y ON x1 x2 x3 x4;

```
MODEL RESULTS

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

SQW2     ON
   SQW1              0.716      0.026     27.362      0.000
   WCOND2            0.158      0.044      3.573      0.000
   WCOND3            0.207      0.044      4.733      0.000
   WCOND4            0.218      0.045      4.832      0.000

Intercepts
   SQW2              0.441      0.054      8.182      0.000

Residual Variances
   SQW2              0.124      0.008     15.859      0.000
```

Tip: For the standardized coefficients see:

```
STANDARDIZED MODEL RESULTS


STDYX Standardization !for continuous predictors (like sqw1)

STDY Standardization !for binary predictors (like wcond2 wcond3
wcond4)
```

Task 4: Adding interaction effects

Adding interaction is fairly straightforward in Mplus. All you need to do is "multiply" the terms you want to interact (using *) in the DEFINE: command.

Take the previous model and add an interaction term between the experimental condition and exercise before the intervention.

```
VARIABLE:

names = wcond class w1sn02 w2sn02 w1int w2int w1att
        w2att z1pbc z2pbc sqw1 sqw2;

usevariables = sqw1 sqw2 wcond2 wcond3 wcond4
sqw1C2 sqw1C3 sqw1C4;

missing = ; !no missing data


DEFINE:
wcond2 = 0;
if (wcond == 2) then wcond2 = 1;
wcond3 = 0;
if (wcond == 3) then wcond3 = 1;
wcond4 = 0;
if (wcond == 4) then wcond4 = 1;


sqw1C2 = sqw1*wcond2;
sqw1C3 = sqw1*wcond3;
sqw1C4 = sqw1*wcond4;
```

Tip: Remember to always add any new variables created in DEFINE:, in the USEVARIABLES list. New variables must appear last in the USEVARIABLES.

Now let's add the three interaction terms to the previous multiple regression.

```
ANALYSIS:

estimator = ml; !maximum likelihood

MODEL:

sqw2 on sqw1 wcond2 wcond3 wcond4
sqw1C2 sqw1C3 sqw1C4;


OUTPUT: stand res sampstat;
```

Question time!

Q6. What is the R-squared of this model?

Q7. Does the addition of interactions improve the model?

Task 5: Predictions

You can calculate predicted values by using the values of the coefficients. So, for example, you can predict the score of a pupil who scored 2 in the baseline measure and was in the control group as such:

Since the control group is coded 0, the equation for these pupils reduces to:

$Y_i = \beta_0 + \beta_1 \chi_{i1} Y_i = \beta_0 + \beta_1 \chi_{i1}$

Then

$Y = 0.38351 + 0.75041 * \text{baseline}$
$Y = 0.38351 + 0.75041 * 2$
$Y = 1.88433$

The predicted score after the trial for a child in the control group who scored 2 at baseline is then 1.88433.

Question time!

Q8. What is the predicted score for a child in condition 2 that scored 1.5 at baseline?

Plotting the estimated model

For plots in Mplus you need to add a new command called PLOT

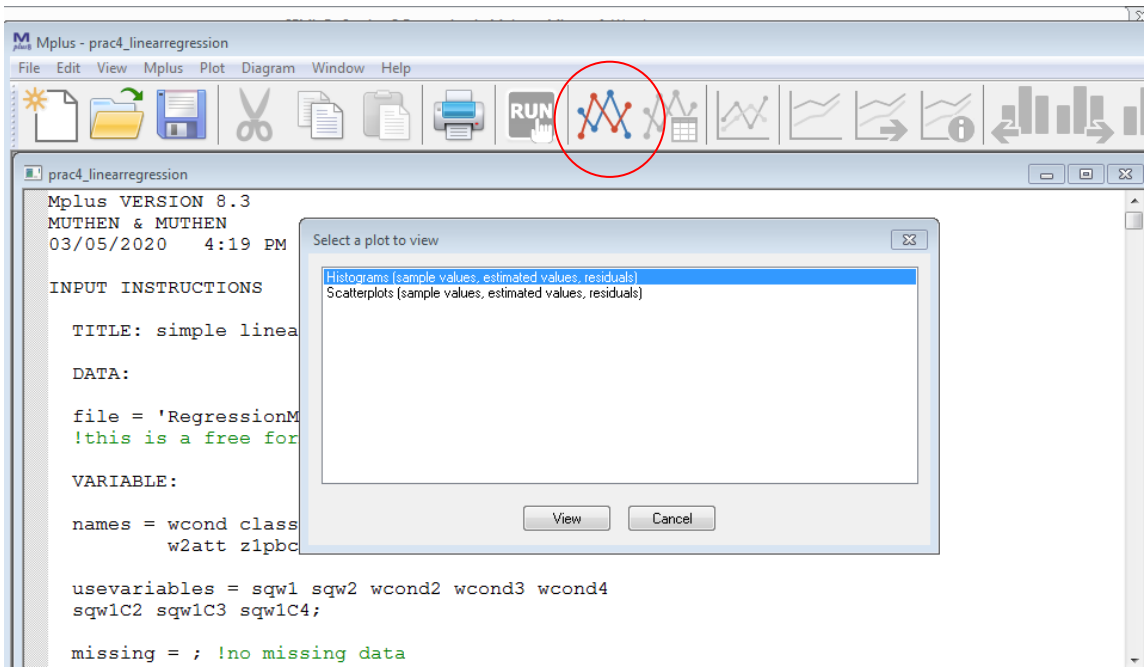
```
MODEL:

sqw2 on sqw1 wcond2 wcond3 wcond4
sqw1C2 sqw1C3 sqw1C4;


OUTPUT: stand res sampstat;

PLOT: type = plot1 plot3;
!for a description of what is available
!for each PLOT see pages 846-848 of the
!Mplus user manual
```

Rerun the model with PLOT and then select the plots in Mplus through the Plot button or through “Plot” in the taskbar:

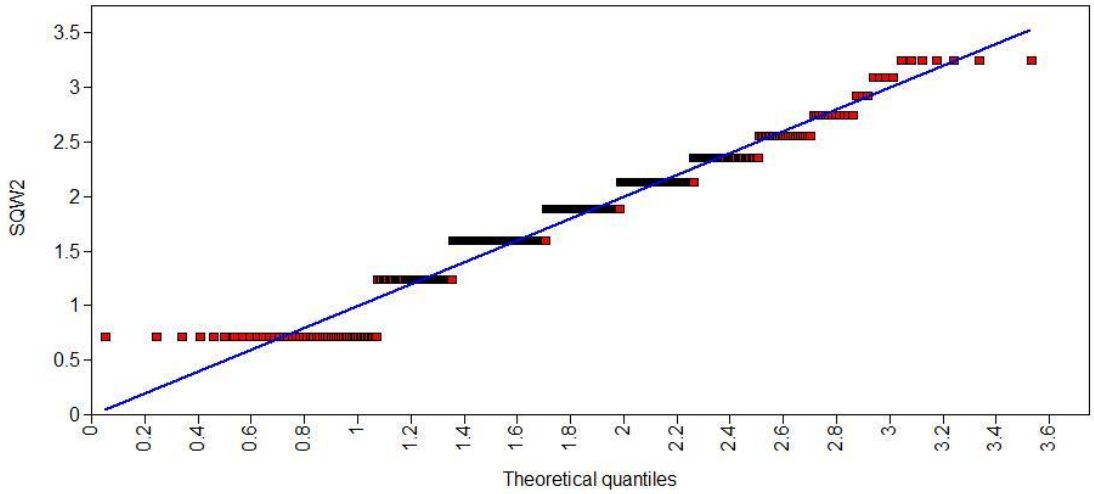


→ Select Histrograms

→ Display Properties

→ Select QQPlots

→ Press OK.



Task 6: Checking assumptions

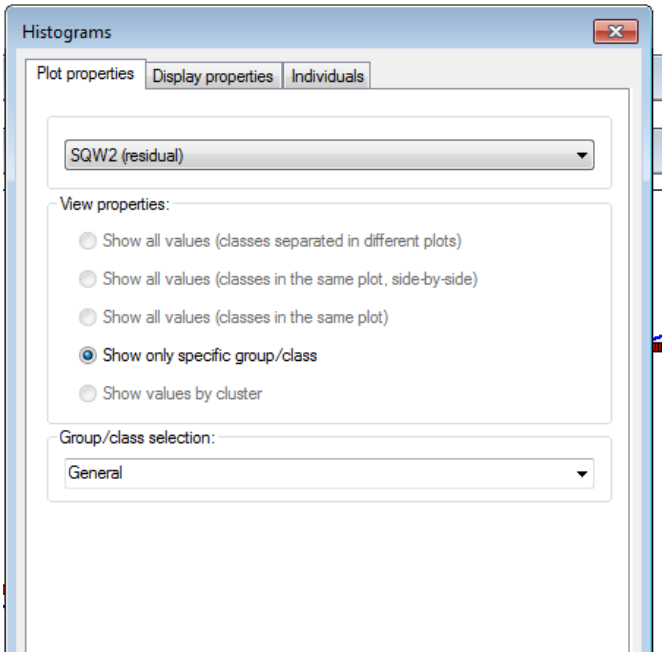
We are going to focus on the two most basic assumption checks that you can do, which are to with the residuals. We will use graphic methods, since they are quite straightforward and easily understood.

6.1. Normality of residuals

A key assumption in linear regression is that our residuals are normally distributed. This means roughly that whatever is left unexplained by our model could be thought of as random variation or “white noise”.

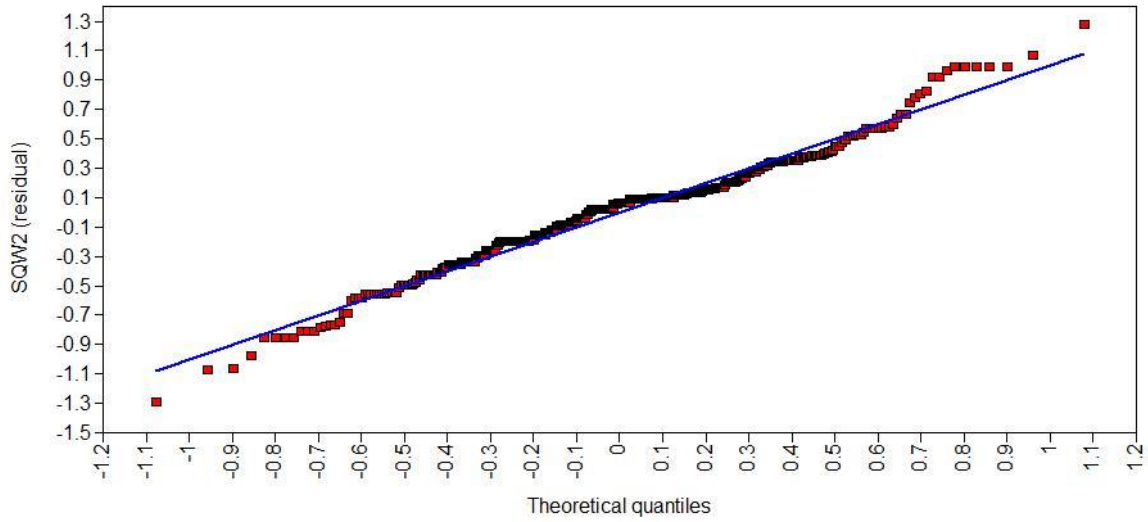
The normality of the residuals can be checked by running a formal statistical test, e.g. the Kolmogorov-Smirnov test, but those methods can be overly sensitive to small departures. Graphic methods, such as drawing a histogram or a Q-Q plot, are widely used instead.

To draw a Q-Q plot from your model results follow the previous steps, but this time we need to plot a QQ plot of the residuals: Go to Plot Properties and select **SQW2 (residual)** from the drop down list.



The result is a Q-Q plot or normal probability plot. It compares the residuals against the normal distribution. Put simply, the residuals (dots on the plot) should all lie on the diagonal line. Any obvious patterns would indicate that our model does not meet the normality assumption.

In the plot above, we can see that the extremes of the distribution depart from normality, which could be indicating that there are outliers. It may also be the case that we need to control for other variables.



6.2. Homoscedasticity

Homoscedasticity means that residuals should distribute evenly in terms of spread across the predicted values of our model. In other words, if we plot residuals vs fitted values, it should look completely random, i.e. a random cloud of dots.

To do this:

→ Press the plots button

→ Select Scatterplots

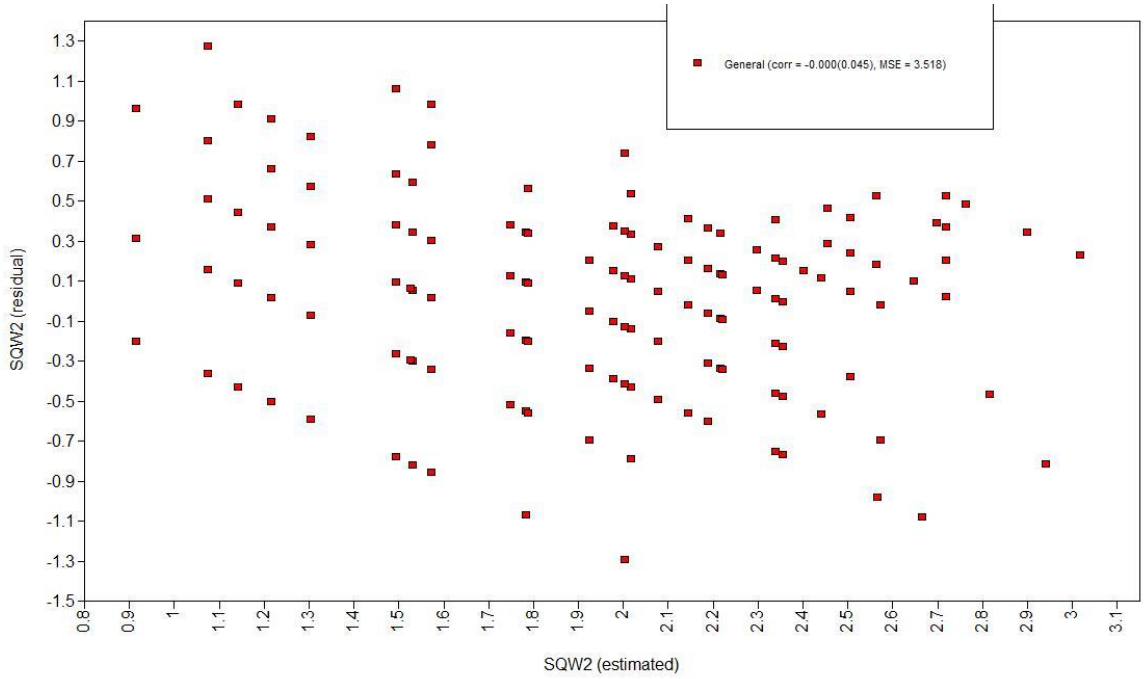
→ View

→ Plot Properties

→ X: SQW2 (estimated)

→ Y: SQW2 (residual)

→ Press OK



What we are interested to see here is no obvious patterns. In this specific plot, there seems to be some non-random pattern (not a huge one). The left-hand side of the distribution (lower predicted values) tends to have more positive residuals, whereas the right-hand side has more negative residuals. How do you address this? Usually, by adding more variables to the model.