From The Expert: Multiple Linear Regression

From the Expert: Multiple Linear Regression

Multiple linear regression (MLR) is used to predict the response variable when there are 2 or more quantitative explanatory variables. For example, the selling price of a house may depend on location, number of bedrooms, number of bathrooms, year built, etc. The model can be represented as:

Where is the dependent variable

are the independent variables
is the intercept
are coefficients

For example:
(slope along axis) represents the expected change in the response per one unit change in when other independent variables hold fixed
(slope along axis) represents the expected change in the response per one unit change in when other independent variables hold fixed

is the error
and k is the number of independent variables.

Multiple regression is a common technique utilized in many fields such as medicine, social sciences, education, and business. Multiple regression reflects real world situations better than simple linear regression. In addition, the predictor variables can be any combination of continuous, categories or dichotomous. In cases of category variables, additional coding techniques need to be performed.

The assumptions for simple linear regression (e.g. linearity, normality, independence, and constant variance) are also applied for MLR.

Here are suggestions to deal with multiple x variables:

- Investigate each individual variable (e.g. mean, standard deviation, min,max, outliers, histogram, stem plot) to understand each variable.
- Examine the relationship between different pairs of variables using correlation and scatter plots. The variables with strong relationships with y will be kept.
- Perform a regression using all explanatory variables

The regression line is:

is the *y* intercept

Slopes () are the weight of each independent variable, and adjusted for the other independent variables in the model.

The residual for *i*-th observation is:

The standard deviation (estimate of) of *y* is

1. Analysis of Variance (ANOVA) table for MLR

Source	Sum of Squares	df	Mean Square	F
Regression	SSR= = SST-SSE	k		df=k,N-k-1
Error	SSE= = SST-SSR	N-k-1		
Total	SST= =SSR+SSE	N-1		

Variation allocation for MLR
Total sum of squares (*SST*) is the sum of squared deviation from the grand mean. SST is a measure of the total variability in *y*

Sum of squares regression (*SSR*) is total sum squared deviation of the regression (predicted value) from the grand mean. This is the variation that is accounted for by the regression.

Sum of squares error/residual (*SSE*) is the total sum squared deviation of the predicted values from the observed value. This is the unexplained variation.

Sum of squares partitioning: SST = SSR + SSE

- Coefficient of determination `R^2` measures the quality of a regression model

is the variation in *y* that can be explained by the linear regression of *y* on or the amount of linear association between response *y* and multiple explanatory variables. Rules of thumb for `R^2` depend on the field. For example, used in the Social Sciences:

- 0 no linear relationship
- .10 small (r is about 0.3)
- .25 moderate (r is about 0.5)
- .50 strong (r is about 0.7)
- 1 is perfect linear relationship

Standard error of the estimates

This gives us the idea how close the observations to the predicted values on the regression line. Given that approximately 68.3%, 95.4%, and 99.7% of the observations should be within , , and , respectively.

1.1 Inference on regression coefficients

1.1.1 To test whether there is a linear relationship between all independent variables taken together and dependent variable. This tests all the coefficients.

- Hypotheses:

(no linear relationships)
at least one of the coefficients is not zero (at least one independent variable *x* effects *y*)

- Test statistic (global *F* test):

Regression is significant if ; is the significance level
This means that the predictor variables are assumed to explain a significant fraction of the response variable.

1.1.2 To test which predictor variables are important

- Hypotheses:

0 (no linear relationship)
0 (linear relationship between and *y*)
The alternative hypothesis can be one-tailed or two-tailed. The given example is two-tailed alternative.

- Test statistic:

, with degree of freedom *N-k-1*

1.1.3 Confidence Interval

1.2 Interpretation of regression coefficients

is the mean change in the response variable for one unit change of the predictor variable when other predictors are held fixed. However, this might not reflect reality since one predictor tends to correlate with another so changing one predictor also results in changing another predictor.

1.3 Assessing Goodness of fit

Unlike simple linear regression, where we can observe the plot of the response and the predictor on a single plot, MLR tends to deal with higher dimensional variables. In general, a plot lets us deal with only two or three variables at a time, beyond that the interpretation can be complicated such as finding outliers.

1.4 Multi-colinearity

As mentioned earlier, there is more than one predictor in multiple regression, therefore, it is possible that one predictor may highly correlate with another predictor. This is known as multi-colinearity. Adding and removing a variable from the model with high correlation can greatly affect coefficients and significance of other predictors. As a result, the model tends to be unstable and complicates the interpretation.

result, the model tends to be unstable and complicates the interpretation.

Muli-colinearity is a situation when one or more independent variables have a linear relationship. Suppose there are two independent variables (e.g. and). If , and are constants. Then, we have multi-colinearity. This causes a problem because including both and in the model will not give additional information than including just one of them. Among the solutions, the simple one is deleting the predictor variables from the model. So, which predictor variables should we remove? There are several building models to deal with problems such as backward removal, forward entry, and stepwise.

Here are some methods to detect multicollinearity:

- Correlation matrix check for large correlations among independent variables

Regress each of the *x*(s) on all the other *x*(s) and examine whether there are any strong linear dependencies. This is also referred to as auxiliary regressions. If any of these \`R^2\`s is greater than \`R^2\` from the main model. This may indicate the problem.

- Calculate Variance Inflation Factors (VIF).   =

When is the \`R^2\` when regress on the remaining independent variables.
 measures how much the variance of your coefficients increased by multicollinearity compared to variance of coefficients when were independent of other explanatory variables.

2. Dummy Coding (Variables)

In the real world, many variables are discrete, for instance, gender (male, female), season (spring, summer, fall, winter), etc. Besides continuous predictor variables, MLR can also deal with categorical predictor variable. There are different ways to code categorical variables for regression: dummy coding, effects coding, and contrast coding. All of these methods, a categorical predictor with C level will be coded into C-1 ***different variables***.
The focus here is on Dummy coding. For this method, one group is selected as the reference group and a value of 0 is assigned to each of C-1 indicator variables. For example,

Dummy coding for Gender
(C=2)

Gender	
Male	1
Female	0

Dummy coding for Treatment
(C=3)

Group		
Treatment 1	0	1
Treatment 2	1	0
Control	0	0

Dummy coding for Field of Study
(C=4)

Field			
Education	0	0	1
Social Sciences	0	1	0
Sciences	1	0	0
Humanities	0	0	0

The choice of the reference group is arbitrary. General guidelines for selecting a reference group are:

- The reference group should provide a useful comparison such as a control group, a standard treatment, or a base group with expected highest or lowest score
- The reference group should be a meaningful group not ‘other’ category
- The sample size of the reference group should not small when compare to other groups for unequal sample sizes groups

Consider the dummy coding for field of study (4), three new dummy variables are created. The value of 0 and 1 is assigned to each field of study. In this case, humanities is selected as a reference group so all three dummy variables have a value of 0. For other three fields, one of the dummy variables has a value of 1 and 0 for other dummy variables. For instance,

When the field of study is Humanities
When , the field of study is Sciences.
When , the field of study is Social Sciences.
When , the field of study is Education.

Assume that the regression model is to predict the average salary. Therefore, the predicted salary from dummy-coded field of study equation using OLS for this data is:

When the appropriate values of dummy variables are substituted in this equation, different regression lines for Humanities, Sciences, Social sciences, and Education are obtained:

For Humanities: Predicted_average_salary =
=

For Sciences: Predicted_average_salary =
=

For Social Sciences: Predicted_average_salary =
=

For Education: Predicted_average_salary =
=

Parameter Interpretations:
is the average salary for Humanities (the reference group)

is the difference in average salary between Humanities and Sciences

is the difference in average salary between Humanities and Social Sciences

is the difference in average salary between Humanities and Education

The T-test of is testing whether the mean salary of the reference group (Humanities) is different from zero.

The T-test of is testing whether the mean salary of Sciences is significant different from Humanities.

The T-test of is testing whether the mean salary of Social Sciences is significant different from Humanities.

The T-test of is testing whether the mean salary of Education is significant different from Humanities.

If we want to test whether the average salary of Education (Sciences, Social Sciences) is different from zero, we need to re-code Education (Sciences, Social Sciences) as the reference group.
Example:

3. Interaction Terms

The MLR model used so far assumed that the effect on from one predictor (e.g.) is independent of the value of other predictors (). This means that one unit change in associated with y is the same regardless the values of other variables . It can be seen that each independent variable has appeared in the model separately (as an additive term also known as a ***main effect***) in regression.

However, it is possible that two or more predictor variables have interaction effects on the response variable, for instance, the effects of environment and exercise on stress. The linear model with two variables, the main effects and an interaction is in the form:

With an interaction, the slope of depends on the level of , and slope of , depends on the level of . For example,

- When is hold fixed, the regression can be represented as:

This suggests that for a given level of , the response change by for each unit change in .

- When is hold fixed, the regression can be represented as:

This suggested that for a given level of , the response change by , for each unit change in .

4. Methods for building regression models

In the real world there are many predictor variables, you might want to consider the formal process of building the regression model. Many statistical packages offer several choices or combine different methods for automated model building. This can be helpful for evaluating the importance of particular predictors. Two categories of model building are stepwise (include or retain within a model) and blocking (a group of variables to be considered in a particular step) method.

There are 3 basic stepwise methods for building a model

1. Backward removal - For this method, all predictor variables are included in the model. Next, remove the predictor one by one starting by the one with the least unique variance in the dependent variable that has the smallest partial correlation, and then the least variance of the remaining model and so on. The removal

unique variance in the dependent variable that has the smallest partial correlation, and then the least variance of the remaining model and so on. The removal processes continue until it is no longer a fit model (user can specify the criteria).

2. Forward entry - Predictor variables are included in the model one at a time. Start with the prediction that has the biggest absolute correlation with the dependent variable. As for the following, the predictor with the largest partial correlation with the predictor (i.e. variable that explained unique variance in the dependent variable the most) will be selected. The criterion of entering the model is based on improving the model fit or individual significance of the predictor.
3. Stepwise - This is a combination method of forward entry and backward removal. Predictors are added one at a time based on improving the model fit. When a new predictor is added, the predictors that already appeared in model are evaluated. The predictors that do not significantly improve the fit model will be removed.

5. Common Mistakes

- Not checking for linear relationship
  - simply use the scatter plot to visualize the relationship
  - if relationship is not linear, consider transformation (e.g. curvilinear regression)
- Relying too much on automated results from statistical analysis packages without verifying visually (e.g. scatter plots)
- Attaching numerical importance to regression parameters
  - Even small regression parameters may be meaningful
  - Units changing (e.g. meter to centimeters) can affect their magnitude greatly
- Not identifying confidence intervals for regression parameters
  - They came from sample not population
- Not identifying the coefficient of determination `R^2`
  - It is difficult to understand the quality of the regression model if `R^2` is left out
- Mixing up the coefficient (`R^2`) of determination with coefficient of correlation (*R*)
- Using predictor variables that are highly correlated
  - Should only include if regression increases significantly
- Utilizing regression for prediction further than the measured range
  - Linear relationship may not valid beyond the range
- Too many predictor variables
  - Use smaller subsets of predictor variables
- Collecting only a small subset of the operation range
  - May overlook other forms (e.g. non-linearity)

MLR with R examples can be found at:

MLR with R examples can be found at:

6. Other reminders:

- If independent variables are not measured in the same scale (e.g. meters, dollars, years), we cannot compare . However, standardized coefficients (i.e. transformed variables to mean of zero, and variance of 1) can be used to compare the strength of the predictors. By comparing standardized coefficients, we will know which independent variables are more important.

- If the interaction effect term is included in the model, the component variables should also be in the model (e.g. female*income, should include female and income in the model)

- Sometimes interaction terms are highly correlated with its components (multicollinearity). Be careful in eliminating a variable based only on its t-value

- Make sure there are enough cases

Some examples of Multiple Linear Regression using R:

- Example 1

- Example 2

Reference:

Lindquist, M (2011). Applied Linear Regression Analysis (Muliple Linear Regression)
Retrieved from: https://web.archive.org/web/20181222053620/http://www.stat.columbia.edu/~martin/W2024/R6.pdf

Mellor-Crummey, J (2005). Other Regression Models (slide). Rice University.
Retrieved from: https://web.archive.org/web/20190926002230/https://www.cs.rice.edu/~johnmc/comp528/lecture-notes/Lecture10.pdf

Williams, R. (2015) Review of Multiple Regression. University of Notre Dame.
Retrieved from: https://web.archive.org/web/20190926002450/https://www3.nd.edu/~rwilliam/stats2/l02.pdf

Reflect in ePortfolio Download Print

Activity Details

Task: View this topic