

Statistics Overview: Regression Analyses

Simple Linear Regression Analysis

Purpose

Simple linear regression analysis is used to determine if there is a linear relationship between two continuous variables, a predictor variable and a response variable. In a regression analysis, the predictor (independent) variable (x) is controlled or experimentally manipulated and data are collected on the response (dependent) variable (y). Consequently, a regression analysis estimates the response of the dependent variable to variation in the independent variable. In any type of regression analysis, a best-fit (least squares) linear regression equation is calculated such that the distances between the observed data points and the predicted values estimated by the regression equation are minimized. The simple linear regression equation is that of a line. Much like the t -test, a basic assumption of a regression analysis is that the variables are normally distributed for both the population and the sample. While there are alternative tests that can be used when this assumption are violated, we will not go into them here.

Graphical Representation

A scatter plot is often used to examine the linear relationship between the response and predictor variables. Each data point represents an individual measure of the response variable and the corresponding value of the predictor variable. In other words, it is a graph of the response variable plotted against the predictor variable. In addition to the data points, it is customary to provide a graphical representation of the best-fit line for the data.

Statistical hypotheses

H_0 : There is no statistically significant linear relationship between the response variable and the predictor variable.

H_A : There is a statistically significant linear relationship between the response variable and the predictor variable.

Report

1) Model parameters

The regression equation of the best-fit line is given in the format $y = \beta_0 + \beta_1x$, where β_0 is the intercept and β_1 is the slope of the line for the predicted linear relationship between y (response variable) and x (predictor variable).

2) Test statistic

We will not detail the method for calculating the test-statistic here. It is sufficient to know that the value you need from the statistical report is the F -statistic and it is based on deviations of the observations from the best-fit (least squares) line

3) Degrees of freedom

$(DFM, DFE) = (p, n-p) = [\text{degrees of freedom of model (Regression), degrees of freedom of error (Residual)}]$, where n = sample size and p = number predictor variables

4) p-value

$p \leq 0.05$ (always use this baseline p -value for the purpose of this course). In R, specific p -values are reported and you should assess the reported p -value against the 0.05 level.

5) Coefficient of determination

R^2 = proportion of variation in response variable explained by or due to the predictor variable. Values fall between 0 and 1 and a low value indicates that the predictor variable explains very little variation in the response variable.

Interpretation

In order to determine whether to reject the null hypothesis, it is necessary to compare the calculated F-statistic to a critical F-value. The critical F-value can be found in an F Distribution table (found in any statistics book) or may be indicated in the results report produced by the statistical software. In the case of R, the software compares the two values automatically and the critical F-value is not indicated in the report, just the p-value.

Calculated $F >$ critical F , $p \leq 0.05$:

We can reject the null hypothesis; we find significant statistical evidence for a linear relationship between the response variable and the predictor variable. The probability that we would find a linear relationship between the two variables due to chance is less than or equal to 5%, which is an acceptable level of error for ecological experiments.

Calculated $F \leq$ critical F , $p > 0.05$:

We fail to reject the null hypothesis; we fail to find significant statistical evidence for a linear relationship between the response variable and the predictor variable. The probability that we would find a linear relationship between the two variables due to chance is greater than or equal to 5%, which exceeds the acceptable level of error for ecological experiments.

Coefficient of determination (R^2):

A high R^2 indicates that the data points fall very closely along the best-fit line and that the independent variable is a good predictor of the dependent variable. In fact, R^2 can be interpreted as the proportion of variation in the response variable explained by the predictor variable. A low R^2 indicates that the data points are scattered away from the best-fit line and that the independent variable is a poor predictor of the dependent variable. It is possible to find evidence for a statistically significant relationship between two variables ($p < 0.05$) but a very low R^2 . In this case, you should indicate that while the linear relationship between the variables is significant, the independent variable is a poor predictor of the dependent variable. Unlike the F-statistic significance level (p-value), there is no standard cut-off value for a “low” versus “high” R^2 , but instead is at the discretion of the researcher.

In addition to providing an interpretation of the statistics, it is always necessary to indicate the direction of the relationship (positive or negative) and the biological significance of the relationship. Your interpretation of the model should also take into account both the F-value and the R^2 value. In particular, a model in which there is a significant linear relationship between the two variables (large F-value, low p-value) but a low R^2 is often indicative of unmeasured independent variables that also have an effect on the dependent variable. Simultaneous evaluation of multiple variables in a more inclusive model (Multiple Regression) may improve the fit and increase the R^2 of the overall model.

Multiple Regression Analysis

Purpose

Multiple regression analysis is very similar to a simple linear regression analysis and is used to determine if there is a relationship between a response variable and multiple predictor variables simultaneously. In a multiple regression, a best-fit (least squares) regression equation is calculated such that the distances between the observed data points and the predicted values estimated by the multiple regression equation are minimized. This method is also used to estimate the significance of the relationship between each predictor variable and the response variable while controlling for variation that can be attributed to the other predictor variables.

Graphical Representation

The number of dimensions of the multiple regression is determined by the number of predictor variables e.g., a plane for two predictor variables or a cube for three. It becomes impossible to visualize this multi-dimensional space with more than three predictor variables. You will not be asked to graph anything more than scatter plot of one response variable against one predictor variable.

Statistical hypotheses

Full Model

H_0 : There is no statistically significant relationship between the response variable and the predictor variables as a group.

H_A : There is a statistically significant relationship between the response variable and the predictor variables as a group.

Residual Models

H_0 : There is no statistically significant relationship between the response variable and the specific predictor variable while controlling for the effects of the other predictor variables.

H_A : There is a statistically significant relationship between the response variable and the specific predictor variable while controlling for the effects of the other predictor variables.

Report

1) Model parameters

The regression equation is in the format $y = \beta_0 + \beta_1x_1 + \beta_2x_2 \dots \beta_nx_n$, where β_0 is a constant, β_1 is the regression coefficient for the predicted relationship between x_1 (first predictor variable) and y (response variable), β_2 is the regression coefficient for the predicted relationship between x_2 (second predictor variable) and y (response variable), up to β_n , the regression coefficient for the predicted relationship between x_n (nth predictor variable) and y (response variable).

2) Test statistics

You must report F-statistic for the full model including all of the predictor variables with the appropriate degrees of freedom. You also need to report the t-statistic calculated separately for each of the predictor variables with the appropriate degrees of freedom.

3) Degrees of freedom

F-statistic: (DFM, DFE) = (p, n-p) = [degrees of freedom of model (Regression), degrees of freedom of error (Residual)], where n = sample size and p = number independent variables

t-statistic: (1, n-1). t-statistic: (DFM, DFE) = (1, n-1) = [degrees of freedom of model (Regression), degrees of freedom of error (Residual)], where n = sample size.

4) p-value

$p \leq 0.05$ (always use this baseline p-value for the purpose of this course). In R, specific p-values are reported and you should assess the reported p-value against the 0.05 level.

5) Coefficient of determination

R^2 = proportion of variation in response variable explained by or due to all of the predictor variables included in the model. Values fall between 0 and 1 and a low value indicates that the independent variables explain very little variation in the dependent variable.

Interpretation

For the multiple regression analysis, it is necessary to first consider the statistical significance of the full model before going on to interpret the significance of the individual predictor variables. If you fail to find a statistically significant relationship between the response variable and all the predictor variables simultaneously, then you fail to reject the null hypothesis for the full model and consequently must conclude that the model as a whole is not a good fit for the data. It is not appropriate to interpret the effects of the individual predictor variables from a full model that has been rejected. However, if you do find evidence for a significant relationship between the response variable and all the predictor variables as a group, then it is appropriate to interpret the effect of each of the individual predictor variables (residual models).

Full Model

Calculated $F > \text{critical } F$, $p \leq 0.05$

We can reject the null hypothesis for the full model; we find significant statistical evidence for a relationship between the response variable and all the predictor variables simultaneously (as a group). The probability that we would find a relationship among the variables due to chance is less than or equal to 5%, which is an acceptable level of error for ecological experiments.

Calculated $F \leq \text{critical } F$, $p > 0.05$

We fail to reject the null hypothesis; we fail to find significant statistical evidence for a relationship between the response variable and all the predictor variables simultaneously (as a group). The probability that we would find a relationship among the variables due to chance is less than or equal to 5%, which is an acceptable level of error for ecological experiments.

Residual Models

Calculated $t > \text{critical } t$, $p \leq 0.05$

We can reject the null hypothesis; we find significant statistical evidence for a relationship between the response variable and the predictor variable controlling for variation due to the other predictor variables. The probability that we would find a relationship between the variables due to chance is less than or equal to 5%, which is an acceptable level of error for ecological experiments.

Calculated $t \leq \text{critical } t$, $p > 0.05$

We fail to reject the null hypothesis; we fail to find significant statistical evidence for a relationship between the response variable and the predictor variable controlling for variation due to the other predictor variables. The probability that we would find a relationship between the variables due to chance is less than or equal to 5%, which is an acceptable level of error for ecological experiments.

Coefficient of determination (R^2):

A high R^2 indicates that the data points are close to the values predicted by the multiple regression equation and that as a group the independent variables are a good predictor of the dependent variable. In a multiple regression, R^2 is interpreted as the proportion of variation in the response variable explained by all the predictor variables simultaneously. A low R^2 indicates that the data points are scattered away from the values predicted by the multiple regression equation and that the independent variables are a poor predictor of the dependent variable. It is possible to find evidence for a statistically significant relationship between the response variable and all the predictor variables ($p < 0.05$) but a very low R^2 . In this case, you should indicate that while the relationship is significant, the independent variables are a poor predictor of the dependent variable.

In addition to providing an interpretation of the statistics, it is always necessary to indicate the direction of the relationships (positive or negative) and the biological significance of the relationships. Your interpretation of the model should also take into account both the F-value and the R^2 value. In particular, a model in which there is a significant relationship among the variables (large F-value) but a low R^2 is often indicative of still unmeasured independent variables that also have an effect on the dependent variable.