OCTOBER 27, 2020 BY ZACH

Introduction to Multiple Linear Regression

When we want to understand the relationship between a single predictor variable and a response variable, we often use simple linear regression.

However, if we'd like to understand the relationship between *multiple* predictor variables and a response variable then we can instead use **multiple linear regression**.

If we have *p* predictor variables, then a multiple linear regression model takes the form:

$$\mathbf{Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon}$$

where:

- **Y**: The response variable
- $\mathbf{X_j}$: The j^{th} predictor variable
- $\boldsymbol{\beta_j}$: The average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- **ε**: The error term

The values for β_0, β_1, B_2, … , β_p are chosen using **the least square method**, which minimizes the sum of squared residuals (RSS):

$$\mathbf{RSS = \Sigma(y_i - \hat{y}_i)^2}$$

where:

- **Σ**: A greek symbol that means *sum*
- $\mathbf{y_i}$: The actual response value for the i^{th} observation
- $\mathbf{\hat{y}_i}$: The predicted response value based on the multiple linear regression model

The method used to find these coefficient estimates relies on matrix algebra and we will not cover the details here. Fortunately, any statistical software can calculate these coefficients for you.

How to Interpret Multiple Linear Regression Output

Suppose we fit a multiple linear regression model using the predictor variables *hours studied* and *prep exams taken* and a response variable *exam score*.

The following screenshot shows what the multiple linear regression output might look like for this model:

Note: *The screenshot below shows multiple linear regression output for Excel, but the numbers shown in the output are typical of the regression output you'll see using any statistical software.*

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.857
R Square	0.734
Adjusted R Square	0.703
Standard Error	5.366
Observations	20

ANOVA

	df	SS	MS	F	Significance F
Regression	2	1350.76	675.38	23.46	0.00
Residual	17	489.44	28.79		
Total	19	1840.20			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	67.67	2.82	24.03	0.00	61.73	73.61
hours	5.56	0.90	6.18	0.00	3.66	7.45
prep_exams	-0.60	0.91	-0.66	0.52	-2.53	1.33

From the model output, the coefficients allow us to form an estimated multiple linear regression model:

Exam score = 67.67 + 5.56*(hours) – 0.60*(prep exams)

The way to interpret the coefficients are as follows:

- Each additional one unit increase in hours studied is associated with an average increase of **5.56** points in exam score, *assuming prep exams is held constant.*
- Each additional one unit increase in prep exams taken is associated with an average decrease of **0.60** points in exam score, *assuming hours studied is held constant.*

We can also use this model to find the expected exam score a student will receive based on their total hours studied and prep exams taken. For example, a student who studies for 4 hours and takes 1 prep exam is expected to score a **89.31** on the exam:

Exam score = 67.67 + 5.56*(4) -0.60*(1) = **89.31**

Here is how to interpret the rest of the model output:

- **R-Square:** This is known as the coefficient of determination. It is the proportion of the variance in the response variable that can be explained by the explanatory variables. In this example, 73.4% of the variation in the exam scores can be explained by the number of hours studied and the number of prep exams taken.
- **Standard error:** This is the average distance that the observed values fall from the regression line. In this example, the observed values fall an average of 5.366 units from the regression line.
- **F:** This is the overall F statistic for the regression model, calculated as regression MS / residual MS.
- **Significance F:** This is the p-value associated with the overall F statistic. It tells us whether or not the regression model as a whole is statistically significant. In other words, it tells us if the two explanatory variables combined have a statistically significant association with the response variable. In this case the p-value is less than 0.05, which indicates that the explanatory variables hours studied and prep exams taken combined have a statistically significant association with exam score.
- **Coefficient P-values.** The individual p-values tell us whether or not each explanatory variable is statistically significant. We can see that hours studied is statistically significant ($p = 0.00$) while prep exams taken ($p = 0.52$) is not statistically significant at $\alpha = 0.05$. Since prep exams taken is not statistically significant, we may end up deciding to remove it from the model.

How to Assess the Fit of a Multiple Linear Regression Model

There are two numbers that are commonly used to assess how well a multiple linear regression model “fits” a dataset:

1. R-Squared: This is the proportion of the variance in the response variable that can be explained by the predictor variables.

The value for R-squared can range from 0 to 1. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable.

The higher the R-squared of a model, the better the model is able to fit the data.

2. Standard Error: This is the average distance that the observed values fall from the regression line. The smaller the standard error, the better a model is able to fit the data.

If we’re interested in making predictions using a regression model, the standard error of the regression can be a more useful metric to know than R-squared because it gives us an idea of how precise our predictions will be in terms of units.

For a complete explanation of the pros and cons of using R-squared vs. Standard Error for assessing model fit, check out the following articles:

- What is a Good R-squared Value?
- Understanding the Standard Error of a Regression Model

Assumptions of Multiple Linear Regression

There are four key assumptions that multiple linear regression makes about the data:

1. Linear relationship: There exists a linear relationship between the independent variable, x, and the dependent variable, y.

2. Independence: The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

3. Homoscedasticity: The residuals have constant variance at every level of x.

4. Normality: The residuals of the model are normally distributed.

For a complete explanation of how to test these assumptions, check out this article.

Multiple Linear Regression Using Software

The following tutorials provide step-by-step examples of how to perform multiple linear regression using different statistical software:

How to Perform Multiple Linear Regression in R
How to Perform Multiple Linear Regression in Python
How to Perform Multiple Linear Regression in Excel
How to Perform Multiple Linear Regression in SPSS

How to Perform Multiple Linear Regression in Stata
How to Perform Linear Regression in Google Sheets

Published by Zach

View all posts by Zach

PREV

How to Perform Simple Linear Regression in Python (Step-by-Step)

NEXT

Introduction to Logistic Regression

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment *

Name *

Email *

Website

POST COMMENT

SEARCH

Search …

ABOUT

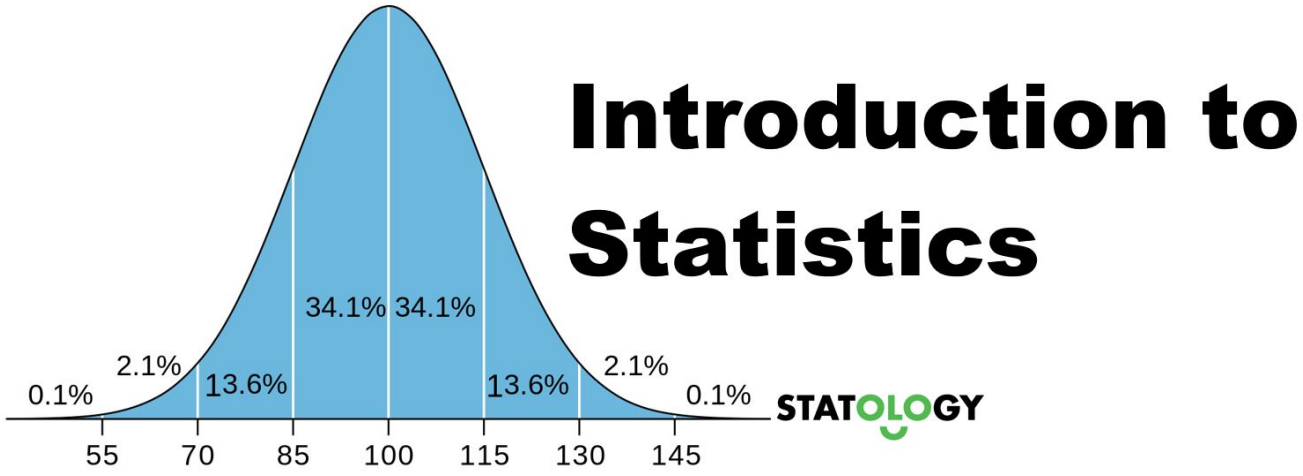
Statology is a site that makes learning statistics easy by explaining topics in simple and straightforward ways. **Learn more about us**.

STATOLOGY STUDY

Statology Study is the ultimate online statistics study guide that helps you study and practice all of the core concepts taught in any elementary statistics course and makes your life so much easier as a student.

INTRODUCTION TO STATISTICS COURSE

Introduction to Statistics is our premier online video course that teaches you all of the topics covered in introductory statistics. **Get started** with our course today.



RECENT POSTS

Pandas: How to Groupby Range of Values

Pandas: How to Use groupby() with size()

Pandas: Use Groupby to Calculate Mean and Not Ignore NaNs