What are the significant variables in regression models?

Authored by

stats writer

June 23, 2024

Regression models are statistical tools used to analyze the relationship between a dependent variable and one or more independent variables. The significant variables in regression models are those that have a significant impact on the dependent variable. These variables play a crucial role in predicting the outcome of the dependent variable and their inclusion or exclusion can greatly affect the accuracy and reliability of the model. Some of the important variables in regression models include the strength of the relationship between the dependent and independent variables, the level of significance, and the coefficient of determination. Other factors such as multicollinearity, heteroscedasticity, and autocorrelation can also greatly influence the results of a regression model. Therefore, it is essential to carefully select and analyze the significant variables in order to build a reliable and accurate regression model.

Determine Significant Variables in Regression Models

One of the main questions you'll have after fitting a is: *Which variables are significant?*

There are two methods you should not use to determine variable significance:

1. The value of the regression coefficients

A regression coefficient for a given predictor variable tells you the average change in the response variable associated with a one unit increase in that predictor variable.

However, each predictor variable in a model is usually measured on a different scale so it doesn't make sense to compare the absolute values of the regression

coefficients to determine which variables are most important.

2. The p-values of the regression coefficients

The p-values of the regression coefficients can tell you if a given predictor variable has a statistically significant association with the response variable, but they can't tell you if a given predictor variable is practically significant in the real world.

P-values can also be low due to a large sample size or low variability, which doesn't actually tell us whether or not a given predictor variable is practically significant.

However, there are two methods you should use to determine variable significance:

1. Standardized Regression Coefficients

Typically when we perform multiple linear regression, the resulting regression coefficients in the model output are unstandardized, meaning they use the raw data to find the line of best fit.

However, it's possible to standardize each predictor

variable and the response variable (by subtracting the mean value of each variable from the original values and then dividing by the variables standard deviation) and then perform regression, which results in standardized regression coefficients.

By standardizing each variable in the model, each variable becomes measured on the same scale. Thus, it makes sense to compare the absolute values of the regression coefficients in the output to understand which variables have the greatest effect on the response variable.

2. Subject Matter Expertise

While p-values can tell you if there is a statistically significant effect between a given predictor variable and the response variable, subject matter expertise is needed to confirm whether or not a predictor variable is actually relevant and should actually be included in a model.

The following example shows how to determine significant variables in a regression model in practice.

Example: How to Determine Significant Variables in Regression Model

Age	Sq. Footage	Price
4	2600	$ 280,000
7	2800	$ 340,000
10	1700	$ 195,000
15	1300	$ 180,000
16	1500	$ 150,000
18	1800	$ 200,000
24	1200	$ 180,000
28	2200	$ 240,000
30	1800	$ 200,000
35	1900	$ 180,000
40	2100	$ 260,000
44	1300	$ 140,000

Suppose we then perform multiple linear regression, using age and square footage as the predictor variables and price as the response variable.

We receive the following output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	34736.543	37184.321	0.934	0.375
Age	-409.833	612.458	-0.669	0.520
Sq. Footage	100.866	15.747	6.405	0.000

The regression coefficients in this table are unstandardized, meaning they used the raw data to fit this regression model.

Upon first glance, it appears that age has a much larger effect on house price since it's coefficient in the regression table is -409.833 compared to just 100.866 for the predictor variable square footage.

However, the standard error is much larger for age compared to square footage, which is why the corresponding p-value is actually large for age (p=0.520) and small for square footage (p=0.000).

The reason for the extreme differences in regression coefficients is because of the extreme differences in scales for the two variables:

The values for age range from 4 to 44.The values for square footage range from 1,200 to 2,800.

Suppose we instead standardize the raw data:

Raw Data

Age	Sq. Footage	Price
4	2600	$ 280,000
7	2800	$ 340,000
10	1700	$ 195,000
15	1300	$ 180,000
16	1500	$ 150,000
18	1800	$ 200,000
24	1200	$ 180,000
28	2200	$ 240,000
30	1800	$ 200,000
35	1900	$ 180,000
40	2100	$ 260,000
44	1300	$ 140,000

Standardized Data

Age	Sq. Footage	Price
-1.425	1.479	1.175
-1.195	1.873	2.212
-0.965	-0.296	-0.295
-0.581	-1.084	-0.555
-0.505	-0.690	-1.074
-0.351	-0.099	-0.209
0.109	-1.281	-0.555
0.415	0.690	0.483
0.569	-0.099	-0.209
0.952	0.099	-0.555
1.335	0.493	0.829
1.642	-1.084	-1.247

If we then perform multiple linear regression using the standardized data, we'll get the following regression output:

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	0.000	0.123	0.000	1.000
Age	-0.092	0.138	-0.669	0.520
Sq. Footage	0.885	0.138	6.405	0.000

The regression coefficients in this table are standardized, meaning they used standardized data to fit this regression model.

The way to interpret the coefficients in the table is as follows:

A one standard deviation increase in age is associated with a 0.092 standard deviation decrease in house price, assuming square footage is held constant.A one standard deviation increase in square footage is associated with a 0.885 standard deviation increase in house price, assuming age is held constant.

Now we can see that square footage has a much larger effect on house price than age.

Note: The p-values for each predictor variable are the exact same as the previous regression model.

When deciding on the final model to use, we now know that square footage is much more important for predicting the price of a house compared to age.

Ultimately we would need to use subject matter expertise to determine which variables to include in the final model based on existing knowledge about real estate and house prices.

The following tutorials provide additional information about regression models:

How to Read and Interpret a Regression Table

How to Interpret Regression Coefficients