Multiple Regression

by Sophia

WHAT'S COVERED

This tutorial will cover the topic of multiple regression. Our discussion breaks down as follows:

1. Multiple Regression

1. Multiple Regression

Multiple regression is going to allow you to predict a response based on more than one explanatory variable, although they have to be independent.

EXAMPLE In many school districts, teacher salaries are dependent on two variables: years of experience and number of postgraduate hours accumulated.

It's possible that a teacher with a lot of years of experience might not have a high number of postgrad hours. It’s also possible that someone with a lot of postgrad hours doesn't have a whole lot of experience. Consider the table below with those three variables--salary, years of experience, and postgrad hours--listed for Mr. Backman, Mr. Jones, Ms. Nordstrom, Mr. Osters, and Ms. Williams.

Teacher	Salary	Years	Hours
Backman	38,000	4	14
Jones	42,000	3	45
Nordstrom	59,000	10	55
Osters	44,000	6	28
Williams	48,000	5	39

We can use this information to come up with three different linear regressions models:

- Model A: Salary vs. Years
- Model B: Salary vs. Hours

- Model C: Salary vs. Both Years and Hours

Model A		
Variables	**Regression Line**	**Coefficient of Determination (r^2)**
Explanatory: Years **Response:** Salary	$\widehat{salary} = 31{,}164 + 2{,}684(years)$ A starting salary for someone with no years of experience is $31,164. For every additional year that a person works, they are predicted to make an additional $2,685 on average.	$r^2 = 0.83$ If you look at the r-squared for this, it's fairly high at 0.83. It's clear there's something of an association here between salary and years.

Model B		
Variables	**Regression Line**	**Coefficient of Determination (r^2)**
Explanatory: Hours **Response:** Salary	$\widehat{salary} = 31{,}384 + 409(hours)$ A starting salary for someone with no postgrad hours is $31,384. For each additional postgrad hour, they are predicted to make an additional $409 on average.	$r^2 = 0.65$ The r-squared here isn't as high, so there's a little bit less of an association between postgraduate hours and salary than the one with years.

Model C		
Variables	**Regression Line**	**Coefficient of Determination (r^2)**
Explanatory: Years and Hours **Response:** Salary	$\widehat{salary} = 26{,}807 + 1{,}970(years) + 23(hours)$ A starting salary for someone with no years of experience and no postgrad hours is $26,807. For every additional year that a person works, they are predicted to make an additional $1,970 on average. For each additional postgrad hour, they are predicted to make an additional $23 on average.	$r^2 = 0.97$ The r-squared value is higher than either of the two individual linear regressions.

For multiple regression, if those variables are independent, then you can do a regression on both variables, like in Model C.

The predicted salary is going to have some part that has a constant, some coefficient for the number of years that the teacher has, and some coefficient for the number of postgrad hours that that teacher has accumulated.

Look at the r-squared value for Model C. It's higher than either of the two individual linear regressions. Every time you add an independent variable, the r-squared would continue to increase. It will always go up when you add another variable because more of the variability in salary is going to be explained by an additional variable.

Look how well these models did. These lists below indicate the residuals for each model, which is how far off each model was in predicting the teacher's salary.

Teacher	Salary	Years	Hrs	Model A	Model B	Model C
Backman	38,000	4	14	-3,904	890	80
Jones	42,000	3	45	2,781	-7,789	-1,112
Nordstrom	59,000	10	55	986	5,121	-210
Osters	44,000	6	28	-3,274	1,164	-1,094
Williams	48,000	5	39	3,411	665	2,335

If you look at Model A, the residuals indicate that the predicted values were somewhat off from the actual values. Model A under-predicted Mr. Backman's salary by nearly $4,000 and over-predicted Mr. Jones' salary by about $2,800.

If you look at Model B, these residuals are fairly big. Mr. Jones' salary was under-predicted by nearly $8,000 in Model B, and Ms. Nordstrom's salary was over-predicted by over about $5,000.

If you look at Model C, on average, these residuals are much smaller than those of Model A or Model B. There was only one teacher that had a better prediction from either Model A or Model B than from Model C. Overall, Model C, the one from multiple regression, is the most accurate model.

TERM TO KNOW

Multiple Regression

Using more than one explanatory variable to predict the value of the response variable.

SUMMARY

Multiple regression is going to allow us to use more than one explanatory variable to predict the response. Those explanatory variables must be independent. This allows for certain variables to have a larger effect on the response than others, but still shows what those effects are and allows us to explain more of the variation in the response, increasing the r-squared value. By adding a second explanatory variable independent of the first, or a third independent of the first two, etc., the value of r-squared will increase.

Good luck!

Source: THIS TUTORIAL WAS AUTHORED BY JONATHAN OSTERS FOR SOPHIA LEARNING. PLEASE SEE OUR **TERMS OF USE**.

TERMS TO KNOW

Multiple Regression
Using more than one explanatory variable to predict the value of the response variable.