

Regression and Correlation

SPSS

A major objective of many statistical investigations is to establish relationships that make it possible to predict one variable in terms of others. Thus, studies are made to explain how much the economy will grow if illiteracy rates fall or what socio-economic factors might boost immigration.

Let us assume we are trying to determine the gas mileage of a car. We can reasonably presume that the car's horsepower will be a decisive factor in predicting the mileage. In this case, we will use simple regression to determine whether these two variables are significantly related, and the direction and strength of their relationship. However, if we think that other variables might affect mileage (such as the weight of the vehicle, its engine displacement), we will use multiple regression. From a statistical point of view, note that we will refer to gas mileage as the *dependent variable*; it will always be the Y variable in our regression, the variable we aim to predict. The remaining variables (weight, horsepower, engine displacement) will be the *independent variables* or the *explanatory variables*; they are the X 's in our equation, the variables that will help us explain gas mileage.

1 Simple Regression

Simple regression analyzes the relationship between two variables.

1.1 Scatterplot

The relationship of two variables can be portrayed in a **scatterplot**. A scatterplot is merely a plot of the data points of two variables. It is conventional to show the dependent variable on the vertical axis and the independent variable on the horizontal axis. The relationship between the two variables is estimated as a straight line relationship, defined by the equation: $Y = aX + b$, where b is the intercept or the constant and a , the slope. The line is mathematically calculated such that the sum of distances from each observation to the line is minimized¹. By definition, the slope indicates the change in Y as a result of a unit change in X . The straight line is also called the *regression line* or the *fit line* and a is referred to as the *regression coefficient*.

¹ The method of calculating the regression coefficient (the slope) is called *ordinary least squares*, or OLS. OLS estimates the slope by minimizing the sum of squared differences between each predicted ($aX + b$) and the actual value of Y . One reason for squaring these distances is to ensure that all distances are positive.

A positive regression coefficient indicates a positive relationship between the variables; the fit line will be upward sloping. A negative regression coefficient indicates a negative relationship between the variables; the fit line will be downward sloping.

To draw a scatterplot using SPSS, go to Graphs/Scatter/Dot... Then, select the variables you are interested in, say the gas mileage (miles per gallon) as your dependent variable and the horsepower as your independent variable, and OK. If you want to add the fit line to your chart, double click on the diagram. A new window – the chart editor – will appear. Then, go to Elements/Fit Line at Total... Then, close the Properties window and the Chart Editor. You should observe your scatter diagram with the fit line.

1.2 Fit of Model

The test of significance of the regression slope is a key test of hypothesis regression analysis that tells us whether the slope a is statistically different from 0 . To determine whether the slope equals zero, a t-test is performed. Without going into too much detail, the idea of this test is to test a hypothesis $H_0: a = 0$, i.e. the two variables are unrelated, versus $H_1, a \neq 0$, i.e. the two variables are related. When observations on the scatterplot lie closely around the fit line, the regression line is more likely to be statistically significant: in other words, it will be more likely that the two variables are – positively or negatively – related. SPSS calculates the slope, the intercept, standard error of the slope, and the level at which the slope is statically significant.

Let us go back to our example. We aim to determine whether gas mileage and horsepower are related. Go to Analyze/Regression/Linear... Gas mileage is our dependent variable and horsepower, our independent variable. Click OK. A series of tables will show up in our output window.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.771(a)	.595	.594	4.974

a Predictors: (Constant), Horsepower

In the first table, the R^2 , also called the coefficient of determination, is of great interest. It measures the proportion of the total variation in Y about its mean explained by the regression of Y on X . In this case, our regression explains 59.5 % of the variation of gas mileage. We will explain the Adjusted R^2 in the next section. Typically, values of R^2 below 0.2 are considered weak, between 0.2 and 0.4, moderate, and above 0.4, strong.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	14169.756	1	14169.756	572.709	.000 ^a
	Residual	9649.237	390	24.742		
	Total	23818.993	391			

a. Predictors: (Constant), Horsepower

b. Dependent Variable: Miles per Gallon

In the second table, we will focus on the *F*-statistic. By computing this statistic, we test the hypothesis that none of the explanatory variables help explain variation in *Y* about its mean. The information to pay attention to here is the probability – Sig. in the table. If this probability is below *0.05*, we conclude that the *F*-statistic is large enough so that we can reject the hypothesis that none of the explanatory variables help explain variation in *Y*. This test is like a test of significance of the R^2 .

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	39.855	.730		54.578	.000
	Horsepower	-.157	.007	-.771	-23.931	.000

a. Dependent Variable: Miles per Gallon

Finally, the last table will help us determine whether gas mileage and horsepower are significantly related, and the direction and strength of their relationship. Let us ignore the standardized coefficients. The first important thing to note is that the sign of the coefficient of horsepower is negative. It confirms our intuition (powerful cars consume more gasoline) and our visual analysis of the scatterplot. Furthermore, the probability reported in the left column is very low. This implies that the slope *a* is statistically significant. To be less abstract, let us recall what those coefficients mean: they are the slope and the intercept of the regression line, i.e. $Y = -0.157X + 39.855$. What does this mean? It simply means that when horsepower increases by one unit (i.e. *1* horse), mileage will – on average – fall by *0.157* miles per gallon.

In sum, R^2 is high, probabilities are low: WE ARE HAPPY!

2 Multiple Regression

Multiple regression is a natural extension of simple regression. It is the right tool whenever you think that a variable is explained by several different variables. In our empirical work, we reasonably assume that gas mileage (*Y*) is not only explained by

horsepower (X_1) but also the weight (X_2) of the car and its engine displacement (X_3). Therefore, we will test the following equation:

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + error$$

where Y is the dependent variable, a_0 is the constant, X_1 , X_2 , and X_3 are the independent variables and their respective coefficients a_1 , a_2 , and a_3 , and the error term reflects all other factors that are not in the model.

Before any statistical work, do not forget to draw scatterplots of every independent variable against the dependent variable. Hence, you should observe at least three different scatterplots (mileage versus horsepower, mileage versus weight and mileage versus engine displacement). It is important to do this to see if you can visually detect a relationship between your variables. For instance, you might observe a non-linear relationship between two variables, in which case, you should use different techniques.

2.1 Correlation Matrix

Before starting our multiple regression analysis, it is important to compute the correlation matrix. Go to Analyze/Correlate/Bivariate... We select the three independent variables. We then click OK. The following table will show up in the output window.

Correlations

		Engine Displacement (cu. inches)	Horsepower	Vehicle Weight (lbs.)
Engine Displacement (cu. inches)	Pearson Correlation	1	.897**	.933**
	Sig. (2-tailed)		.000	.000
	N	406	400	406
Horsepower	Pearson Correlation	.897**	1	.859**
	Sig. (2-tailed)	.000		.000
	N	400	400	400
Vehicle Weight (lbs.)	Pearson Correlation	.933**	.859**	1
	Sig. (2-tailed)	.000	.000	
	N	406	400	406

** . Correlation is significant at the 0.01 level (2-tailed).

This preliminary step is important because independent variables should NOT be correlated with one another. If independent variables are correlated, this might affect the robustness of our results. In the fascinating world of statistics, this is referred to as the issue of multicollinearity. To keep things simple, when we use multiple regression analysis, we attach a weight to any of the independent variables in order to explain the variation in Y . If two independent variables are strongly correlated, it becomes very hard to attach a weight to those variables because they basically convey the same information.

As a result, the validity of our empirical work will be greatly affected. In general, assuming two of the independent variables are correlated, the easiest solution is to ignore one of them variables one and to use simply the other one. In this particular case, all variables are strongly correlated with one another. I would recommend you should use only the horsepower (which we did in the previous section) and not use engine displacement nor weight of the vehicle. I will still run our multiple regression to explain its basic philosophy, even though it is very BAD!

2.2 Goodness of Fit

To run the regression, go to Analyze/Regression/Linear... Select mileage as the dependent variable and horsepower, weight, and engine displacement as the independent variables. Click OK. The following tables show up in the output window.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.822 ^a	.676	.673	4.461

a. Predictors: (Constant), Vehicle Weight (lbs.), Horsepower, Engine Displacement (cu. inches)

The R_2 slightly greater than the R_2 we obtained in the former section. However, the variation is not that large considering that we added two new variables. Clearly, this is due to the fact that the new independent variables are strongly correlated and ultimately, do not bring much extra information. One of the flaw of the R_2 is that it is sensitive to the number of included independent variables. Specifically, addition of additional independent variables can only increase the R_2 . In contrast, the Adjusted R_2 accounts for the number of independent variables. It may rise or fall with the addition of more variables. The Adjusted R_2 is greater than the one obtained in the former section. Therefore, the extra information brought by the new variables is greater than the penalty of adding variables (assuming that we did not encounter the issue of multicollinearity).

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16098.158	3	5366.053	269.664	.000 ^a
	Residual	7720.836	388	19.899		
	Total	23818.993	391			

a. Predictors: (Constant), Vehicle Weight (lbs.), Horsepower, Engine Displacement (cu. inches)

b. Dependent Variable: Miles per Gallon

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	
	B	Std. Error	Beta			
1	(Constant)	44.015	1.272		34.597	.000
	Horsepower	-.056	.013	-.273	-4.153	.000
	Vehicle Weight (lbs.)	-.005	.001	-.504	-6.186	.000
	Engine Displacement (cu. inches)	-.006	.007	-.074	-.786	.432

a. Dependent Variable: Miles per Gallon

Finally, as expected from our scatterplots, we find out that all independent variables are negatively correlated with gas mileage. However, the coefficient on engine displacement is not statistically significant. This implies that gas mileage and engine displacement are unrelated.