



Search DSS

[Home](#) → [Online Help](#) → [Analysis](#) → Interpreting Regression Output

  
   
☒ DSS

# Interpreting Regression Output

## Finding Data

- [Data](#)
- [Subject specialists](#)

## Analyzing Data

- [Software](#)
  - [Stata](#)
  - [R](#)
- [Getting Started](#)
- [Consultants](#)

## Citing data

## About Us

### DSS lab consultation schedule

(Monday-Friday)

Sep 1-Nov 3 [By appt. here](#)

Nov 6-Dec 15 [Walk-in, 2-5 pm\\*](#)

Dec 18-Feb 2 [By appt. here](#)

Feb 5-May 4 [Walk-in, 1-5 pm\\*](#)

May 7-May 15 [Walk-in, 2-5 pm\\*](#)

May 16-Aug 31 [By appt. here](#)

For quick questions email [data@princeton.edu](mailto:data@princeton.edu).

\*No appts. necessary during walk-in hrs.

**Note:** the DSS lab is open as long as Firestone is open, no appointments necessary to use the lab computers for your own analysis.



- [Introduction](#)
- [P, t and standard error](#)
- [Coefficients](#)
- [R squared and overall significance of the regression](#)
- [Linear regression \(guide\)](#)
- [Further reading](#)

## Introduction

This guide assumes that you have at least a little familiarity with the concepts of linear multiple regression, and are capable of performing a regression in some software package such as Stata, SPSS or Excel. You may wish to read our companion page [Introduction to Regression](#) first. For assistance in performing regression in particular software packages, there are some resources at [UCLA Statistical Computing Portal](#).

## Brief review of regression

Remember that regression analysis is used to produce an equation that will predict a dependent variable using one or more independent variables. This equation has the form

$$Y = b_1X_1 + b_2X_2 + \dots + A$$

where  $Y$  is the dependent variable you are trying to predict,  $X_1$ ,  $X_2$  and so on are the independent variables you are using to predict it,  $b_1$ ,  $b_2$  and so on are the coefficients or multipliers that describe the size of the effect the independent variables are having on your dependent variable  $Y$ , and  $A$  is the value  $Y$  is predicted to have when all the independent variables are equal to zero.

In the Stata regression shown below, the prediction equation is  $\text{price} = -294.1955 (\text{mpg}) + 1767.292 (\text{foreign}) + 11905.42$  - telling you that price is predicted to increase 1767.292 when the foreign variable goes up by one, decrease by 294.1955 when mpg goes up by one, and is predicted to be 11905.42 when both mpg and foreign are zero.

Stata Results

. reg price mpg foreign

Source	SS	df	MS		
Model	180261702	2	90130850.8	Number of obs =	74
Residual	454803695	71	6405685.84	F( 2, 71) =	14.07
				Prob > F =	0.0000
				R-squared =	0.2838
				Adj R-squared =	0.2637
Total	635065396	73	8699525.97	Root MSE =	2530.9

the coefficients (betas)

price	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
mpg	-294.1955	55.69172	-5.28	0.000	-405.2417	-183.1494
foreign	1767.292	700.158	2.52	0.014	371.2169	3163.368
_cons	11905.42	1158.634	10.28	0.000	9595.164	14215.67

the constant (alpha)

Coming up with a prediction equation like this is only a useful exercise if the independent variables in your dataset have some correlation with your dependent variable. So in addition to the prediction

components of your equation--the coefficients on your independent variables (betas) and the constant (alpha)--you need some measure to tell you how strongly each independent variable is associated with your dependent variable.

When running your regression, you are trying to discover whether the coefficients on your independent variables are really different from 0 (so the independent variables are having a genuine effect on your dependent variable) or if alternatively any apparent differences from 0 are just due to random chance. The null (default) hypothesis is always that each independent variable is having absolutely no effect (has a coefficient of 0) and you are looking for a reason to reject this theory.

## P, t and standard error

The *t* statistic is the *coefficient* divided by its *standard error*. The standard error is an estimate of the *standard deviation* of the coefficient, the amount it varies across cases. It can be thought of as a measure of the precision with which the regression coefficient is measured. If a coefficient is large compared to its standard error, then it is probably different from 0.

How large is large? Your regression software compares the *t* statistic on your variable with values in the *Student's t distribution* to determine the *P* value, which is the number that you really need to be looking at. The Student's *t* distribution describes how the mean of a sample with a certain number of observations (your *n*) is expected to behave.

**Stata Results**

**. reg price mpg foreign**

**P value for the regression as a whole**

Source	SS	df	MS
Model	180261702	2	90130850.8
Residual	454803695	71	6405685.84
Total	635065396	73	8699525.97

Number of obs = 74  
 F( 2, 71) = 14.07  
 Prob > F = 0.0000  
 R-squared = 0.2838  
 Adj R-squared = 0.2637  
 Root MSE = 2530.9

**R squared**

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
price	-294.1955	55.69172	-5.28	0.000	-405.2417 -183.1494
mpg	1767.292	700.158	2.52	0.014	371.2169 3163.368
foreign	11905.42	1158.634	10.28	0.000	9595.164 14215.67
_cons					

**t statistic** **P value**

If 95% of the *t* distribution is closer to the mean than the *t*-value on the coefficient you are looking at, then you have a *P* value of 5%. This is also referred to a significance level of 5%. The *P* value is the probability of seeing a result as extreme as the one you are getting (a *t* value as large as yours) in a collection of random data in which the variable had no effect. A *P* of 5% or less is the generally accepted point at which to reject the null hypothesis. With a *P* value of 5% (or .05) there is only a 5% chance that results you are seeing would have come up in a random distribution, so you can say with a 95% probability of being correct that the variable is having some effect, assuming your model is specified correctly.

The *95% confidence interval* for your coefficients shown by many regression packages gives you the same information. You can be 95% confident that the real, underlying value of the coefficient that you are estimating falls somewhere in that 95% confidence interval, so if the interval does not contain 0, your *P* value will be .05 or less.

Note that the size of the *P* value for a coefficient says nothing about the size of the effect that variable is having on your dependent variable - it is possible to have a highly significant result (very small *P*-value) for a miniscule effect.

## Coefficients

In simple or multiple linear regression, the size of the coefficient for each independent variable gives you the size of the effect that variable is having on your dependent variable, and the sign on the

coefficient (positive or negative) gives you the direction of the effect. In regression with a single independent variable, the coefficient tells you how much the dependent variable is expected to increase (if the coefficient is positive) or decrease (if the coefficient is negative) when that independent variable increases by one. In regression with multiple independent variables, the coefficient tells you how much the dependent variable is expected to increase when that independent variable increases by one, *holding all the other independent variables constant*. Remember to keep in mind the units which your variables are measured in.

**Note:** in forms of regression other than linear regression, such as logistic or probit, the coefficients do not have this straightforward interpretation. Explaining how to deal with these is beyond the scope of an introductory guide.

## R-Squared and overall significance of the regression

The R-squared of the regression is the fraction of the variation in your dependent variable that is accounted for (or predicted by) your independent variables. (In regression with a single independent variable, it is the same as the square of the correlation between your dependent and independent variable.) The R-squared is generally of secondary importance, unless your main concern is using the regression equation to make accurate predictions. The P value tells you how confident you can be that each individual variable has some correlation with the dependent variable, which is the important thing.

Another number to be aware of is the P value for the regression as a whole. Because your independent variables may be correlated, a condition known as multicollinearity, the coefficients on individual variables may be insignificant when the regression as a whole is significant. Intuitively, this is because highly correlated independent variables are explaining the same part of the variation in the dependent variable, so their explanatory power and the significance of their coefficients is "divided up" between them.

## Further Reading

- [Linear Regression 101](#)
- [Stats topics](#)
- [Resources at the UCLA Statistical Computing Portal](#)

© 2007 The Trustees of Princeton University. All rights reserved. [data@princeton.edu](mailto:data@princeton.edu);

NOTE: Information is for Princeton University. Feel free to use the documentation but we can not answer questions outside of Princeton

This page last updated on: