What is the annotated output for a regression analysis in Stata?

Authored by

stats writer

June 29, 2024

The annotated output for a regression analysis in Stata is a comprehensive summary of the results obtained from running a regression model in the Stata statistical software. It includes a detailed display of the regression equation, the estimated coefficients, their standard errors, and significance levels, as well as various diagnostic statistics such as R-squared, F-statistic, and p-values. Additionally, the annotated output provides a thorough explanation of the data used, the methodology applied, and the assumptions made in the regression analysis. This annotated output serves as a valuable tool for researchers and analysts to interpret and communicate the findings of their regression analysis accurately.

Regression Analysis | Stata Annotated Output

This page shows an example regression analysis with footnotes explaining the
output. These data were collected on 200 high schools students and are
scores on various tests, including science, math, reading and social studies (socst).
The variable female is a dichotomous variable coded 1 if the student was
female and 0 if male.

use https://stats.idre.ucla.edu/stat/stata/notes/hsb2
(highschool and beyond (200 cases))

regress science math female socst read

Source | SS df MS Number of obs = 200
----------+---------- F(4, 195) = 46.69

Model | 9543.72074 4 2385.93019 Prob > F = 0.0000
Residual | 9963.77926 195 51.0963039 R-squared = 0.4892
----------+---------- Adj R-squared = 0.4788
Total | 19507.5 199 98.0276382 Root MSE = 7.1482

--
science | Coef. Std. Err. t P>|t|
----------+---
math | .3893102 .0741243 5.25 0.000 .243122 .5354983
female | -2.009765 1.022717 -1.97 0.051 -4.026772 .0072428
socst | .0498443 .062232 0.80 0.424 -.0728899 .1725784
read | .3352998 .0727788 4.61 0.000 .1917651 .4788345
_cons | 12.32529 3.193557 3.86 0.000 6.026943 18.62364
--

Anova Table

Sourcea | SSb dfc MSd
----------+----------
Model | 9543.72074 4 2385.93019
Residual | 9963.77926 195 51.0963039
----------+----------

Total | 19507.5 199 98.0276382

a. Source - This is the source of variance, Model, Residual, and Total. The Total variance is partitioned into the variance which can be explained by the independent variables (Model) and the variance which is not explained by the independent variables (Residual, sometimes called Error). Note that the Sums of Squares for the Model and Residual add up to the Total Variance, reflecting the fact that the Total Variance is partitioned into Model and Residual variance.

b. SS - These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual. These can be computed in many ways.
Conceptually, these formulas can be expressed as:
SSTotal The total variability around the mean. S(Y - Ybar)2.
SSResidual The sum of squared errors in prediction. S(Y - Ypredicted)2.
SSModel The improvement in prediction by using

the predicted value of Y over just using the mean of Y. Hence, this would
be the squared differences between the predicted value of Y and the mean of Y,
S(Ypredicted - Ybar)2. Another
way to think of this is the SSModel is SSTotal - SSResidual. Note that the
SSTotal = SSModel + SSResidual. Note that SSModel / SSTotal is equal to .4892, the value of R-Square. This is because R-Square is the
proportion of the variance explained by the independent variables, hence can be computed
by SSModel / SSTotal.

c. df - These are the
degrees of freedom associated with the sources of variance. The total
variance has N-1 degrees of freedom. In this case, there were N=200
students, so the DF
for total is 199. The model degrees of freedom corresponds to the number
of predictors minus 1 (K-1). You may think this would be 4-1 (since there were

4
independent variables in the model, math, female, socst and read).
But, the intercept is automatically included in the model (unless you explicitly omit the
intercept). Including the intercept, there are 5 predictors, so the model has
5-1=4
degrees of freedom. The Residual degrees of freedom is the DF total minus the DF
model, 199 - 4 is 195.

d. MS - These are the Mean
Squares, the Sum of Squares divided by their respective DF. For the Model,

9543.72074 / 4 = 2385.93019. For the Residual, 9963.77926 / 195 =

51.0963039. These are
computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square
Residual to test the significance of the predictors in the model.

Overall Model Fit

Number of obse = 200
F(4, 195)f = 46.69
Prob > Ff = 0.0000
R-squaredg = 0.4892
Adj R-squaredh = 0.4788
Root MSEi = 7.1482

e. Number of obs - This is the number of observations used in the regression analysis.

f. F and Prob > F - The F-value is the Mean Square Model (2385.93019) divided by the Mean Square Residual (51.0963039), yielding F=46.69. The p-value associated with this F value is very small (0.0000).
These values are used to answer the question "Do the independent variables reliably predict the dependent variable?. The p-value is compared to your alpha level (typically 0.05) and, if smaller, you can conclude "Yes, the independent variables reliably predict the dependent

**variable". You could say
that the group of variables math and female can be used to
reliably predict science (the dependent variable). If the p-value were greater than
0.05, you would say that the group of independent variables does not show a
statistically significant relationship with the dependent variable, or that the group of
independent variables does not reliably predict the dependent variable. Note that
this is an overall significance test assessing whether the group of independent
variables when used together reliably predict the dependent variable, and does
not address the ability of any of the particular independent variables to
predict the dependent variable. The ability of each individual independent
variable to predict the dependent variable is addressed in the table below where
each of the individual variables are listed.**

g. R-squared - R-Squared is the proportion

of variance in the dependent variable (science) which can be predicted from the independent variables (math,female, socst and read). This value indicates that 48.92% of the variance in science scores can be predicted from the variables math,female, socst and read. Note that this is an overall measure of the strength of association, and does not reflect the extent to which any particular independent variable is associated with the dependent variable.

h. Adj R-squared - Adjusted R-square. As predictors are added to the model, each predictor will explain some of the variance in the dependent variable simply due to chance. One could continue to add predictors to the model which would continue to improve the ability of the predictors to explain the dependent variable, although some of this increase in R-square would be simply due to chance variation in that particular sample. The

adjusted R-square attempts to yield a more honest value to estimate the
R-squared for the population. The value of R-square was .4892, while the value
of Adjusted R-square was .4788 Adjusted R-squared is computed using the formula
1 - ((1 - Rsq)((N - 1) /(N - k - 1)). From this formula, you can see that when the
number of observations is small and the number of predictors is large, there
will be a much greater difference between R-square and adjusted R-square
(because the ratio of (N - 1) / (N - k - 1) will be much greater than 1). By contrast,
when the number of observations is very large compared to the number of
predictors, the value of R-square and adjusted R-square will be much closer
because the ratio of (N - 1)/(N - k - 1) will approach 1.

i. Root MSE - Root MSE is the standard
deviation of the error term, and is the square root of the Mean Square Residual
(or Error).

Parameter Estimates

---

sciencej | Coef.k Std. Err.l tm P>|t|m n

---+---

math | .3893102 .0741243 5.25 0.000 .243122 .5354983

female | -2.009765 1.022717 -1.97 0.051 -4.026772 .0072428

socst | .0498443 .062232 0.80 0.424 -.0728899 .1725784

read | .3352998 .0727788 4.61 0.000 .1917651 .4788345

_cons | 12.32529 3.193557 3.86 0.000 6.026943 18.62364

---

j. science - This column shows the dependent variable at the top (science) with the predictor variables below it (math,female, socst, read and _cons). The last variable (_cons) represents the constant, also referred to in textbooks as the Y intercept, the height of the regression line when it crosses the Y axis. In other words, this is the predicted value of science when all other variables are 0.

k. Coef. - These are the values for the regression equation for
predicting the dependent variable from the independent variable. The regression
equation is presented in many different ways, for example:

Ypredicted = b0 + b1*x1 + b2*x2 + b3*x3 + b4*x4

The column of estimates (coefficients or
parameter estimates, from here on labeled coefficients) provides the values for
b0, b1, b2, b3 and b4 for this equation. Expressed in terms of the variables used
in this example, the regression equation is

sciencePredicted = 12.32529 +
.3893102*math +
-2.009765*female+.0498443*socst+.3352998*read

These estimates tell you about the
relationship between the independent variables and the dependent variable.
These estimates tell the amount of increase in science scores that would be predicted

by a 1 unit increase in the predictor. Note: For the independent variables
which are not significant, the coefficients are not significantly different from
0, which should be taken into account when interpreting the coefficients. (See
the columns with the t-value and p-value about testing whether the coefficients
are significant).
math - The coefficient (parameter estimate) is

.3893102. So, for every unit (i.e., point, since this is the metric in
which the tests are measured)
increase in math, a .3893102 unit increase in science is predicted,
holding all other variables constant. (It does not matter at what value you hold
the other variables constant, because it is a linear model.) Or, for
every increase of one point on the math test, your science score is predicted to be
higher by .3893102 points. This is significantly different from 0.

female - For every unit increase in female, there is a

-2.009765 unit decrease in
the predicted science score, holding all other variables constant. Since female is coded 0/1 (0=male,
1=female) the interpretation can be put more simply. For females the predicted
science score would be 2 points lower than for males. The variable
female is technically not statistically significantly different from 0,
because the p-value is greater than .05. However, .051 is so close to .05
that some researchers would still consider it to be statistically significant.
socst - The coefficient for socst is .0498443.
This means that for a 1-unit increase in the social studies score, we expect an
approximately .05 point increase in the science score. This is not
statistically significant; in other words, .0498443 is not different from 0.
read - The coefficient for read is .3352998.
Hence, for every unit increase in reading score we

**expect a .34 point increase
in the science score. This is statistically significant.**

**l. Std. Err. - These are the standard
errors associated with the coefficients. The standard error is used for testing
whether the parameter is significantly different from 0 by dividing the
parameter estimate by the standard error to obtain a t-value (see the column
with t-values and p-values). The standard errors can also be used to form a
confidence interval for the parameter, as shown in the last two columns of this
table.**

**m. t and P>|t| - These columns provide the t-value and 2-tailed p-value used in testing the null hypothesis that the
coefficient (parameter) is 0. If you use a 2-tailed test, then you would compare each p-value to your pre-selected value of alpha. Coefficients having p-values less than alpha are statistically significant. For example, if you chose alpha to be 0.05,**

coefficients having a p-value of 0.05 or less would be statistically significant (i.e., you can reject the null hypothesis and say that the coefficient is significantly different from 0). If you use a 1-tailed test (i.e., you hypothesize that the parameter will go in a particular direction), then you can divide the p-value by 2 before comparing it to your pre-selected alpha level.

The coefficient for female (-2.009765) is technically not significantly different from 0 because with a 2-tailed test and alpha of 0.05, the p-value of 0.051 is greater than 0.05. However, if you used a 1-tailed test, the p-value is now (0.051/2=.0255), which is less than 0.05 and then you could conclude that this coefficient is less than 0. CAUTION: We do not recommend changing from a two-tailed test to a one-tailed test *after* running your regression. This would be statistical cheating! You must know the direction of your hypothesis *before* running your regression.

The coefficient for math (3893102) is significantly different from 0 using alpha of 0.05 because its p-value is 0.000, which is smaller than 0.05.

The coefficient for socst (.0498443) is not statistically

significantly different from 0 because its p-value is definitely larger than 0.05.

The coefficient for read (.3352998) is statistically significant because its p-value of 0.000 is less than .05.

The constant (_cons) is significantly different from 0 at the 0.05 alpha level. However, having a significant intercept is seldom interesting.

n. - This shows a 95%
confidence interval for the coefficient. This is very useful as it helps you
understand how high and how low the actual population value of the parameter
might be. The confidence intervals are related to the p-values such that
the coefficient will not be statistically significant if the confidence interval
includes 0. If you look at the confidence interval for female, you will
see that it just includes 0 (-4 to .007). Because .007 is so close to 0,
the p-value is close to .05. If the upper confidence level had been a

**little smaller, such that it did not include 0, the coefficient for female
would have been statistically significant. Also, consider the coefficients for
female (-2) and read (.34). Immediately you see that the estimate for
female is so much bigger, but examine
the confidence interval for it (-4 to .007). Now examine the confidence
interval for read (.19 to .48). Even though female has a bigger coefficient
(in absolute terms)
it could be as small as -4. By contrast, the lower confidence level for read is
.19, which is still above 0. So, even though female has a bigger
coefficient, read is significant and even the smallest value in the
confidence interval is still higher than 0. The same cannot be said about the
coefficient for socst. Such confidence intervals help you to put the estimate
from the coefficient into perspective by seeing how much the value could vary.**