What does the annotated output of a SAS regression analysis reveal?

Authored by

stats writer

June 29, 2024

The annotated output of a SAS regression analysis is a comprehensive report that reveals the statistical results and interpretations of a regression model. It provides detailed information on the variables used in the model, their coefficients, and their significance levels. The output also includes measures of the model's goodness of fit, such as R-squared and adjusted R-squared, as well as diagnostic tests for assumptions and outliers. Additionally, the annotated output may contain visual aids such as graphs and charts to aid in the interpretation of the results. Overall, the annotated output of a SAS regression analysis serves as a valuable tool for understanding the relationship between variables and making informed decisions based on the statistical findings.

Regression Analysis | SAS Annotated Output

This page shows an example regression analysis with footnotes explaining the output. These data (hsb2demo) 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.

In the code below, the data = option on the proc reg statement tells SAS where to find the SAS data set to be used in the analysis. On the model statement, we specify the regression model that we want to run, with the dependent variable (in this case, science) on

the left of the
equals sign, and the independent variables on the right-hand side. We use
the clb option after the slash on the model statement to get the
95% confidence limits of the parameter estimates. The quit
statement is included because proc reg is an interactive procedure, and
quit tells SAS that not to expect another proc reg immediately.

```
%let path=C:temp;
libname idre "&path";

proc reg data = idre.hsb2demo;
model science = math female socst read / clb;
run;
quit;
```

The REG Procedure
Model: MODEL1
Dependent Variable: science science score

Analysis of Variance

Sum of Mean
Source DF Squares Square F Value Pr > F

Model 4 9543.72074 2385.93019 46.69 <.0001
Error 195 9963.77926 51.09630
Corrected Total 199 19507

Root MSE 7.14817 R-Square 0.4892
Dependent Mean 51.85000 Adj R-Sq 0.4788
Coeff Var 13.78624

Parameter Estimates

Parameter Standard
Variable Label DF Estimate Error t Value Pr > |t|

Intercept Intercept 1 12.32529 3.19356 3.86 0.0002
math math score 1 0.38931 0.07412 5.25 <.0001
female 1 -2.00976 1.02272 -1.97 0.0508
socst social studies score 1 0.04984 0.06223 0.80 0.4241
read reading score 1 0.33530 0.07278 4.61 <.0001

Parameter Estimates

Variable Label DF 95% Confidence Limits

Intercept Intercept 1 6.02694 18.62364
math math score 1 0.24312 0.53550
female 1 -4.02677 0.00724
socst social studies score 1 -0.07289 0.17258
read reading score 1 0.19177 0.47883

Anova Table

Analysis of Variance
Sum of Mean
Sourcea DFb Squaresc Squared F Valuee Pr > Fe
Model 4 9543.72074 2385.93019 46.69 <.0001
Error 195 9963.77926 51.09630
Corrected Total 199 19507

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. 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.

c. Sum of Squares - 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.

d. Mean Square - 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.

e. F Value and Pr > 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, female, socst and read 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.

Overall Model Fit

Root MSEf 7.14817 R-Squarei 0.4892
Dependent Meang 51.85000 Adj R-Sqj 0.4788
Coeff Varh 13.78624

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

g. Dependent Mean - This is the mean of the dependent variable.

h. Coeff Var - This is the coefficient of variation, which is a unit-less
measure of variation in the data. It is the root MSE divided by the mean of the dependent variable, multiplied by 100: (100*(7.15/51.85) = 13.79).

i. R-Square - R-Square 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.

j. Adj R-Sq - 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 less 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.**

Parameter Estimates

Parameter Estimates

Parameter Standard

Variablek Labell DFm Estimaten Erroro t Valuep Pr > |t|p

Intercept Intercept 1 12.32529 3.19356 3.86 0.0002
math math score 1 0.38931 0.07412 5.25 <.0001
female 1 -2.00976 1.02272 -1.97 0.0508
socst social studies score 1 0.04984 0.06223 0.80 0.4241
read reading score 1 0.33530 0.07278 4.61 <.0001

Parameter Estimates

Variablek Labell DFm 95% Confidence Limitsq

Intercept Intercept 1 6.02694 18.62364
math math score 1 0.24312 0.53550
female 1 -4.02677 0.00724
socst social studies score 1 -0.07289 0.17258
read reading score 1 0.19177 0.47883

k. Variable - This column shows the predictor variables (constant, math,female, socst, read).
The first variable (constant) 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.

l. Label - This column gives the label for the variable. Usually, variable labels are added when the data set is created so that it is clear what the variable is (as the name of the variable can sometimes be ambiguous). SAS has labeled the variable Intercept for us by default. Note that this variable is not added to the data set.

m. DF - This column give the degrees of freedom associated with each independent variable. All continuous variables have one degree of freedom, as do binary variables (such as female).

n. Parameter Estimates - 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.

o. Standard Error - 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.

p. t Value and Pr > |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 preselected 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 predict
that the parameter will go in a particular direction), then you can divide the p-value by 2 before comparing it to your preselected alpha level. With a 2-tailed
test and alpha of 0.05, you may reject the null hypothesis that the coefficient
for female is equal to 0. The coefficient of -2.009765 is significantly greater
than 0. However, if you used a 2-tailed test and alpha of 0.01, the p-value
of .0255
is greater than 0.01 and the coefficient for female would not be significant at
the 0.01 level. Had you predicted that this coefficient would be positive (i.e.,
a one tail test), you would be able to divide the p-value by 2 before comparing
it to alpha. This would yield a one-tailed p-value of 0.00945, which is less
than 0.01 and then you could conclude that this coefficient is greater than 0
with a one tailed alpha of 0.01.

The coefficient for math 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 (0.0498443) is not statistically significantly different from 0 because
its p-value is definitely larger than 0.05.
The coefficient for read (0.3353) 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.

q. 95% Confidence Limits - 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.