This page shows an example simple regression analysis with footnotes explaining the output. The analysis uses a data file about scores obtained by elementary schools, predicting **api00** from **enroll** using the following Stata commands.

```
use https://stats.idre.ucla.edu/stat/stata/webbooks/reg/elemapi
regress api00 enroll
```

The output of this command is shown below, followed by explanations of the output.

Output

```
      Source[a] |       SS[b]       df[c]       MS[d]            Number of obs[e] =     400
-------------+------------------------------           F(  1,   398)[f] =   44.83
       Model |  817326.293     1  817326.293           Prob > F[f]      =  0.0000
    Residual |  7256345.70   398  18232.0244           R-squared[g]     =  0.1012
-------------+------------------------------           Adj R-squared[h] =  0.0990
       Total |  8073672.00   399  20234.7669           Root MSE[i]      =  135.03


------------------------------------------------------------------------------
      api00[j] |      Coef.[k]   Std. Err.[l]      t[m]    P>|t|[m]     [95% Conf. Interval][n]
-------------+----------------------------------------------------------------
      enroll |  -.1998674   .0298512    -6.70   0.000    -.2585532   -.1411817
       _cons |   744.2514   15.93308    46.71   0.000     712.9279    775.5749
------------------------------------------------------------------------------
```

Footnotes

a. 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. 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. These are the Sum of Squares associated with the three sources of variance, Total, Model & Residual. These can be computed in many ways. Conceptually, these formulas can be expressed as: SSTotal. The total variability around the mean. $\Sigma(Y – Ybar)^2$. SSResidual. The sum of squared errors in prediction. $\Sigma(Y – Ypredicted)^2$. 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, Σ(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 .10, 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. 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=400 observations, so the DF for total is 399. The model degrees of freedom corresponds to the number of predictors minus 1 (K-1). You may think this would be 1-1 (since there was 1 independent variable in the model statement, enroll). But, the intercept is automatically included in the model (unless you explicitly omit the intercept). Including the intercept, there are 2 predictors, so the model has 2-1=1 degree of freedom. The Residual degrees of freedom is the DF total minus the DF model, 399 – 1 is 398.

d. These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, 817326.293 / 1 is equal to 817326.293. For the Residual, 7256345.7 / 398 equals 18232.0244. 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 predictor(s) in the model.

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

f. The F Value is the Mean Square Model (817326.293) divided by the Mean Square Residual (18232.0244), yielding F=44.83. 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 variable enroll can be used to reliably predict api00 (the dependent variable). If the p value were greater than 0.05, you would say that the independent variable does not show a significant relationship with the dependent variable, or that the independent variable does not reliably predict the dependent variable.

g. R-Square is the proportion of variance in the dependent variable (api00) which can be predicted from the independent variable (enroll). This value indicates that 10% of the variance in api00 can be predicted from the variable enroll.

h. 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 .10, while the value of Adjusted R-square was .099. 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 (N-1)/(N-k-1) will be much greater than 1 and adjusted R-squared will be much smaller than unadjusted R-squared. 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 (N-1)/(N-k-1) will approach 1.

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

j. This column shows the dependent variable at the top (api00) with the predictor variables below it (enroll). 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.

k. 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**

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

api00Predicted = **744.25 – .20*enroll**

This estimate tells you about the relationship between the independent variable and the dependent variable. This estimate indicates the amount of increase in api00 that would be predicted by a 1 unit increase in the predictor.   Note: If an independent variable is not significant, the coefficient is not significantly different from 0, which should be taken into account when interpreting the coefficient. (See the columns with the t value and p value about testing whether the coefficients are significant). enroll – The coefficient (parameter estimate) is -.20. So, for every unit increase in enroll, a -.20 unit decrease in api00 is predicted.

l. 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 2 columns of this table.

m. 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 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 can reject the null hypothesis that the coefficient for enroll is equal to 0. The coefficient of -.20 is significantly different from 0. 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. Such confidence intervals help you to put the estimate from the coefficient into perspective by seeing how much the value could vary.

Click here to report an error on this page or leave a comment

How to cite this page (https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/)