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|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        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

```
      Source[a] |       SS[b]       df[c]       MS[d]
-------------+------------------------------
       Model |  9543.72074     4  2385.93019
    Residual |  9963.77926   195  51.0963039
```

```
-------------+------------------------------
       Total |     19507.5   199  98.0276382
```

a. **Source** – Looking at the breakdown of variance in the outcome variable, these are the categories we will examine: 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).

b. **SS** – These are the Sum of Squares associated with the three sources of variance, Total, Model and Residual.

c. **df** – These are the degrees of freedom associated with the sources of variance. The total variance has N-1 degrees of freedom. The model degrees of freedom corresponds to the number of coefficients estimated minus 1. Including the intercept, there are 5 coefficients, 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 =195.

d. **MS** – These are the Mean Squares, the Sum of Squares divided by their respective DF.

Overall Model Fit

```
Number of obs[e] =     200
F(  4,   195)[f] =   46.69
Prob > F[g]      =  0.0000
R-squared[h]     =  0.4892
Adj R-squared[i] =  0.4788
Root MSE[j]      =  7.1482
```

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

f. **F(4, 195)** – This is the F-statistic is the Mean Square Model (2385.93019) divided by the Mean Square Residual (51.0963039), yielding F=46.69. The numbers in parentheses are the Model and Residual degrees of freedom are from the ANOVA table above.

g. **Prob > F** – This is the p-value associated with the above F-statistic. It is used in testing the null hypothesis that all of the model coefficients are 0.

h. **R-squared** – R-Squared is the proportion of variance in the dependent variable (**science**) which can be explained by the independent variables (**math, female**, **socst** and **read**). 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.

i. **Adj R-squared** – This is an adjustment of the R-squared that penalizes the addition of extraneous predictors to the model. Adjusted R-squared is computed using the formula 1 – ((1 – Rsq)((N – 1) /(N – k – 1)) where k is the number of predictors.

j. **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

science[k]	Coef.[l]	Std. Err.[m]	t[n]	P>|t|[o]	[95% Conf. Interval][p]
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

k. **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 or intercept.

l. **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 provides the values for b0, b1, b2, b3 and b4 for this equation.

math – The coefficient is

.3893102. So for every unit increase in **math**, a .3893102 unit increase in **science** is predicted, holding all other variables constant.

female – For every unit increase in **female**, we expect a 2.009765 unit decrease in the **science** score, holding all other variables constant. Since **female** is coded 0/1 (0=male, 1=female) the interpretation is

more simply: for females, the predicted science score would be 2 points lower than for males.

socst – The coefficient for **socst** is .0498443. So for every unit increase in **socst**, we expect an approximately .05 point increase in the science score, holding all other variables constant.

read – The coefficient for **read** is .3352998. So for every unit increase in **read**, we expect a .34 point increase in the science score.

m. **Std. Err.** – These are the standard errors associated with the coefficients.

n. **t** – These are the t-statistics used in testing whether a given coefficient is significantly different from zero.

o. **P>|t|** – This column shows the 2-tailed p-values used in testing the null hypothesis that the coefficient (parameter) is 0. Using an alpha of 0.05:

The coefficient for **math** is significantly different from 0 because its p-value is 0.000, which is smaller than 0.05.

The coefficient for **female** (-2.01) is not statictically significant at the 0.05 level since the p-value is greater than .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.

p. [**95% Conf. Interval**] – These are the 95% confidence intervals for the coefficients. The confidence intervals are related to the p-values such that the coefficient will not be statistically significant at alpha = .05 if the 95% confidence interval includes zero. These confidence intervals can 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/)

© 2019 UC REGENTS (http://www.ucla.edu/terms-of-use/)