ANNOTATED STATA OUTPUT MULTIPLE REGRESSION ANALYSIS

This page shows an example multiple regression analysis with footnotes explaining the output. The analysis uses a data file about scores obtained by elementary schools, predicting **api00** from **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll** using the following Stata commands.

```
use https://stats.idre.ucla.edu/stat/stata/olc/reg/elemapi2
regress api00 ell meals yr_rnd mobility acs_k3 acs_46 full emer 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]   =    395
----------------+------------------------------   F(9, 385)[f]       =    232.41
          Model |  6740702.01          9   748966.89   Prob > F          =    0.0000
       Residual |  1240707.78        385  3222.61761   R-squared[g]       =    0.8446
```

Total	7981409.79	394	20257.3852

Adj R-squared[h] = 0.8409
Root MSE[i] = 56.768

api00[j]	Coef.[k]	Std. Err.[l]	t[m]	P>|t|[m]	[95% Conf. Interval][n]	
ell	-.8600707	.2106317	-4.08	0.000	-1.274203	-.4459382
meals	-2.948216	.1703452	-17.31	0.000	-3.28314	-2.613293
yr_rnd	-19.88875	9.258442	-2.15	0.032	-38.09219	-1.685309
mobility	-1.301352	.4362053	-2.98	0.003	-2.158995	-.4437088
acs_k3	1.3187	2.252683	0.59	0.559	-3.110401	5.747801
acs_46	2.032456	.7983213	2.55	0.011	.462841	3.602071
full	.609715	.4758205	1.28	0.201	-.3258169	1.545247
emer	-.7066192	.6054086	-1.17	0.244	-1.89694	.4837019
enroll	-.012164	.0167921	-0.72	0.469	-.0451798	.0208517
_cons	758.9418	62.28601	12.18	0.000	636.4785	881.4051

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 (Residual). 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 and 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, $S(Ypredicted - Ybar)^2$. 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 .84, 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 (DF). In this case, there were N=395 observations, so the DF for total is 394. The model degrees of freedom corresponds to the number of predictors minus 1 (K-1). You may think this would be 9-1 (since there were 9 independent variables in the model: **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll**). But, the intercept is automatically included in the model (unless you explicitly omit the intercept). Including the intercept, there are 10 predictors, so the model has 10-1=9 degrees of freedom. The Residual degrees of freedom is the DF total minus the DF model, 394 – 9 is 385.

d. These are the Mean Squares, the Sum of Squares divided by their respective DF. For the Model, 6740702.01 / 9 is equal to 748966.89. For the Residual, 124070779 / 385 equals 3222.6176. These are computed so you can compute the F ratio, dividing the Mean Square Model by the Mean Square Residual (or Error) to test the significance of the predictors in the model.

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

f. The F Value is the Mean Square Model (748966.89) divided by the Mean Square Residual (3222.61761), yielding F=232.41. 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 **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full** , and **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 group of independent variables do not show a significant relationship with the dependent variable, or that the group of independent variables do 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 variables. 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-Square is the proportion of variance in the dependent variable (**api00**) which can be predicted from the independent variables (**ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full emer**, and **enroll**). This value indicates that 84% of the variance in **api00** can be predicted from the variables **ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full**, **emer** and **enroll**. 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. Adjusted R-square. As predictors are added to the model, each predictor will explain some of the

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 .8446, while the value of Adjusted R-square was .8409. Adjusted R-squared is computed using the formula 1 – ((1-R-sq)(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.

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 (**ell**, **meals**, **yr_rnd**, **mobility**, **acs_k3**, **acs_46**, **full emer** and **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** +**b2*x2** + **b3*x3 . . .**

The column of estimates (coefficients or parameter estimates, from here on labeled coefficients) provides the values for b0, b1, b2, b3, b4, b5, b6, b7, b8 and b9 for this equation. Expressed in terms of the variables used in this example, the regression equation is

api00Predicted =778.83-.86*ell-2.95*meals-19.89*yr_rnd-1.30*mobility+1.32*acs_k3+2.03*acs_46+.61*full-.71*emer-.01*enroll

These estimates tell you about the relationship between the independent variables and the dependent variable. These estimates tell the amount of increase in api00 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.) **ell** – The coefficient (parameter estimate) is -.86. So, for every unit increase in ell, a .86 unit decrease in **api00** is predicted. Or, for every increase of one percentage point of **api00**, **ell** is predicted to be lower by .86. This is significantly different from 0. **meals** – For every unit increase in **meals**, there is a 2.95 unit decrease in the predicted **api00**.

yr_rnd – For every unit increase of **yr_rnd**, the predicted value of **api00** would be 19.89 units lower. **mobility** – For every unit increase in **mobility**, **api00** is predicted to be 1.30 units lower. **acs_k3** – For every unit increase in **acs_k3**, **api00** is predicted to be 1.32 units higher. **acs_46** – For every unit increase in **acs_46**, **api00** is predicted to be 2.03 units higher. **full** – For every unit increase in **full**, **api00** is predicted to be .61 units higher. **emer** – For every unit increase in **emer**, **api00** is predicted to be .71 units lower. **enroll** – For every unit increase in **enroll**, **api00** is predicted to be .01 units lower.

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 **ell** is equal to 0. The coefficient of -.86 is significantly different from 0. Using a 2-tailed test and alpha of 0.01, the p-value of 0.000 is smaller than 0.01 and the coefficient for **ell** would still be significant at the 0.01 level. Had you predicted that this coefficient would be positive (i.e., a 1-tailed test), you would be able to divide the p-value by 2 before comparing it to alpha. This would yield a 1-tailed p-value of 0.000, which is less than 0.01, and then you could conclude that this coefficient is greater than 0 with a 1-tailed alpha of 0.01. The coefficient for **meals** is significantly different from 0 using alpha of 0.05 because its p-value of 0.000 is smaller than 0.05. The coefficient for **yr_rnd** (-19.89) is significantly different from 0 because its p-value is definitely smaller than 0.05 and even 0.01. The coefficient for **mobility** is significantly different from 0 using alpha of 0.05 because its p-value of 0.003 is smaller than 0.05. The coefficient for **acs_k3** is not significantly different from 0 using alpha of 0.05 because its p-value of .559 is greater than 0.05. The coefficient for **acs_46** is significantly different from 0 using alpha of 0.05 because its p-value of 0.011 is smaller than 0.05. The coefficient for **full** is not significantly different from 0 using alpha of 0.05 because

its p-value of .201 is greater than 0.05. The coefficient for **emer** is not significantly different from 0 using alpha of 0.05 because its p-value of .244 is greater than 0.05. The coefficient for **enroll** is not significantly different from 0 using alpha of 0.05 because its p-value of .469 is greater than 0.05. 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. Consider the coefficients for **ell** (-.86) and **meals** (-2.95). Immediately you see that the estimate for **meals** is so much bigger, but examine the confidence interval for it (-3.28 to -2.61). Now examine the confidence interval for **ell** (-1.27 to -.45). Even though **meals** has a larger coefficient, it could be as small as -3.28. By contrast, the lower confidence level for **ell** is -1.27.

