Simple Linear Regression

Simple Linear Regression tells you the amount of variance accounted for by one variable in predicting another variable.

```
. reg sexfreq age marital racenew happy attend

      Source |       SSc      df       MS              Number of obs =    1052
-------------+------------------------------           F(  5,  1046) =  101.49d
       Modela|  1382.66217     5  276.532435           Prob > F      =  0.0000
    Residualb|  2850.06406  1046  2.72472664           R-squared     =  0.3267e
-------------+------------------------------           Adj R-squared =  0.3234
       Total |  4232.72624  1051  4.02733229           Root MSE      =  1.6507f

------------------------------------------------------------------------------
     sexfreq |      Coef.   Std. Err.      tg    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0554982   .0030401   -18.26   0.000    -.0614636   -.0495327
     marital |  -1.372326   .1071273   -12.81   0.000    -1.582535   -1.162117
     racenew |  -.2153774   .1252702    -1.72   0.086    -.4611869    .0304321
       happy |  -.2618401    .084642    -3.09   0.002    -.4279275   -.0957527
      attend |  -.0674279   .0196243    -3.44   0.001    -.1059354   -.0289204
       _cons |      8.298   .2966264    27.97   0.000     7.715949     8.88005
------------------------------------------------------------------------------
```

a. The output for Model displays information about the variation accounted for by the model.

b. The output for Residual displays information about the variation that is not accounted for by your model. And the output for Total is the sum of the information for Regression and Residual.

c. A model with a large model sum of squares in comparison to the residual sum of squares indicates that the model accounts for most of variation in the dependent variable. Very high residual sum of squares indicate that the model fails to explain a lot of the variation in the dependent variable, and you may want to look for additional factors that help account for a higher proportion of the variation in the dependent variable. In this example, we see that 32.5% of the total sum of squares is made up from the regression sum of squares. You may notice that the R^2 for this model is also .325 (this is not a coincidence!).

d. If the significance value of the F statistic is small (smaller than say 0.05) then the independent variables do a good job explaining the variation in the dependent variable. If the significance value of F is larger than say 0.05 then the independent variables do not explain the variation in the dependent variable. For this example, the model does a good job explaining the variation in the dependent variable.

e. The "R Square" tells us how much of the variance of the dependent variable can be explained by the independent variable(s). In this example, 32.7% of the variance in frequency of sex is explained by differences in all included variables.

f. The standard error of the estimate (aka, the root mean square error), is the standard deviation of the error term, and is the square root of the Mean Square Residual (or Error).

g. The t statistics can help you determine the relative importance of each variable in the model. The t-statistic is calculated by divided the variable's unstandardized coefficient by its standard error. As a guide regarding useful predictors, look for t values well below -1.96 or above +1.96. As you can see, the race variable is the only t-value not within this range (note also that this is the only variable that is not significant).

Regression Equation

$$SEXFREQ_{predicted} = 8.298 - .055*age - 1.372*marital - .215*race - .262*happy - .067*attend$$

If we plug in values in for the independent variables (age = 35 years; marital = currently married-1; race = white-1; happiness = very happy-1; church attendance = never attends-0), we can predict a value for frequency of sex:

$$SEXFREQ_{predicted} = 8.298 - .055*35 - 1.372*1 - .215*1 - .262*1 - .067*0$$
$$= 4.524$$

As this variable is coded, a 35-year old, White, married person with high levels of happiness and who never attends church would be expected to report their frequency of sex between values 4 (weekly) and 5 (2-3 times per week).

If we plug in 70 years, instead, we find that frequency of sex is predicted at 2.599, or approximately 1-2 times per month.

Model Interpretation

Constant = The predicted value of "frequency of sex", when all other variables are 0. In this example, a value of 8.298 is not interpretable, since the valid responses for frequency of sex range from 0-6. Important to note, values of 0 for all variables is not interpretable either (i.e., age cannot equal 0).

Age = For every unit increase in age (in this case, year), frequency of sex will decrease by .055 units.

Marital Status = For every unit increase in marital status, frequency of sex will decrease by 1.372 units. Since marital status has only two categories, we can conclude that currently married persons have more sex than currently unmarried persons.

Race = For every unit increase in race, frequency of sex will decrease by .215 units. Since race has only two categories, we can conclude that non-White persons have more sex than White persons.

Happiness = For every unit increase in happiness, frequency of sex will decrease by .262 units. Recall that happiness is coded such that higher scores indicate less happiness. For this example, then, higher levels of happiness predict higher frequency of sex.

Church Attendance = For every unit increase in church attendance, frequency of sex decreases by .067 units.

Simple Linear Regression (with Age-Squared Variable)

It is known that some variables are often non-linear, or curvilinear. Such variables may be age or income. In this example, we include the original age variable and an age squared variable.

```
. reg sexfreq age marital racenew happy attend agesquar

      Source |       SS       df       MS              Number of obs =    1052
-------------+------------------------------           F(  6,  1045) =   85.55
       Model |  1394.24116     6  232.373526           Prob > F      =  0.0000
    Residual |  2838.48508  1045  2.71625366           R-squared     =  0.3294
-------------+------------------------------           Adj R-squared =  0.3255
       Total |  4232.72624  1051  4.02733229           Root MSE      =  1.6481

------------------------------------------------------------------------------
     sexfreq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0204952   .0172229    -1.19   0.234    -.0542906    .0133002
     marital |   -1.31425   .1105973   -11.88   0.000    -1.531268   -1.097232
     racenew |  -.2183136   .1250833    -1.75   0.081    -.4637567    .0271295
       happy |  -.2713968   .0846369    -3.21   0.001    -.4374745   -.1053191
      attend |  -.0670664   .0195946    -3.42   0.001    -.1055156   -.0286172
    agesquar |  -.0003523   .0001706    -2.06   0.039    -.0006872   -.0000175
       _cons |   7.466803   .4997854    14.94   0.000     6.486106    8.4475
------------------------------------------------------------------------------
```

The age squared variable is significant, indicating that age is non-linear.

Simple Linear Regression (with interaction term)

In a linear model, the effect of each independent variable is always the same. However, it could be that the effect of one variable depends on another. In this example, we might expect that the effect of age is dependent on sex. In the following example, we include an interaction term, age*sex.

To test for two-way interactions (often thought of as a relationship between an independent variable (IV) and dependent variable (DV), moderated by a third variable), first run a regression analysis, including both independent variables (IV and moderator) and their interaction (product) term. It is highly recommended that the independent variable and moderator are standardized before calculation of the product term, although this is not essential. The product term should be significant in the regression equation in order for the interaction to be interpretable.

For this example, two dummy variables were created, for ease of interpretation. Sex was recoded such that 1=Male and 0=Female. Marital status was recoded such that 1=Currently married and 0=Not currently married. The interaction term is a cross-product of these two dummy variables.

Regression Model (*without* interactions)

```
. reg sexfreq age racenew happy attend male married

      Source |       SS       df       MS              Number of obs =    1052
-------------+------------------------------           F(  6,  1045) =   87.90
       Model |  1419.69041     6  236.615068            Prob > F      =  0.0000
    Residual |  2813.03583  1045  2.69190031            R-squared     =  0.3354
-------------+------------------------------           Adj R-squared =  0.3316
       Total |  4232.72624  1051  4.02733229            Root MSE      =  1.6407

------------------------------------------------------------------------------
     sexfreq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0547746    .003028   -18.09   0.000    -.0607164   -.0488329
     racenew |  -.2307659   .1245824    -1.85   0.064    -.4752261    .0136942
       happy |  -.2421688   .0842976    -2.87   0.004    -.4075806    -.076757
      attend |  -.0562567   .0197369    -2.85   0.004    -.0949852   -.0175281
        male |   .3852511    .103874     3.71   0.000     .1814257    .5890764
     married |   1.317949   .1074847    12.26   0.000     1.107038    1.528859
       _cons |   5.295466   .2582801    20.50   0.000     4.788659    5.802273
------------------------------------------------------------------------------
```

Regression Model (*with* interactions)

```
. reg sexfreq age racenew happy attend male married interact

      Source |       SS       df       MS              Number of obs =    1052
-------------+------------------------------           F(  7,  1044) =   77.31
       Model |  1444.99234     7  206.427478            Prob > F      =  0.0000
    Residual |  2787.73389  1044  2.67024319            R-squared     =  0.3414
-------------+------------------------------           Adj R-squared =  0.3370
       Total |  4232.72624  1051  4.02733229            Root MSE      =  1.6341

------------------------------------------------------------------------------
     sexfreq |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0532626   .0030556   -17.43   0.000    -.0592584   -.0472668
     racenew |  -.2625718   .1245097    -2.11   0.035    -.5068896   -.0182541
       happy |  -.2508243   .0840049    -2.99   0.003    -.4156619   -.0859866
      attend |  -.0525717   .0196938    -2.67   0.008    -.0912157   -.0139278
        male |   .6725887   .1393423     4.83   0.000     .3991658    .9460117
     married |   1.630128    .147462    11.05   0.000     1.340772    1.919484
    interact |  -.6434146   .2090208    -3.08   0.002    -1.053563   -.2332659
       _cons |   5.135983   .2624047    19.57   0.000     4.621082    5.650883
------------------------------------------------------------------------------
```

Regression Equation

$SEXFREQ_{predicted}$ = 5.136 -.053*age - .263*race -.251*happy - .053*attend + 1.630*married + (.673 - .643*married) * male

Interpretation

Main Effects

The married coefficient represents the main effect for females (the 0 category). The effect for females is then 1.63, or the "marital" coefficient. The effect for males is 1.63 - .643, or .987.

The gender coefficient represents the main effect for unmarried persons (the 0 category). The effect for unmarried is then .673, or the "sex" coefficient. The effect for married is .673 - .643, or .03.

Interaction Effects

For a simple interpretation of the interaction term, plug values into the regression equation above.

Married Men =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**1* + (.673 - . 643**1*) * *1*	= 5.455
Married Women =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**1* + (.673 - . 643**1*) * *0*	= 5.425
Unmarried Men =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**0* + (.673 - . 643**0*) * *1*	= 3.825
Unmarried Women =	$SEXFREQ_{predicted}$ = 5.136 -.053*35 - .263*1 -.251*1 - .053*0 + 1.630**0* + (.673 - . 643**0*) * *0*	= 3.795

In this example (age = 35 years; race = white-1; happiness = very happy-1; church attendance = never attends-0), we can see that (1) for both married and unmarried persons, males are reporting higher frequency of sex than females, and (2) married persons report higher frequency of sex than unmarried persons. The interaction tells us that the gender difference is greater for married persons than for unmarried persons.