Regression: what to look for

Adding the rest of predictor variables:

`regress csat expense percent income high college, robust`

Robust standard errors (to control for heteroskedasticity)

Output variable (Y)

Predictor variables (X)

```
. regress csat expense percent income high college, robust

Linear regression                               Number of obs =        51
                                                F(  5,    45) =     50.90
                                                Prob > F      =    0.0000
                                                R-squared     =    0.8243
                                                Root MSE      =    29.571

------------------------------------------------------------------------------
             |               Robust
        csat |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     expense |   .0033528    .004781     0.70   0.487    -.0062766    .0129823
     percent |  -2.618177    .2288594   -11.44   0.000    -3.079123   -2.15723
      income |   .1055853    1.207246     0.09   0.931    -2.325933    2.537104
        high |   1.630841    .943318      1.73   0.091    -.2690989    3.530781
     college |   2.030894    2.113792     0.96   0.342    -2.226502    6.28829
       _cons |   851.5649    57.28743    14.86   0.000     736.1821    966.9477
------------------------------------------------------------------------------
```

1 This is the p-value of the model. It indicates the reliability of X to predict Y. Usually we need a p-value lower than 0.05 to show a statistically significant relationship between X and Y.

2 R-square shows the amount of variance of Y explained by X. In this case the model explains 82.43% of the variance in SAT scores.

3 Adj R^2 (not shown here) shows the same as R^2 but adjusted by the # of cases and # of variables. When the # of variables is small and the # of cases is very large then Adj R^2 is closer to R^2. This provides a more honest association between X and Y.

4 Two-tail p-values test the hypothesis that each coefficient is different from 0. To reject this, the p-value has to be lower than 0.05 (you could choose also an alpha of 0.10). In this case, `expense`, `income`, and `college` are not statistically significant in explaining SAT; `high` is almost significant at 0.10. `Percent` is the only variable that has some significant impact on SAT (its coefficient is different from 0)

5 The t-values test the hypothesis that the coefficient is different from 0. To reject this, you need a t-value greater than 1.96 (at 0.05 confidence). You can get the t-values by dividing the coefficient by its standard error. The t-values also show the importance of a variable in the model. In this case, `percent` is the most important.

6 csat = 851.56 + 0.003*expense – 2.62*percent + 0.11*income + 1.63*high + 2.03*college

7 Root MSE: root mean squared error, is the sd of the regression. The closer to zero better the fit.