Lab Worksheet IX

Multiple regression analysis

In today's worksheet we are going to do some multiple regression analysis on respondents' self-reported happiness using the *European Social Survey*. You can find the dataset **essuk16.dta** on LEARN. Since it is the same dataset we used in the previous worksheet, the same caveats established then also apply here (e.g., regression assumptions are not considered for simplicity's sake) . Remember three key Stata commands widely used to explore datasets:

```
describe
summarize
codebook, compact
```

We should be paying attention to things like the number of variables in the dataset, what they measure, how they are coded, whether we have missing cases and how these are coded. Last time we focused on the variables "happy", "trstprt", "trstplt", and "atchctr." This time let's try:

```
tabulate agea
tabulate region
tabulate gender
```

From the first tabulation we can see that respondents' ages range from 15 to 94. The command `summarize` would have given us this information more easily:

```
summarize agea
```

However, remember that tabulations are not always the best way to represent and explore data. Since "agea" is a continuous variable we could plot it in a histogram:

```
histogram agea, frequency
```

Also, you can explore two variables at the same time using the command `histogram`. For example, we could plot the distribution of respondents' ages in each region:

```
histogram agea, frequency by(region)
```

On the other hand, "region" is a categorical variable, so we could plot it with bars like this:

```
graph bar (percent), over(region, label(angle(45)))
```

The variable "gndr" is also categorical; however if you try plotting it in a bar chart or pie chart you might find it redundant because the variable only has two categories. Perhaps it is more convenient representing "gndr" with a simple tabulation (but again, this is a matter of personal preference). Stata allows you to explore data in a myriad of ways. For instance, consider the following:

```
table gndr, contents(freq mean agea) row
```

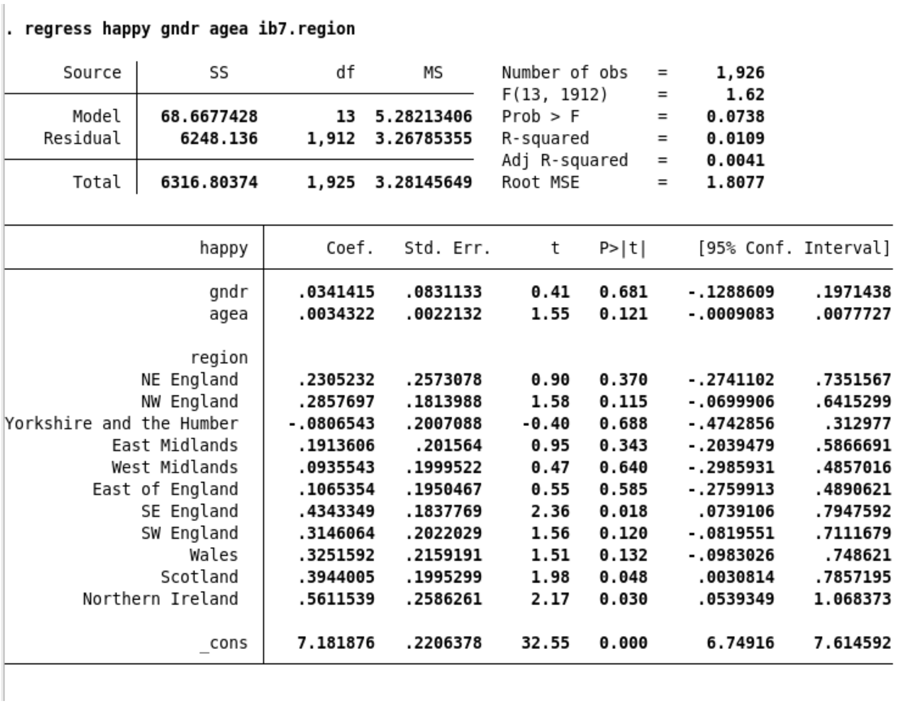
With the `table` command and the `contents` option we can explore gender and age at the same time. The tabulation above shows the mean (average) age for males and females. Tabulating means (not only frequencies or percentages) could sometimes be very useful. If you were to ask the data “are males happier than females?” you would probably want to start by looking at the average level of happiness:

```
table gndr, contents(freq mean happy) row
```

and you would find that female respondents are ever so slightly happier than male respondents on average (7.66 versus 7.6).

Now that we have explored a few variables let’s run some regressions. Start with:

```
regress happy gndr agea ib7.region
```



```
. regress happy gndr agea ib7.region

      Source |       SS           df       MS      Number of obs   =     1,926
-------------+----------------------------------   F(13, 1912)     =      1.62
       Model |  68.6677428        13  5.28213406   Prob > F        =    0.0738
    Residual |    6248.136     1,912  3.26785355   R-squared       =    0.0109
-------------+----------------------------------   Adj R-squared   =    0.0041
       Total |  6316.80374     1,925  3.28145649   Root MSE        =    1.8077

------------------------------------------------------------------------------------
             happy |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
              gndr |   .0341415   .0831133     0.41   0.681    -.1288609    .1971438
              agea |   .0034322   .0022132     1.55   0.121    -.0009083    .0077727
                   |
            region |
        NE England |   .2305232   .2573078     0.90   0.370    -.2741102    .7351567
        NW England |   .2857697   .1813988     1.58   0.115    -.0699906    .6415299
Yorkshire and the Humber |  -.0806543   .2007088    -0.40   0.688    -.4742856     .312977
     East Midlands |   .1913606    .201564     0.95   0.343    -.2039479    .5866691
     West Midlands |   .0935543   .1999522     0.47   0.640    -.2985931    .4857016
   East of England |   .1065354   .1950467     0.55   0.585    -.2759913    .4890621
        SE England |   .4343349   .1837769     2.36   0.018     .0739106    .7947592
        SW England |   .3146064   .2022029     1.56   0.120    -.0819551    .7111679
             Wales |   .3251592   .2159191     1.51   0.132    -.0983026     .748621
          Scotland |   .3944005   .1995299     1.98   0.048     .0030814    .7857195
  Northern Ireland |   .5611539   .2586261     2.17   0.030     .0539349    1.068373
                   |
             _cons |   7.181876   .2206378    32.55   0.000      6.74916    7.614592
------------------------------------------------------------------------------------
```

Figure 1: Multiple linear regression model with happiness as outcome.

Remember that linear regression requires continuous variables as your dependent variable. In the model above, variables “happy” and “agea” are considered continuous, whereas “gndr” and “region” are categorical. It is important to tell Stata that it needs to treat specific variables as categorical, for your model will need $X-1$ *dummy variables* (where X is the total number of categories in a variable). Take "region" as an example: this variables has twelve categories (twelve regions), therefore your model will need eleven dummy variables and one *base group*. In the command above, Stata has been instructed to treat "London" as base group (`ib7.` does that, if you wanted the second category to be the base group instead, you can type `ib2.` before the variable name).

The value of R^2 tells you the proportion of variance in your dependent variable ("happy") that is explained by the model. In this case, we can see that this model does not really explain much, only 1.1 percent of variance or information. Moreover, if we look at the p-value for the F-test of overall significance (reported as $Prob > F$ in Stata) we can see that it is not statistically significant itself. This is telling us that our model does not fit any better than a model with no predictors (also called *intercept-only*).

Most coefficients are not statistically significant at the 0.05 level (we already knew this from the F-test of overall significance). A non-statistically significant coefficient is telling you that we cannot be sure the coefficient (i.e., the value of the parameter) is actually different from zero. If you look at the confidence interval of a non-statistically significant coefficient, you will see that the interval includes zero in its range. This means you should not interpret that coefficient. However, do acknowledge in your papers and reports that you have non-statistically significant coefficients by simply saying so.

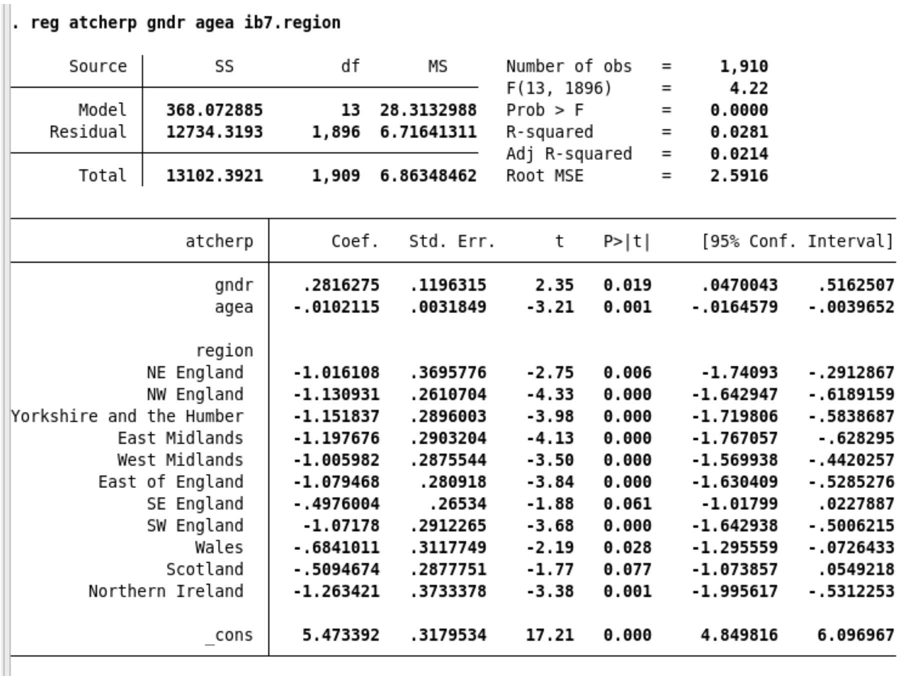
Let's now run the following regression model:

```
regress atcherp gndr agea ib7.region
```

Notice that this new model loses 16 observations (not many) when compared to our previous "happiness model" (the observations go from 1,926 to 1,910). However, now our F-test is statistically significant ($Prob > F = 0.0000$), meaning that our model fits better than an empty or intercept-only model. The model explains about 2.8 percent of information in the dependent variable "atcherp" (a 0-to-10 scale of emotional attachment to Europe).

It is of paramount importance to bear in mind that the interpretation of coefficients of categorical variables is totally different from the interpretation of coefficients of continuous variables. Below the coefficients of the variables "agea" (continuous) and "gndr" (categorical) are interpreted for you:

- "agea": keeping the rest of variables constant, for each one-unit increase in a respondent's age we expect a decrease of 0.01 points in emotional attachment to Europe.
- "gndr": keeping the rest of variables constant, female respondents present an emotional attachment to Europe 0.281 points higher than that of male respondents.



```
. reg atcherp gndr agea ib7.region

      Source |       SS           df       MS      Number of obs   =     1,910
-------------+----------------------------------   F(13, 1896)     =      4.22
       Model |  368.072885        13  28.3132988   Prob > F        =    0.0000
    Residual |  12734.3193     1,896  6.71641311   R-squared       =    0.0281
-------------+----------------------------------   Adj R-squared   =    0.0214
       Total |  13102.3921     1,909  6.86348462   Root MSE        =    2.5916

------------------------------------------------------------------------------------------
                  atcherp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------------------+---------------------------------------------------------------
                     gndr |   .2816275   .1196315     2.35   0.019     .0470043    .5162507
                     agea |  -.0102115   .0031849    -3.21   0.001    -.0164579   -.0039652
                          |
                   region |
               NE England |  -1.016108   .3695776    -2.75   0.006     -1.74093   -.2912867
               NW England |  -1.130931   .2610704    -4.33   0.000    -1.642947   -.6189159
Yorkshire and the Humber  |  -1.151837   .2896003    -3.98   0.000    -1.719806   -.5838687
            East Midlands |  -1.197676   .2903204    -4.13   0.000    -1.767057    -.628295
            West Midlands |  -1.005982   .2875544    -3.50   0.000    -1.569938   -.4420257
          East of England |  -1.079468    .280918    -3.84   0.000    -1.630409   -.5285276
               SE England |  -.4976004     .26534    -1.88   0.061     -1.01799    .0227887
               SW England |   -1.07178   .2912265    -3.68   0.000    -1.642938   -.5006215
                    Wales |  -.6841011   .3117749    -2.19   0.028    -1.295559   -.0726433
                 Scotland |  -.5094674   .2877751    -1.77   0.077    -1.073857    .0549218
         Northern Ireland |  -1.263421   .3733378    -3.38   0.001    -1.995617   -.5312253
                          |
                    _cons |   5.473392   .3179534    17.21   0.000     4.849816    6.096967
------------------------------------------------------------------------------------------
```

Figure 2: Multiple linear regression model with emotional attachment to Europe as outcome.

Firstly, notice that when interpreting regression coefficients in multiple regression *you must* keep constant the variables you are not interpreting. Secondly, when interpreting continuous variables we speak of “one-unit increase” if the coefficient is positive or of “one-unit decrease” if the coefficient is negative. However, when interpreting categorical variables (dummy variables) we compare groups to the base group. In the example above, there are only two categories: males and females. Because the variable is coded 1 "male” and 2 “female”, Stata takes by default the first value, “male”, as base group. The coefficient of “gndr”, therefore, is really the coefficient of female respondents in comparison to that of male respondents. Here too a positive coefficient means an increase, but it is an increase over the base group’s effect on the dependent variable. To clarify this point further consider the variable “region” below (remember the base group was “London”):

- “NE England”: keeping the rest of variables constant, respondents from the North East (England) are 1 point less emotionally attached to Europe than respondents from London.
- “Scotland”: keeping the rest of variables constant, respondents from Scotland are 0.51 points less emotionally attached to Europe than respondents from London.

Because all regional categories (i.e., all the dummy variables in the model) present a negative coefficient, we know that London is the region most emotionally attached to Europe. The interpretation of all these coefficients will therefore be in "negative" terms (less than the base group). Needless to say that if we did not know that "London" is the base group we could not interpret these figures. Also, notice that Scotland's coefficient is not statistically significant. What you already know about statistical significance also applies here in the same fashion: because $p > 0.05$ in the case of the category "Scotland", we cannot say for sure the coefficient is different for zero (and we do not interpret it). Finally, the closer the coefficient of a dummy variable is to zero the more similar that category is to the base group.