Stata Lab 6.
Multivariate Regression 2

Dr. Adam William Chalmers

Table of Contents

Multivariate Regression with Dummy Variables

So far we have learned how to run multivariate regression analyses using continuous (and sometimes ordinal) IVs. However, the use of nominal (binary and continuous) IVs is an important extension of regression analysis. We call these **dummy variables** and their interpretation is slightly different from ordinal and continuous variables.

Binary Dummy Variables

Binary dummy IVs take one of two values, and most frequently 0 or 1. These dummies could be at any level and are often used to distinguish between two groups. For instance, at the individual level, a very common dummy variable is 'gender' where men = 0 and women = 1. At the group level, we could create a dummy that distinguishes technology companies = 1 from companies operating in all other sectors = 0. We can also create binary dummies at the country level: USA = 1, all other countries = 0 or Eurozone countries = 1 and non-Eurozone countries = 0.

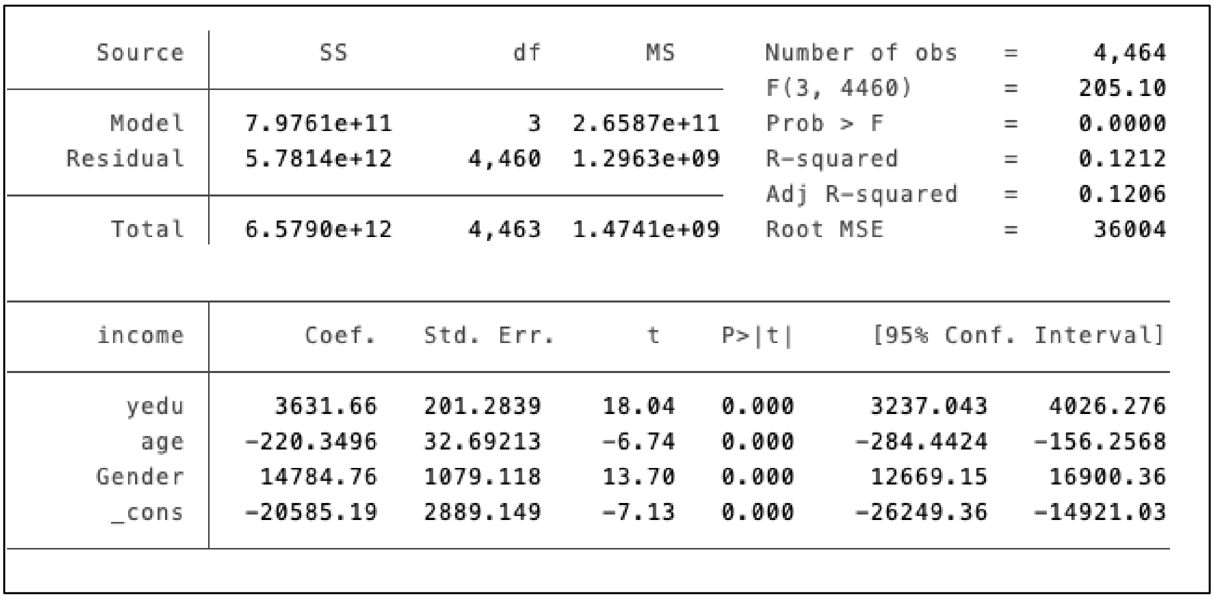
Example 1. Using **Data6.dta**, I want to run a multivariate regression explaining variation in individuals' 'income' (DV) by the IVs 'years of education' (yedu) and age (age). However, I also acknowledge that there may be important differences between men and women. Indeed, research suggests that being male is correlated with higher income levels. We therefore want to include a dummy variable for 'Gender' in our analysis.

First, we need to **create a new variable** for 'Gender', where Men = 1 and Women are the reference category =0. **Reference category** means that we will compare results for men against this category. Stata automatically takes the lower number to be the reference category. Hence, we code Women as 0 and Men as 1 in this case. This is pretty clear when we have binary dummies but will get more complicated with categorical dummies.

```
. recode sex (1=1 'Men')(2=0 'Women'), gen(Gender)
```

Next, run a multivariate regression analysis including all variables as well as our dummy variable Men.

```
. regress income yedu age Gender
```



```
      Source |       SS           df       MS      Number of obs   =     4,464
-------------+----------------------------------   F(3, 4460)      =    205.10
       Model |  7.9761e+11         3  2.6587e+11   Prob > F        =    0.0000
    Residual |  5.7814e+12     4,460  1.2963e+09   R-squared       =    0.1212
-------------+----------------------------------   Adj R-squared   =    0.1206
       Total |  6.5790e+12     4,463  1.4741e+09   Root MSE        =     36004

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |    3631.66   201.2839    18.04   0.000     3237.043    4026.276
         age |  -220.3496   32.69213    -6.74   0.000    -284.4424   -156.2568
      Gender |   14784.76   1079.118    13.70   0.000     12669.15    16900.36
       _cons |  -20585.19   2889.149    -7.13   0.000    -26249.36   -14921.03
------------------------------------------------------------------------------
```

Interpreting the results. First, we can see that years of education has a sizeable impact on income. With each additional year of education, we can expect income to increase by $3631.66USD. We can also see that this effect is statistically significant with a p-value <0.001. Next, we find evidence that being younger is correlated with earning more. With each additional year in age, we can expect a decrease in income of $220.3496USD. This effect is also statistically significant with a p-value <0.001. Finally, we find that being Gender, and in particular being a man, has a sizeable positive impact on income. However, here we interpret the results a bit differently. There is no 'one unit increase' in our binary variable 'Men'. Instead, we need to think of the regression coefficient as the mean difference in the DV for Men compared to the reference category, Women. Hence, we can say that the **mean income difference** for Men compared to Women is $14784.76USD. We can also see that this effect is statistically significant with a p-value <0.001.

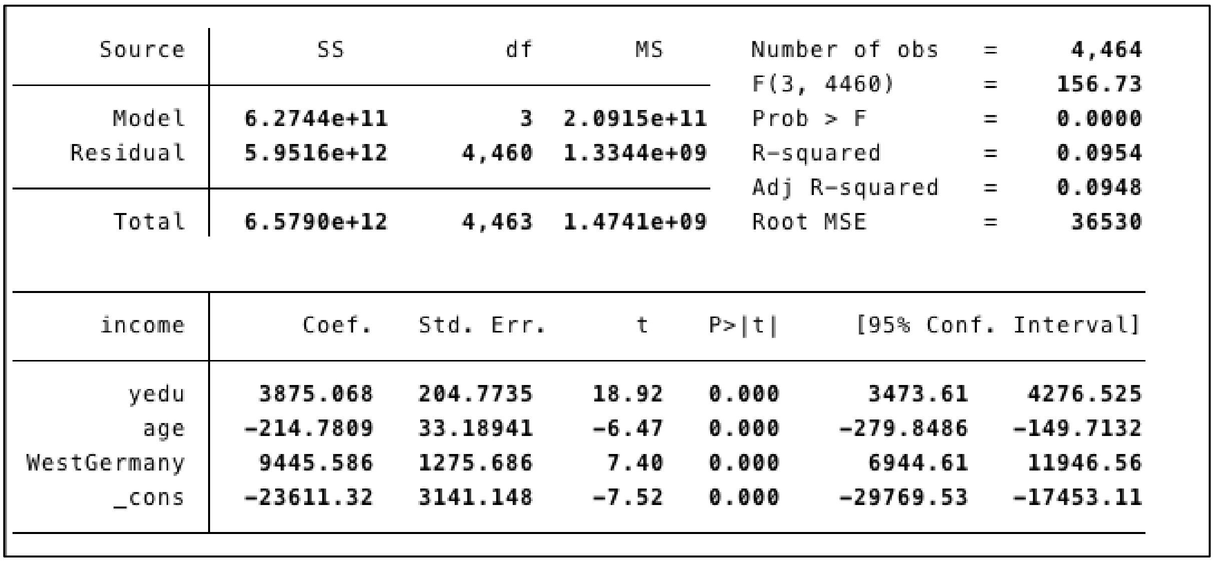
Example 2. We want to replicate the multivariate regression analysis above, but instead of controlling for differences between men and women, we want to include a binary dummy variable for individuals living in West Germany. Our assumption is that the West is more prosperous compared to the East and this might have an important impact on income.

First, we will need to create a dummy variable for 'West Germany'. This new variable will distinguish between West Germany and East Germany based on the geographical location of the different German states.

```
. recode state (0 1 2 3 4 5 6 7 8 9 10 = 1 "West")(11 12 13 14 15 16  = 0 "East"),
gen(WestGermany)
```

In this case, East Germany = 0 and is our reference category. Recall that Stata automatically takes the lower value as the reference category (unless we tell it otherwise). Hence, East = 0 and is the references category for West which = 1.
Next, we run our regression model including our new dummy variable.

```
. regress  income yedu age WestGermany
```



```
      Source |       SS           df       MS      Number of obs   =     4,464
-------------+----------------------------------   F(3, 4460)      =    156.73
       Model |  6.2744e+11         3  2.0915e+11   Prob > F        =    0.0000
    Residual |  5.9516e+12     4,460  1.3344e+09   R-squared       =    0.0954
-------------+----------------------------------   Adj R-squared   =    0.0948
       Total |  6.5790e+12     4,463  1.4741e+09   Root MSE        =     36530

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |   3875.068   204.7735    18.92   0.000      3473.61    4276.525
         age |  -214.7809   33.18941    -6.47   0.000    -279.8486   -149.7132
 WestGermany |   9445.586   1275.686     7.40   0.000      6944.61    11946.56
       _cons |  -23611.32   3141.148    -7.52   0.000    -29769.53   -17453.11
------------------------------------------------------------------------------
```

Interpreting the results. We can see that the **mean income difference** for those living in West Germany compared to those living in East Germany is $9445.586 USD. We can also see that this difference is statistically significant with a p-value <0.001.

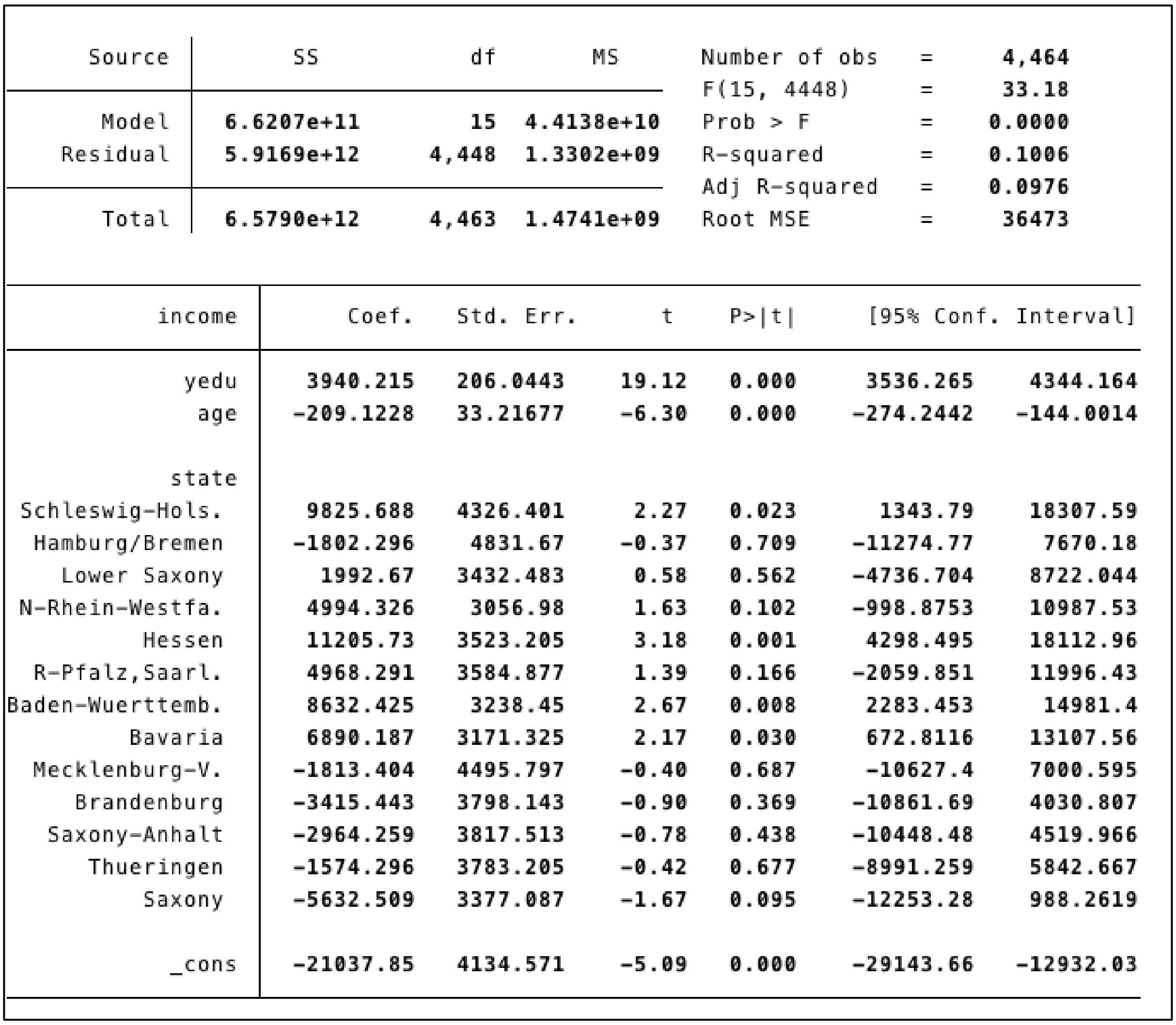
Categorical Dummy Variables

In addition to binary dummy variables we can also include categorical dummy variables in our multivariate regression analyses. Again, these can be measured at any level (individual, group, country, region, international). Please see the lectures and notes from Week 2 if you need a refresher on the nominal categorical level of measurement.

Example 1. Again, we want to replicate the multivariate regression analysis above, but this time we want to control for differences between all of the 16 German Bundesländer (rather than just East and West). Our logic is a bit broader than before when we had very specific expectations about the differences between East and West Germany. Now we are acknowledging that, giving the federal structure of the German Bundesländer that there might be important differences across the Bundesländer that possibly influence income. These differences remain undefined. In fact, using categorical dummies in this way is a good strategy for dealing with variation that is otherwise **not observed** in your other indictors. For instance, there might be 'difficult to observe' cultural differences across the Bundesländer that are otherwise not captured in our analyses. Our categorical dummies are meant to capture these differences.

To include a categorical dummy in your multivariate regression analysis you need to add the prefix 'i.' to the beginning of the variable. This tells Stata to show us results for the different categories in our dummy variable. Stata automatically selects the first category as the reference category, which in this case is Berlin. Here's how it looks for our example outlined above.

```
. regress  income yedu age i.state
```



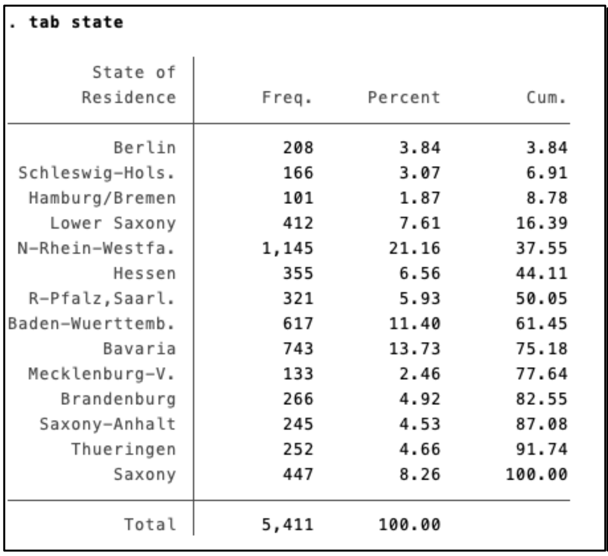
```
      Source |       SS           df       MS      Number of obs   =     4,464
-------------+----------------------------------   F(15, 4448)     =     33.18
       Model |  6.6207e+11        15  4.4138e+10   Prob > F        =    0.0000
    Residual |  5.9169e+12     4,448  1.3302e+09   R-squared       =    0.1006
-------------+----------------------------------   Adj R-squared   =    0.0976
       Total |  6.5790e+12     4,463  1.4741e+09   Root MSE        =     36473

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |   3940.215   206.0443    19.12   0.000     3536.265    4344.164
         age |  -209.1228   33.21677    -6.30   0.000    -274.2442   -144.0014
             |
       state |
 Schleswig-Hols. |   9825.688   4326.401     2.27   0.023      1343.79    18307.59
  Hamburg/Bremen |  -1802.296    4831.67    -0.37   0.709    -11274.77     7670.18
    Lower Saxony |    1992.67   3432.483     0.58   0.562    -4736.704    8722.044
 N-Rhein-Westfa. |   4994.326    3056.98     1.63   0.102    -998.8753    10987.53
          Hessen |   11205.73   3523.205     3.18   0.001     4298.495    18112.96
  R-Pfalz,Saarl. |   4968.291   3584.877     1.39   0.166    -2059.851    11996.43
Baden-Wuerttemb. |   8632.425    3238.45     2.67   0.008     2283.453     14981.4
         Bavaria |   6890.187   3171.325     2.17   0.030     672.8116    13107.56
  Mecklenburg-V. |  -1813.404   4495.797    -0.40   0.687     -10627.4    7000.595
     Brandenburg |  -3415.443   3798.143    -0.90   0.369    -10861.69    4030.807
   Saxony-Anhalt |  -2964.259   3817.513    -0.78   0.438    -10448.48    4519.966
      Thueringen |  -1574.296   3783.205    -0.42   0.677    -8991.259    5842.667
          Saxony |  -5632.509   3377.087    -1.67   0.095    -12253.28    988.2619
             |
       _cons |  -21037.85   4134.571    -5.09   0.000    -29143.66   -12932.03
------------------------------------------------------------------------------
```

Interpreting the results. You can see that we now have results for Germany's various Bundesländer. We can see that some Bundesländer do seem to have significant differences when it comes to explaining income. We would interpret these coefficients in the same way as we do with binary dummy variables. For example, we can see that the **mean income difference** for those living in Hessen compared to those living in the reference category (Berlin) is $11205.73USD. We can also see that this difference is statistically significant with a p-value <0.001. If you look closely you will see that our reference category, Berlin, is not present in our regression results. This is because Berlin has been used as our reference category (the same way that Women were used as the reference category in the example above). As already noted, Stata automatically takes the first category as the reference category. We can, however, manually select our own reference category by adding to the prefix we learned above. For example, to select Schlewig-Hols as the reference category we would change our command to this.

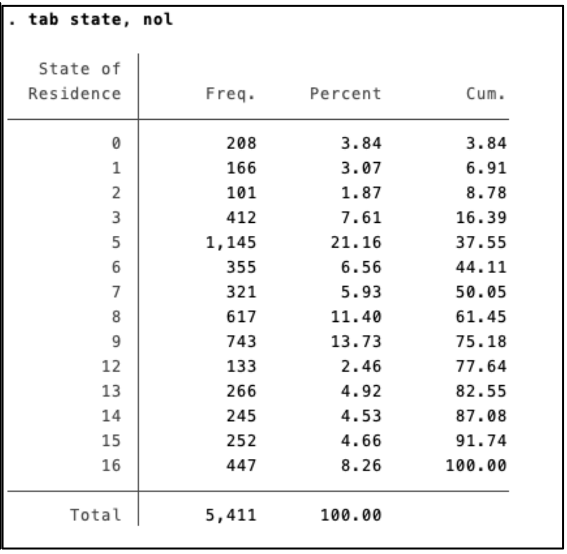
```
. regress  income yedu age ib1.state
```

The '**ib1**' relates to the fact that the value label for Schlewig-Hols = 1.

Recall that we get the numeric values for the value labels by comparing frequency tables, like I've done below.



```
. tab state

    State of |
   Residence |      Freq.     Percent        Cum.
-------------+-----------------------------------
      Berlin |        208        3.84        3.84
Schleswig-Hols. |     166        3.07        6.91
 Hamburg/Bremen |     101        1.87        8.78
  Lower Saxony |      412        7.61       16.39
N-Rhein-Westfa. |   1,145       21.16       37.55
       Hessen |       355        6.56       44.11
 R-Pfalz,Saarl. |     321        5.93       50.05
Baden-Wuerttemb. |    617       11.40       61.45
      Bavaria |       743       13.73       75.18
 Mecklenburg-V. |     133        2.46       77.64
   Brandenburg |      266        4.92       82.55
 Saxony-Anhalt |      245        4.53       87.08
    Thueringen |      252        4.66       91.74
       Saxony |       447        8.26      100.00
-------------+-----------------------------------
       Total |      5,411      100.00
```

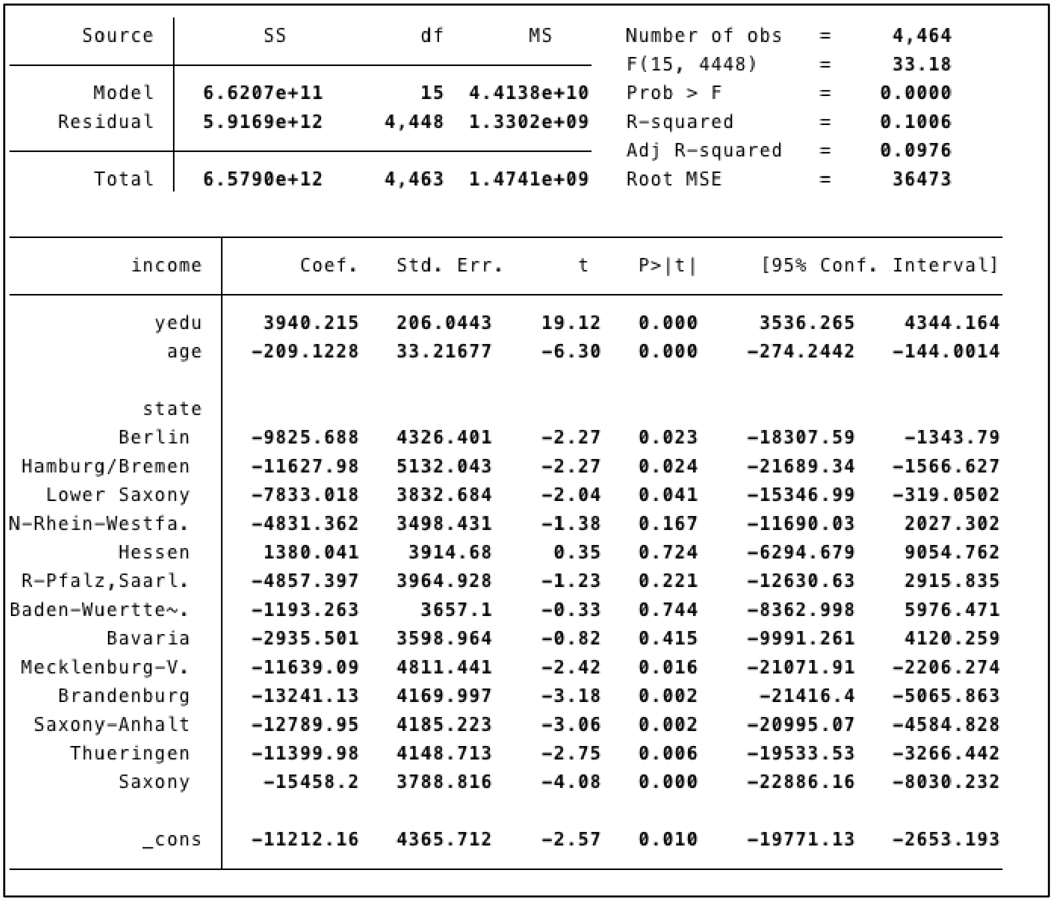


```
. tab state, nol

    State of |
   Residence |      Freq.     Percent        Cum.
-------------+-----------------------------------
           0 |        208        3.84        3.84
           1 |        166        3.07        6.91
           2 |        101        1.87        8.78
           3 |        412        7.61       16.39
           5 |      1,145       21.16       37.55
           6 |        355        6.56       44.11
           7 |        321        5.93       50.05
           8 |        617       11.40       61.45
           9 |        743       13.73       75.18
          12 |        133        2.46       77.64
          13 |        266        4.92       82.55
          14 |        245        4.53       87.08
          15 |        252        4.66       91.74
          16 |        447        8.26      100.00
-------------+-----------------------------------
       Total |      5,411      100.00
```

Below are our regression results.

```
. regress  income yedu age ib1.state
```

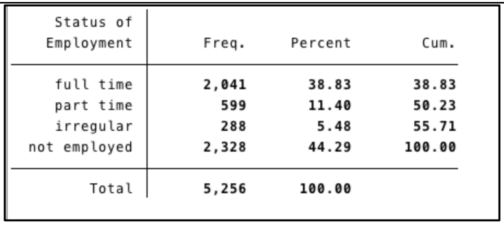
You can now see that Berlin is included in our results table, and Schlewig-Hols has been removed because it is now our reference category.



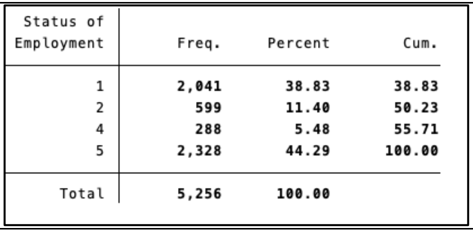
```
      Source |       SS           df       MS      Number of obs   =     4,464
-------------+----------------------------------   F(15, 4448)     =     33.18
       Model |  6.6207e+11        15  4.4138e+10   Prob > F        =    0.0000
    Residual |  5.9169e+12     4,448  1.3302e+09   R-squared       =    0.1006
-------------+----------------------------------   Adj R-squared   =    0.0976
       Total |  6.5790e+12     4,463  1.4741e+09   Root MSE        =     36473

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |   3940.215   206.0443    19.12   0.000     3536.265    4344.164
         age |  -209.1228   33.21677    -6.30   0.000    -274.2442   -144.0014
             |
       state |
      Berlin |  -9825.688   4326.401    -2.27   0.023    -18307.59    -1343.79
 Hamburg/Bremen |  -11627.98   5132.043    -2.27   0.024    -21689.34   -1566.627
  Lower Saxony |  -7833.018   3832.684    -2.04   0.041    -15346.99   -319.0502
N-Rhein-Westfa. |  -4831.362   3498.431    -1.38   0.167    -11690.03    2027.302
       Hessen |   1380.041    3914.68     0.35   0.724    -6294.679    9054.762
 R-Pfalz,Saarl. |  -4857.397   3964.928    -1.23   0.221    -12630.63    2915.835
Baden-Wuertte~. |  -1193.263     3657.1    -0.33   0.744    -8362.998    5976.471
      Bavaria |  -2935.501   3598.964    -0.82   0.415    -9991.261    4120.259
 Mecklenburg-V. |  -11639.09   4811.441    -2.42   0.016    -21071.91   -2206.274
   Brandenburg |  -13241.13   4169.997    -3.18   0.002     -21416.4   -5065.863
 Saxony-Anhalt |  -12789.95   4185.223    -3.06   0.002    -20995.07   -4584.828
    Thueringen |  -11399.98   4148.713    -2.75   0.006    -19533.53   -3266.442
       Saxony |   -15458.2   3788.816    -4.08   0.000    -22886.16   -8030.232
             |
       _cons |  -11212.16   4365.712    -2.57   0.010    -19771.13   -2653.193
------------------------------------------------------------------------------
```

Example 2. We are now interested in controlling for 'Status of employment' (emp). We should first inspect the different categories in this variable.

```
. tab emp
. tab emp, nolabel
```



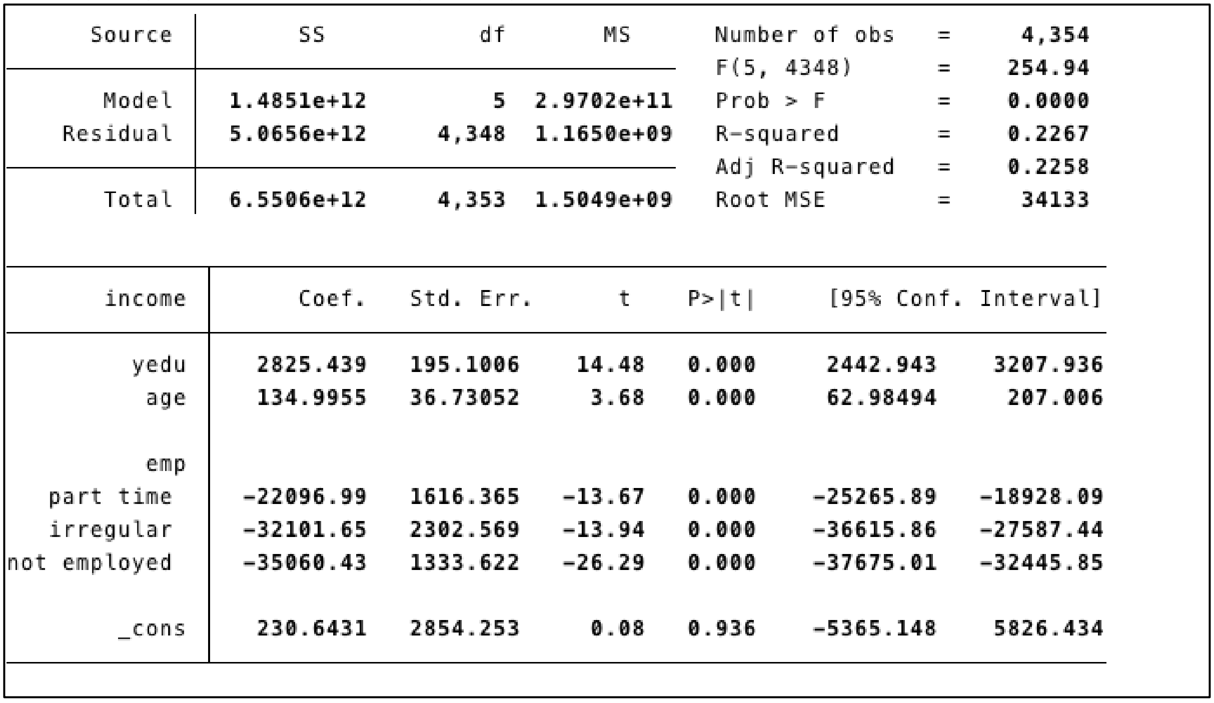
```
   Status of
  Employment      Freq.     Percent        Cum.
------------+-----------------------------------
   full time      2,041       38.83       38.83
   part time        599       11.40       50.23
   irregular        288        5.48       55.71
not employed      2,328       44.29      100.00
------------+-----------------------------------
       Total      5,256      100.00
```



```
   Status of
  Employment      Freq.     Percent        Cum.
------------+-----------------------------------
           1      2,041       38.83       38.83
           2        599       11.40       50.23
           4        288        5.48       55.71
           5      2,328       44.29      100.00
------------+-----------------------------------
       Total      5,256      100.00
```

Our hunch is that there will be important differences in income between people who are employed full time compared to the other categories.

```
. regress  income yedu age i.emp
```

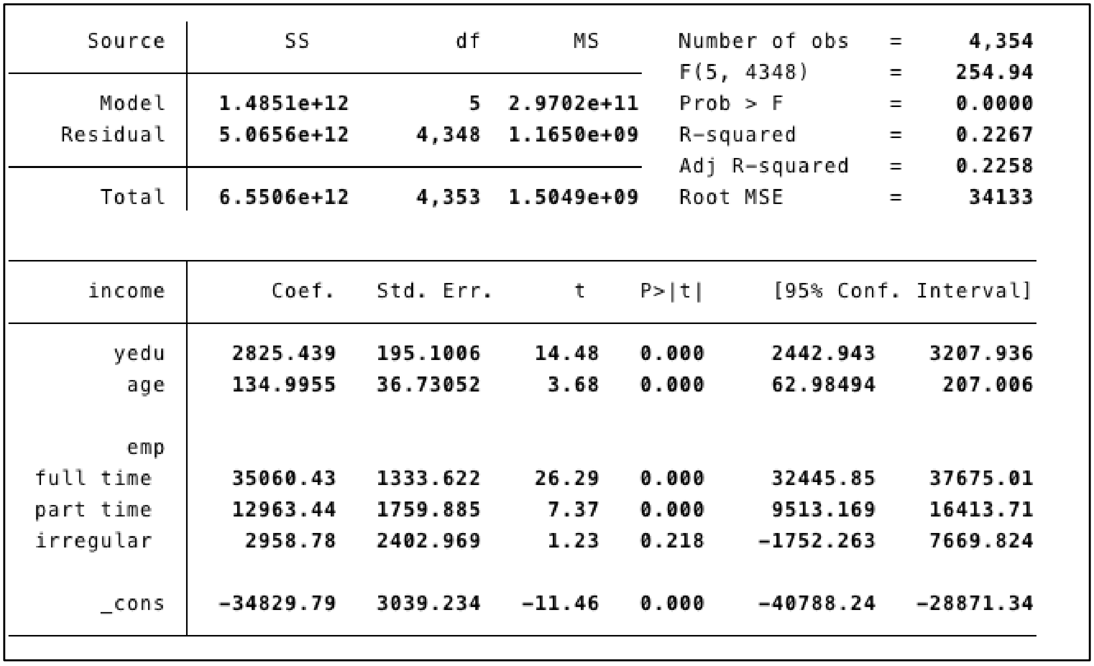


```
      Source |       SS           df       MS      Number of obs   =     4,354
-------------+----------------------------------   F(5, 4348)      =    254.94
       Model |  1.4851e+12         5  2.9702e+11   Prob > F        =    0.0000
    Residual |  5.0656e+12     4,348  1.1650e+09   R-squared       =    0.2267
-------------+----------------------------------   Adj R-squared   =    0.2258
       Total |  6.5506e+12     4,353  1.5049e+09   Root MSE        =     34133

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |   2825.439   195.1006    14.48   0.000     2442.943    3207.936
         age |   134.9955   36.73052     3.68   0.000     62.98494     207.006
             |
         emp |
   part time |  -22096.99   1616.365   -13.67   0.000    -25265.89   -18928.09
   irregular |  -32101.65   2302.569   -13.94   0.000    -36615.86   -27587.44
not employed |  -35060.43   1333.622   -26.29   0.000    -37675.01   -32445.85
             |
       _cons |   230.6431   2854.253     0.08   0.936    -5365.148    5826.434
------------------------------------------------------------------------------
```

Interpreting the results. Even with just a quick glance, we can see that 'part time' employment, 'irregular' employment, and 'not employed' all have mean incomes that are much lower than our reference category, 'full time' employment. This is not a big surprise. After all, we should expect people who are employed full time to make more than those who work part time, irregularly, or are not employed. All of these effects are also statistically significant with a p-value <0.001.

****Pro-tip**: You can choose any reference category you like. However, you should choose one that makes interpreting your results easier. In the case above where we are using the categorical dummy variable 'emp', it might have made more sense to use 'not employed' as our reference category. Then we could see the differences between not being employed and different types of employment. Below are the results that we'd get if we did use 'not employed' as our reference category.

. regress income yedu age ib5.emp



```
      Source |       SS           df       MS      Number of obs   =     4,354
-------------+----------------------------------   F(5, 4348)      =    254.94
       Model |  1.4851e+12         5  2.9702e+11   Prob > F        =    0.0000
    Residual |  5.0656e+12     4,348  1.1650e+09   R-squared       =    0.2267
-------------+----------------------------------   Adj R-squared   =    0.2258
       Total |  6.5506e+12     4,353  1.5049e+09   Root MSE        =     34133

------------------------------------------------------------------------------
      income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yedu |   2825.439   195.1006    14.48   0.000     2442.943    3207.936
         age |   134.9955   36.73052     3.68   0.000     62.98494     207.006
             |
         emp |
  full time  |   35060.43   1333.622    26.29   0.000     32445.85    37675.01
  part time  |   12963.44   1759.885     7.37   0.000     9513.169    16413.71
  irregular  |    2958.78   2402.969     1.23   0.218    -1752.263    7669.824
             |
       _cons |  -34829.79   3039.234   -11.46   0.000    -40788.24   -28871.34
------------------------------------------------------------------------------
```

Categorical Variables: Controlling for Time

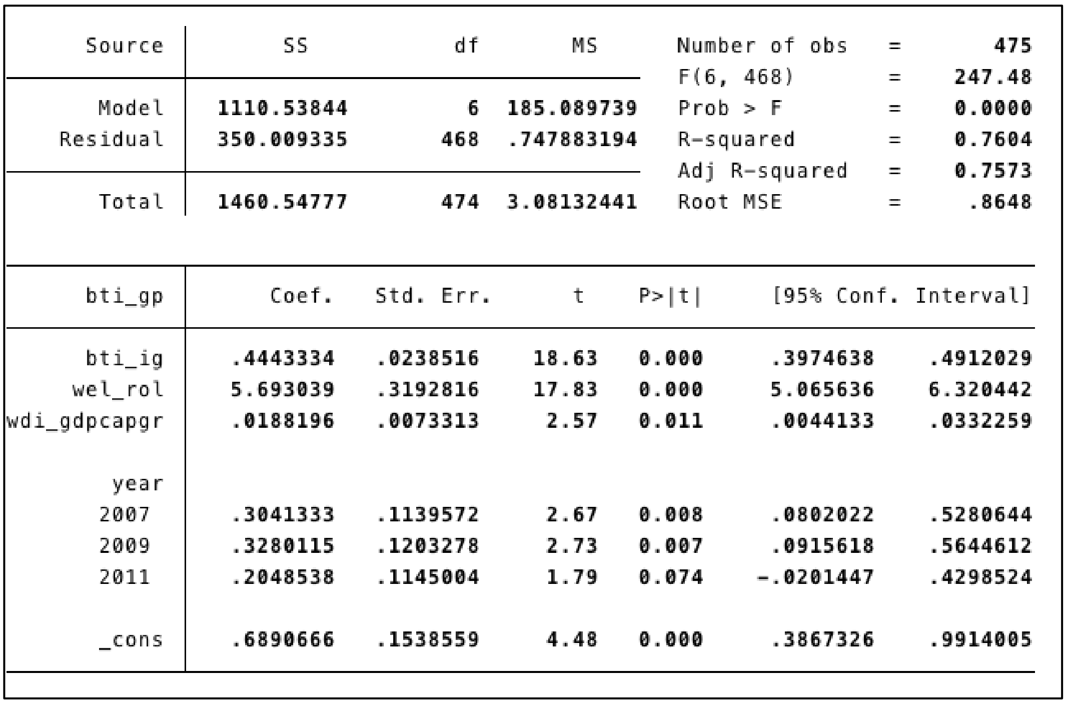
We can also use categorical variables to control for otherwise unobserved variation over **time**. In other words, this means including ‘time dummies’ in our multivariate regression analysis. This is not the same as time-series analysis or multi-level regression models. However, this is one basic way to include a time dimension into our analysis.

Example 1. Using **QOG_timeseries.dta**, we want to examine the following hypothesis:

> H1. The more interest groups in a political system, the more efficient governance will be.

Our DV is governance performance (bti_gp), our main IV is number of interest groups (bti_ig), and we control for GDP (wdi_gdpcapgr) and rule of law (wel_rol). We also include ‘year’ as our categorical dummy variable. As above, use the ‘i’ prefix for year.

. regress bti_gp bti_ig wel_rol wdi_gdpcapgr i.year



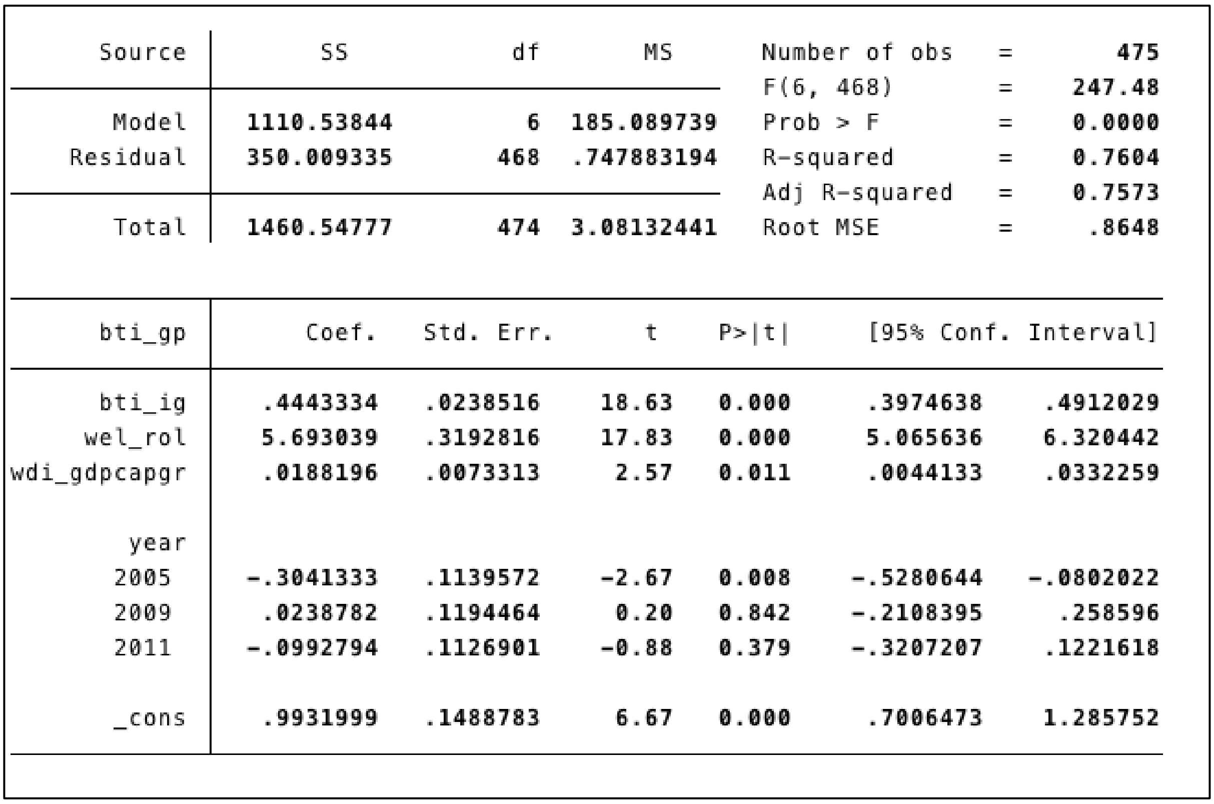
```
      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(6, 468)       =    247.48
       Model |  1110.53844         6  185.089739   Prob > F        =    0.0000
    Residual |  350.009335       468  .747883194   R-squared       =    0.7604
-------------+----------------------------------   Adj R-squared   =    0.7573
       Total |  1460.54777       474  3.08132441   Root MSE        =     .8648

------------------------------------------------------------------------------
      bti_gp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bti_ig |   .4443334   .0238516    18.63   0.000     .3974638    .4912029
     wel_rol |   5.693039   .3192816    17.83   0.000     5.065636    6.320442
wdi_gdpcapgr |   .0188196   .0073313     2.57   0.011     .0044133    .0332259
             |
        year |
        2007 |   .3041333   .1139572     2.67   0.008     .0802022    .5280644
        2009 |   .3280115   .1203278     2.73   0.007     .0915618    .5644612
        2011 |   .2048538   .1145004     1.79   0.074    -.0201447    .4298524
             |
       _cons |   .6890666   .1538559     4.48   0.000     .3867326    .9914005
------------------------------------------------------------------------------
```

Interpreting the results. First, we do see evidence supporting our hypothesis. With each one unit increase in 'interest groups', we can expect a .044 increase in the Governance Performance index. This effect is statistically significant with a p-value <0.001. We have also controlled for differences across years. Specifically, we can see that the **mean difference in the Governance Performance Index** for 2007 compared to the reference category year (which happens to be 2005 due to missing data in 2006) is 0.3041333. We can also see that this effect is statistically significant with a p-value <0.01. If you looked at all the years available in the dataset you will see that they run from 2000 to 2018. However, there are far fewer years included in the regression. This is the case because of the way the data has been collected: there are many years, in other words, where we don't have data on our DV and/or IVs.

****Pro-tip**: the easiest way to figure out the reference category when you have missing data on different years, to continue with our example, is to change the reference category. We know that Stata automatically selects the lowest value category as the reference category. Hence, if you change the reference to anything else and run the regression you will see the old reference category listed in the results.

Changing the reference category for a numeric variable is different than changing it for a string variable. To change our reference category to 2007, for example, we would write the following.

```
. regress bti_gp bti_ig wel_rol wdi_gdpcapgr ib2007.year
```



```
      Source |       SS           df       MS      Number of obs   =       475
-------------+----------------------------------   F(6, 468)       =    247.48
       Model |  1110.53844         6  185.089739   Prob > F        =    0.0000
    Residual |  350.009335       468  .747883194   R-squared       =    0.7604
-------------+----------------------------------   Adj R-squared   =    0.7573
       Total |  1460.54777       474  3.08132441   Root MSE        =     .8648

------------------------------------------------------------------------------
      bti_gp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      bti_ig |   .4443334   .0238516    18.63   0.000     .3974638    .4912029
     wel_rol |   5.693039   .3192816    17.83   0.000     5.065636    6.320442
wdi_gdpcapgr |   .0188196   .0073313     2.57   0.011     .0044133    .0332259
             |
        year |
       2005  |  -.3041333   .1139572    -2.67   0.008    -.5280644   -.0802022
       2009  |   .0238782   .1194464     0.20   0.842    -.2108395     .258596
       2011  |  -.0992794   .1126901    -0.88   0.379    -.3207207    .1221618
             |
       _cons |   .9931999   .1488783     6.67   0.000     .7006473    1.285752
------------------------------------------------------------------------------
```

Stata Exercises. 1

Task 1. We are interested in testing the following hypothesis.

> H1. The more a country is ethnically fractionalised, the less a country will be globalised.

Using **QOG_timeseries.dta**, our DV is the Index of Globalisation (**dr_ig**). Higher values correspond to being more globalised. Our main IV is Ethnic Fractioanlisation (**al_ethnic**). Higher values correspond to being more fractionalised ("Reflects probability that two randomly selected people from a given country will not belong to the same [ethnic] group. The higher the number, the more fractionalized society"). We also include the following IVs: Freedom to trade internationally (**fi_ftradeint_pd**), Inward Foreign Direct Investment (**wdi_fdiin**) and Outward Foreign Direct Investment (**wdi_fdiout**). Finally, we are interested in a country's colonial past and in particular, whether or not a country has one. We therefore need to create a binary dummy for Colonial Past.

- Step 1. Create a binary dummy variable using the variable Colonial Origin (**ht_colonial**), where 1 = colonial past and 0 = no colonial past. Call your variable ColonialPast.
- Step 2. Run a multivariate regression analysis testing H1 and controlling for ColonialPast. Interpret the results for all of your variables including your binary dummy.

Task 2. We want to test the same hypothesis as in Task 1 and include all of the same IVs. However, this time we want to replace our Binary Dummy with a Categorical dummy for 'Regime type' (**ht_regtype1**). Run your regression analysis using the category 'Democracy' as your reference category. Interpret the results for all IVs and the following categories of your Dummy: Monarchy & Military.

Task 3. Introduce a time dummy into the same regression model (be sure to omit the Regime type dummy).

1. Question 1. Compare this model to the model with the Regime type dummy. Which model has more explanatory power (and by how much)?
2. Question 2. What is the reference category for your categorical dummy?
3. Question 3. Interpret the regression results for 2001 & 2015.

Marginal Effects and Predicted Probabilities

Regression coefficients tells us about the magnitude of our expected effects. For instance, with a one unit increase in the IV, we can expect some x increase or decrease in the DV. However, these interpretations can sometimes feel a bit abstract. This is particularly the case when we start to include binary and categorical dummy variables in our analysis. One way to improve our ability to interpret results for multivariate regression analysis is to estimate and present 'marginal effects'. A 'margin' is a statistic computed from predictions from a model of the covariation between an IV and a DV. In plain terms, margins will give us very specific predicted values of our DV when our IV is at another very specific value. In Stata, marginal effects is a post-estimation technique. It is also something that relies heavily on visualisation for interpretation. A good rule of thumb in statistics is that, when possible, present your results visually. I.e., replace confusing and crowded tables with nice visualisations.

The are many basic Stata commands for marginal effects. Here are a few of the main ones:

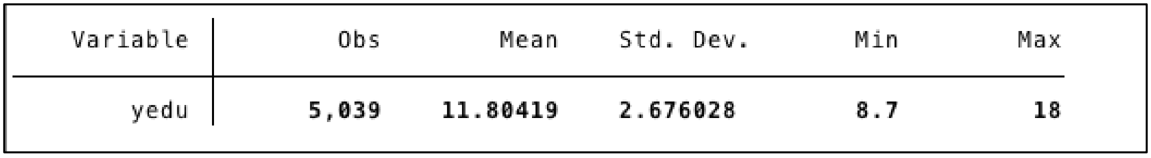
```
. margins, at(var1=(min(increment)max))
. margins, at(var1=(# #))
. margins var1, at(var2=(min(increment)max))
. margins, at (var1  == #)(var1 == #)(var1 == #) …
. margins var1, at (var2  == #)(var2 == #)(var1 == #) …
```

Example. Going back to our first example and using **data6.dta**, above, we want to run a multivariate regression explaining variation in individuals' 'income' (DV) by the IVS 'years of education' (yedu) and 'age' (age) while also controlling for 'Gender'.

```
. recode sex (1=1 'Men')(2=0 'Women'), gen(Gender)
```

This time, we want to supplement our regression results with the post-estimation of marginal effects of yedu on income. This will take several steps and will require you to know the range of values that yedu take.

```
. sum yedu
```



```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        yedu |      5,039    11.80419    2.676028        8.7         18
```

Our IV, yedu, has a minimum value of 8.7 and a maximum value of 18 with a standard deviation of about 2.6.

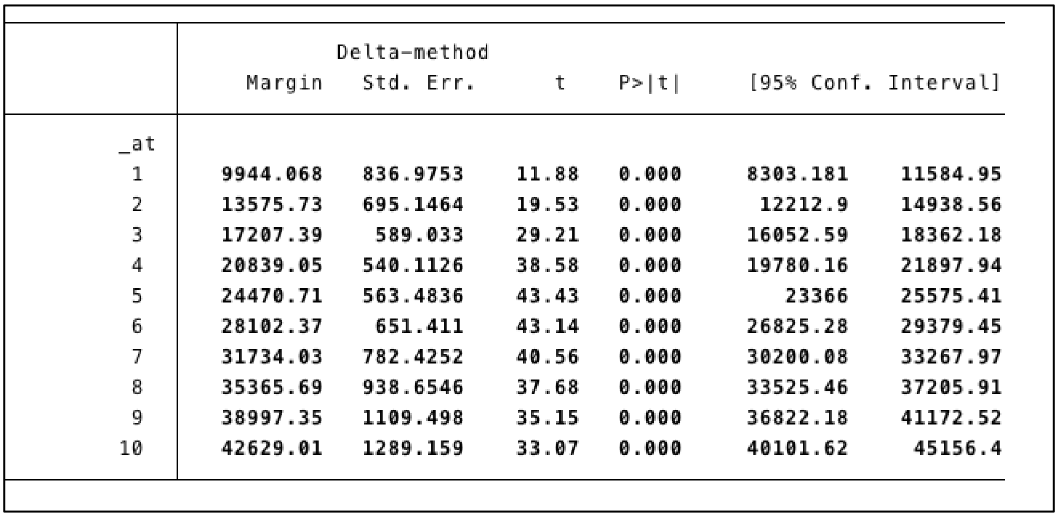
Let's run our regression analysis and then the post-estimation command for margins.

```
. regress  income yedu age Men
. margins, at(yedu=(8.7(1)18))
```

What does this margins command mean?

- 'margins' is a Stata command for predicting marginal effects
- 'yedu' is our IV of interest
- 8.7 is the min value of yedu
- 18 is the max value of yedu
- 1 is the 'increment' that we have told Stata to predict between the min and max values. In this case, 1 means that we have requested predicted results for each 1 year of yedu between our min and max values. We could have just as easily specified '2' for an increment of 2 years, or 0.5 for an increase of 6 months. It is up to you (and is also largely related to how your IV is measured and its min and max values).

Here are the results.



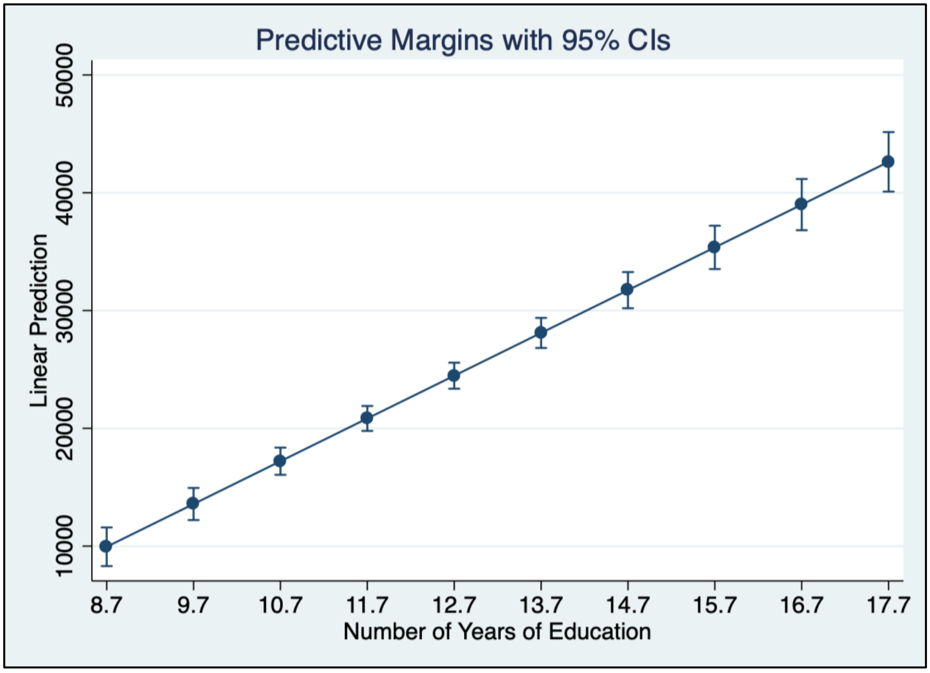
```
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   9944.068   836.9753    11.88   0.000     8303.181    11584.95
          2  |   13575.73   695.1464    19.53   0.000      12212.9    14938.56
          3  |   17207.39    589.033    29.21   0.000     16052.59    18362.18
          4  |   20839.05   540.1126    38.58   0.000     19780.16    21897.94
          5  |   24470.71   563.4836    43.43   0.000        23366    25575.41
          6  |   28102.37    651.411    43.14   0.000     26825.28    29379.45
          7  |   31734.03   782.4252    40.56   0.000     30200.08    33267.97
          8  |   35365.69   938.6546    37.68   0.000     33525.46    37205.91
          9  |   38997.35   1109.498    35.15   0.000     36822.18    41172.52
         10  |   42629.01   1289.159    33.07   0.000     40101.62     45156.4
------------------------------------------------------------------------------
```

Interpreting the results. In the left column (under **_at**) we see our 1 year increments of yedu between our min value and our max value. There are 10 increments. In the '**margin**' column we get the results of a predicted value for income for people for each increment of education. For example, at increment 1 (8.7 years of education), we can predict that the average income will be $9944.068. Critically, this prediction is statistically significant with a p-value <0.001. At increment 2 (9.7 years of education) we can predict that the average income will be $13575.73. This is also statistically significant with a p-value <0.001. We get these predicted values for income for each increment in yedu. Finally, in the two columns on the right we get our **confidence intervals**. The main thing to take from these columns is the range of values (the lower bounds and the upper bounds) with which we can be 95% confident (p-value <0.05) of our results. These values for increment 1 are between $8303.181 (the lower bound) and $11584.95 (the upper bound). Any values outside of these lower and upper bounds are beyond the range in which we can be 95% confident.

Marginsplot

All of the information about marginal effects displayed above is very useful because it gives us a very specific sense of how our IV affects our DV. However, we rarely replicate this table in reports. Instead, the best way to report these results is visually using the 'marginsplot' command. Again, this is part of the post-estimation technique we learned above. It is therefore critical that these commands are executed in the correct order. Here are the commands.

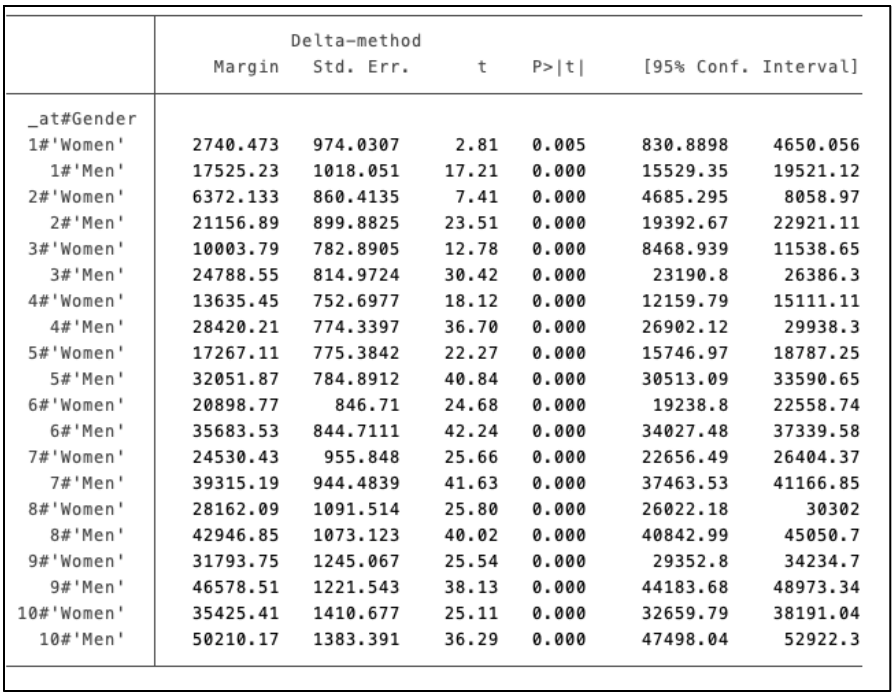
```
. regress  income yedu age Gender
. margins, at(yedu=(8.7(1)18))
. marginsplot
```



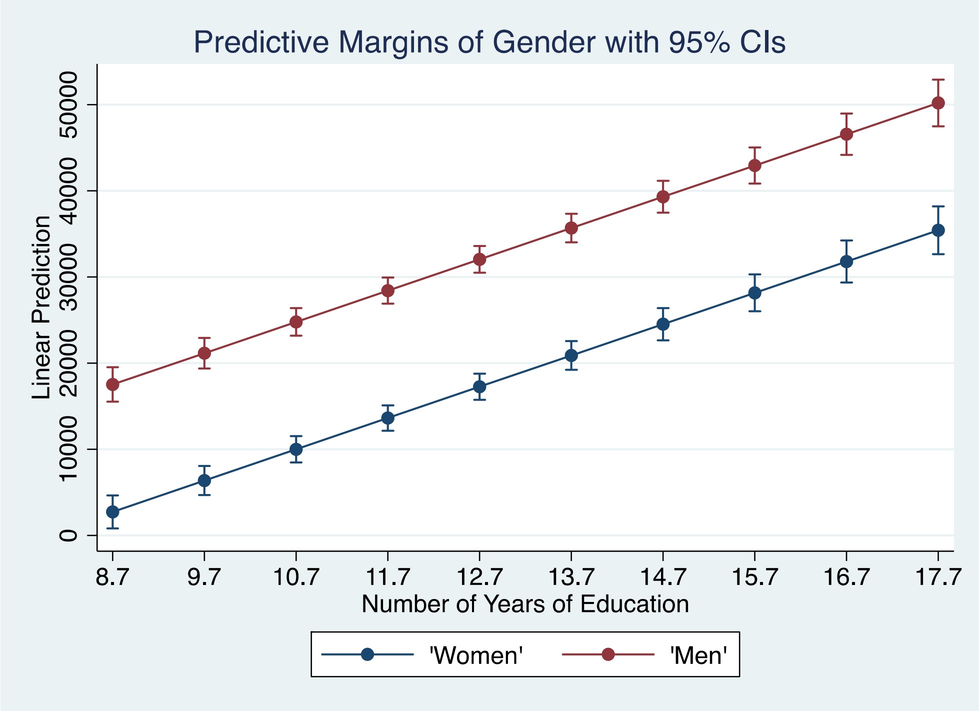
Interpreting the results. All of the values that are plotted here are listed in the table we produced above. However, now we get a very clear and intuitive sense of the impact of Years of education on Income. Along the x-axis we can see our 'increments' for yedu. The dots are the predicted values for income. The capped-lines are the upper and lower bounds of the confidence intervals. A general rule of thumb for confidence intervals is that the tighter (smaller!) the better.

Example. We want to include our binary dummy variable 'Gender' in our predicted values for 'income' at different values of 'yedu'. This time, we need to include the prefix 'i.' for our binary dummy 'Gender' so that Stata includes Men and the reference category, Women.

```
. regress  income yedu age i.Gender
. margins Gender, at(yedu=(8.7(1)18))
. marginsplot
```



```
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  _at#Gender |
  1#'Women'  |   2740.473   974.0307     2.81   0.005     830.8898    4650.056
    1#'Men'  |   17525.23   1018.051    17.21   0.000     15529.35    19521.12
  2#'Women'  |   6372.133   860.4135     7.41   0.000     4685.295     8058.97
    2#'Men'  |   21156.89   899.8825    23.51   0.000     19392.67    22921.11
  3#'Women'  |   10003.79   782.8905    12.78   0.000     8468.939    11538.65
    3#'Men'  |   24788.55   814.9724    30.42   0.000      23190.8     26386.3
  4#'Women'  |   13635.45   752.6977    18.12   0.000     12159.79    15111.11
    4#'Men'  |   28420.21   774.3397    36.70   0.000     26902.12     29938.3
  5#'Women'  |   17267.11   775.3842    22.27   0.000     15746.97    18787.25
    5#'Men'  |   32051.87   784.8912    40.84   0.000     30513.09    33590.65
  6#'Women'  |   20898.77     846.71    24.68   0.000      19238.8    22558.74
    6#'Men'  |   35683.53   844.7111    42.24   0.000     34027.48    37339.58
  7#'Women'  |   24530.43    955.848    25.66   0.000     22656.49    26404.37
    7#'Men'  |   39315.19   944.4839    41.63   0.000     37463.53    41166.85
  8#'Women'  |   28162.09   1091.514    25.80   0.000     26022.18       30302
    8#'Men'  |   42946.85   1073.123    40.02   0.000     40842.99     45050.7
  9#'Women'  |   31793.75   1245.067    25.54   0.000      29352.8     34234.7
    9#'Men'  |   46578.51   1221.543    38.13   0.000     44183.68    48973.34
 10#'Women'  |   35425.41   1410.677    25.11   0.000     32659.79    38191.04
   10#'Men'  |   50210.17   1383.391    36.29   0.000     47498.04     52922.3
------------------------------------------------------------------------------
```

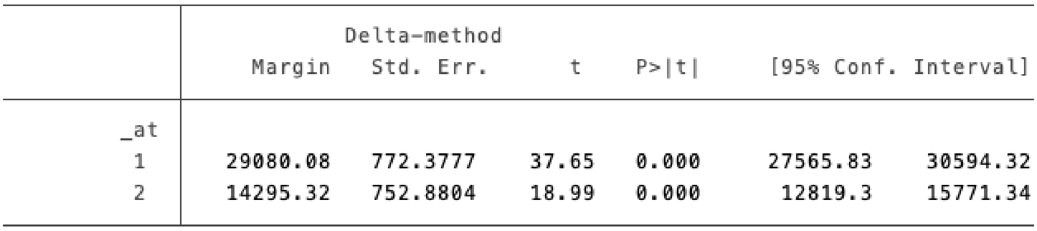


Interpreting the results. While both women and men benefit from more education when it comes to income, men benefit much more. This effect appears to be consistent across all increments of education.

Example. Omitting ‘yedu’ for the time being, we now want to estimate the predicted values of ‘income’ for Men vs Women and graph this using ‘marginsplot’.

```
. regress  income yedu age i.Gender
. margins, at(Gender=(1 0))
```

Note that I have changed the values for the min and max values and have not specified an increment. This is because our binary dummy variable Men only has two values.

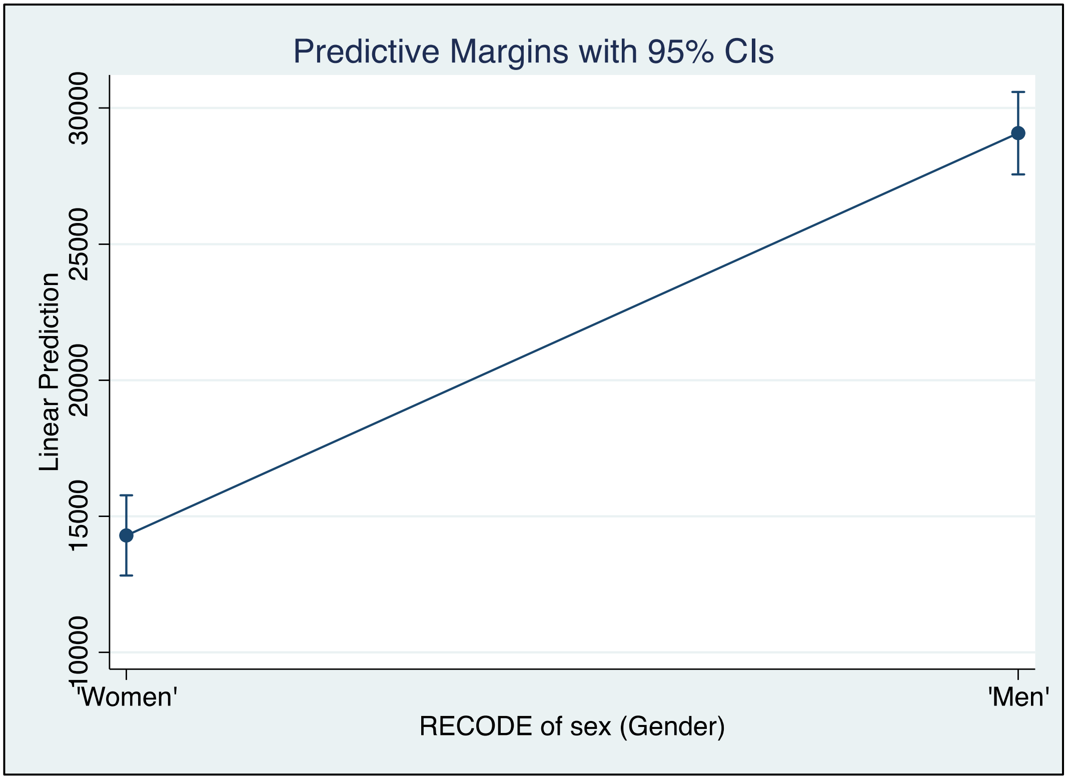


```
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   29080.08   772.3777    37.65   0.000     27565.83    30594.32
          2  |   14295.32   752.8804    18.99   0.000      12819.3    15771.34
```

Interpreting the results. In this case in the **'_at'** column we see the predicted values for 1 = men and 2 = women. We can see that the predicted mean income for men is $29080.08, which is much higher than for women, who have a predicted income of just $14295.32. Both results are statistically significant with p-values of <0.001.

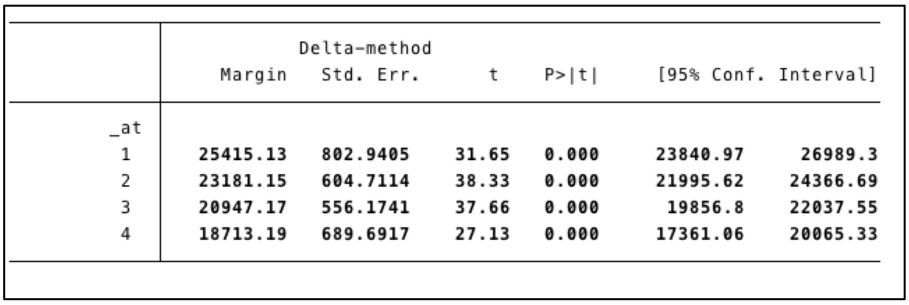
Now, let's visualise the results with the 'marginsplot' command.

```
. marginsplot
```



Example. We now want to examine the predicted values of income for differences across our variable 'age'. However, in this case our focus is on precited values for income for people aged 20, 30, 40, and 50.

```
. regress  income yedu age
. margins, at(age == 20) at(age == 30)  at(age == 40)  at(age == 50)
```

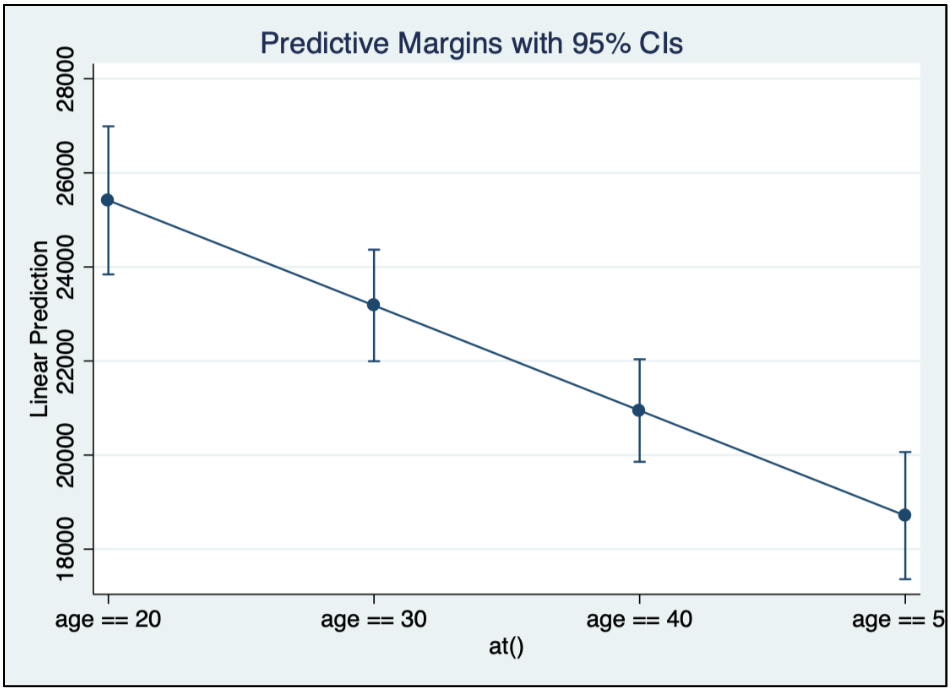


```
             |            Delta-method
             |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         _at |
          1  |   25415.13   802.9405    31.65   0.000     23840.97     26989.3
          2  |   23181.15   604.7114    38.33   0.000     21995.62    24366.69
          3  |   20947.17   556.1741    37.66   0.000      19856.8    22037.55
          4  |   18713.19   689.6917    27.13   0.000     17361.06    20065.33
------------------------------------------------------------------------------
```

Interpreting the results. The column **'_at'** is now at our different age increments. 1 = 20; 2 = 30, 3 = 40, and 4 = 50. We can clearly see that the predicted values for income decrease with age. All results are statistically significant with p-values of <0.001.

Let's also plot these results using 'marginsplot'.

. marginsplot



Stata Exercises. 2

Task 1. Using **QOG_timeseries.dta** and testing the same hypothesis in Task 1-3, complete the following.

- Step 1. Run a multivariate analysis including dr_ig al_ethnic fi_ftradeint_pd wdi_fdiin wdi_fdiout as well as your Binary Dummy variable for 'ColonialPast' and predict marginal effects for the two values of your Binary Dummy.
- Step 2. Visual your results.

Task 2. Run a multivariate analysis including dr_ig (your DV), al_ethnic, fi_ftradeint_pd, wdi_fdiin, and wdi_fdiout. Predict marginal effects for 'al_ethnic' at increments of 0.01 and ranging from 0 to 1.

1. Task 1. What is the predicted value of dr_ig when al_ethnic is at 0.01?
2. Task 2. Is this above our below the mean for dr_ig?
3. Task 3. What is the lower bound for this increment and what is the upper bound?
4. Task 4. Visualise your results.

Task 3. Rerun your regression and margins, but:

- Step 1. This time we want to predict values for al_ethinc for our two values of the Binary Dummy 'ColonialPast'.
- Step 2. Visualise the results.

Task 4. Run a multivariate analysis including 'dr_ig', 'al_ethnic', 'fi_ftradeint_pd', 'wdi_fdiin', 'wdi_fdiout' and this time include your Categorical Dummy variable for Regime Type (ht_regtype1).

- Step 1. Predict marginal values for dr_ig when regime type is Monarchy, Military, and Democracy.
- Step 2. Visualise your results.

Graphing with marginsplot

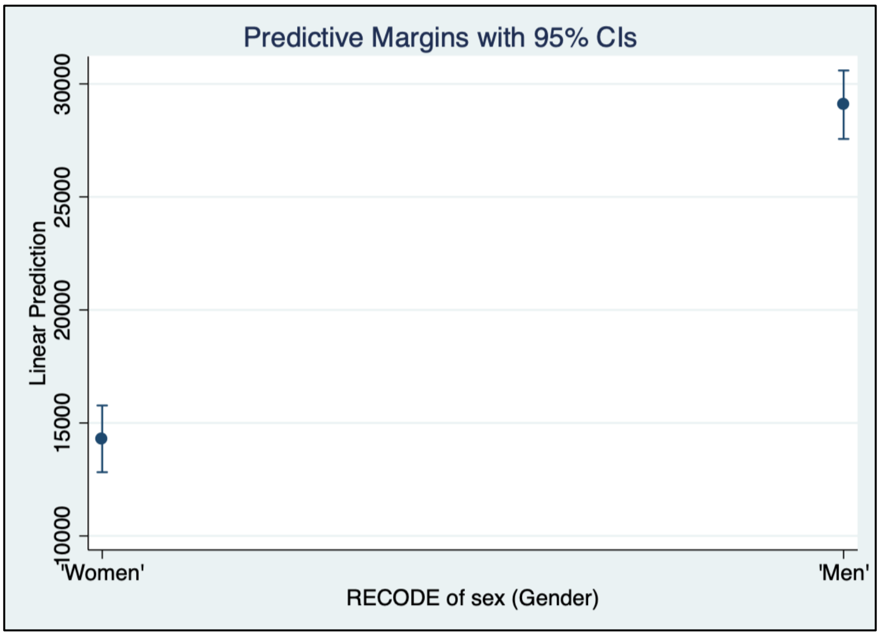
There are many options that you can use to improve your graphs using marginsplot. You can explore these many options using Stata's built in help function. In this section I will point out some basic options for improving your graphing skills in marginsplot.

Scatter and Line

When our variable of interest is **binary** or **categorical** we can graph our margins plot using the 'scatter' option. When the variable of interest is **continuous**, we can graph our margins plot using the 'line' option and also adjusting the confidence intervals. Making these adjustments requires us to use the 'recast' and 'recastci' options. Let's take our examples from what we already learned above.

Scatter and Binary Variable

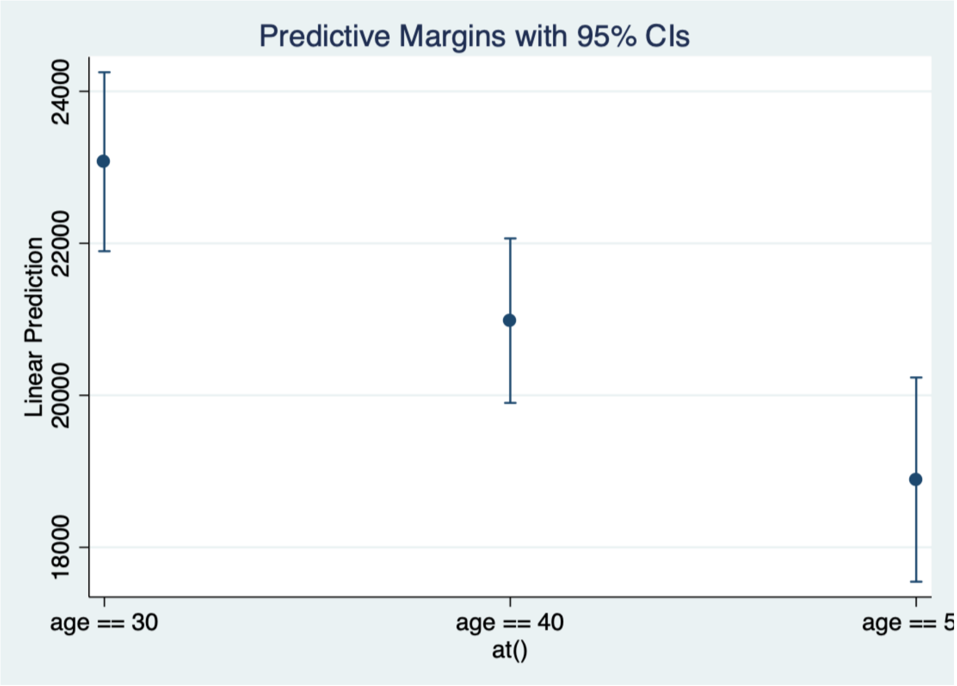
```
. regress  income yedu age i.Gender
. margins, at(Gender=(1 0))
. marginsplot, recast(scatter)
```



This is the same result as we got before, not there is no line joining the two dots. This is the proper way to visualise these marginal effects.

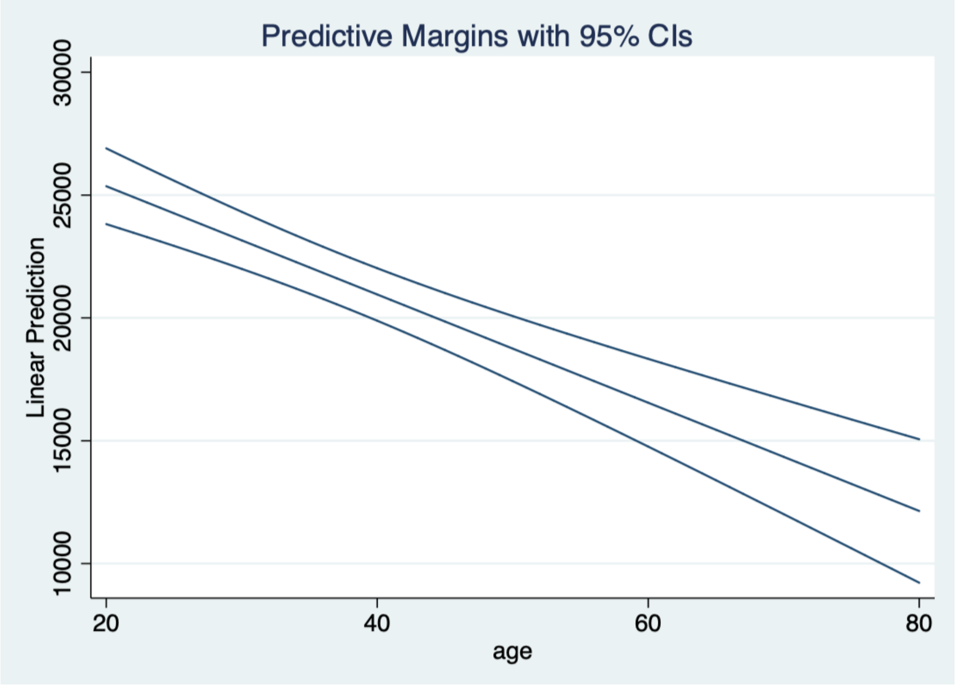
Scatter and Categorical Variable

```
. regress  income yedu age i.state
. margins, at(age == 30)  at(age == 40)  at(age == 50)
. marginsplot, recast(scatter)
```



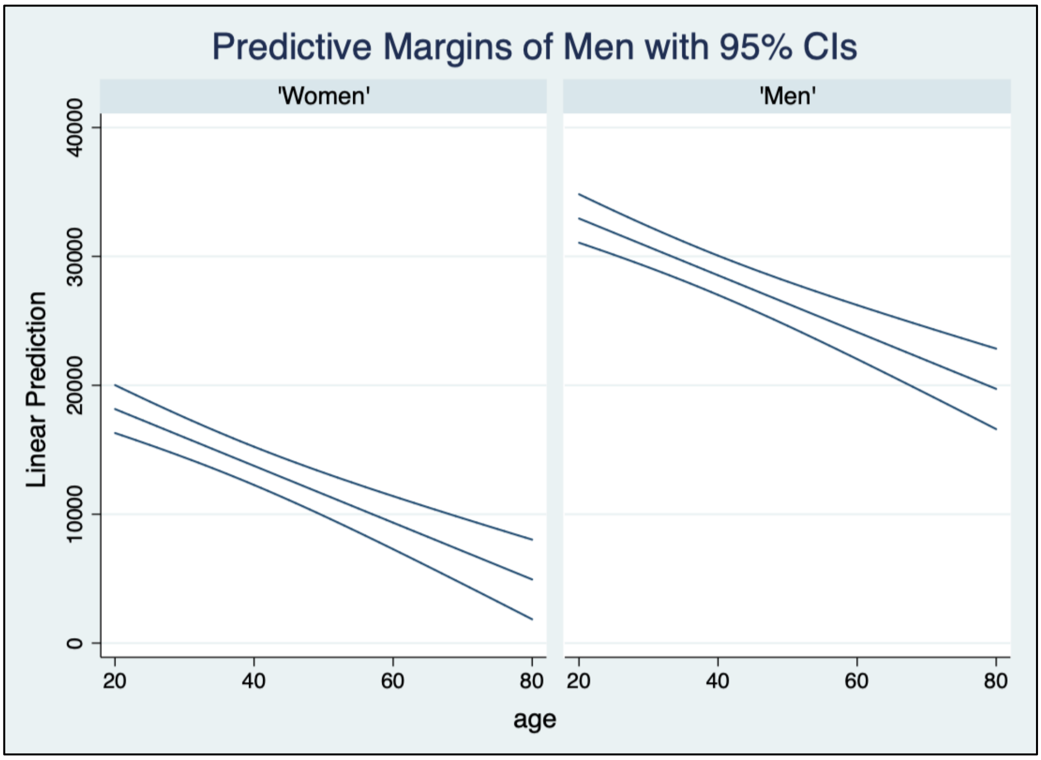
Line and Continuous Variable.

```
. regress  income yedu age i.Men
. margins, at(age=(20(1)80))
. marginsplot, recastci(rline) recast(line)
```

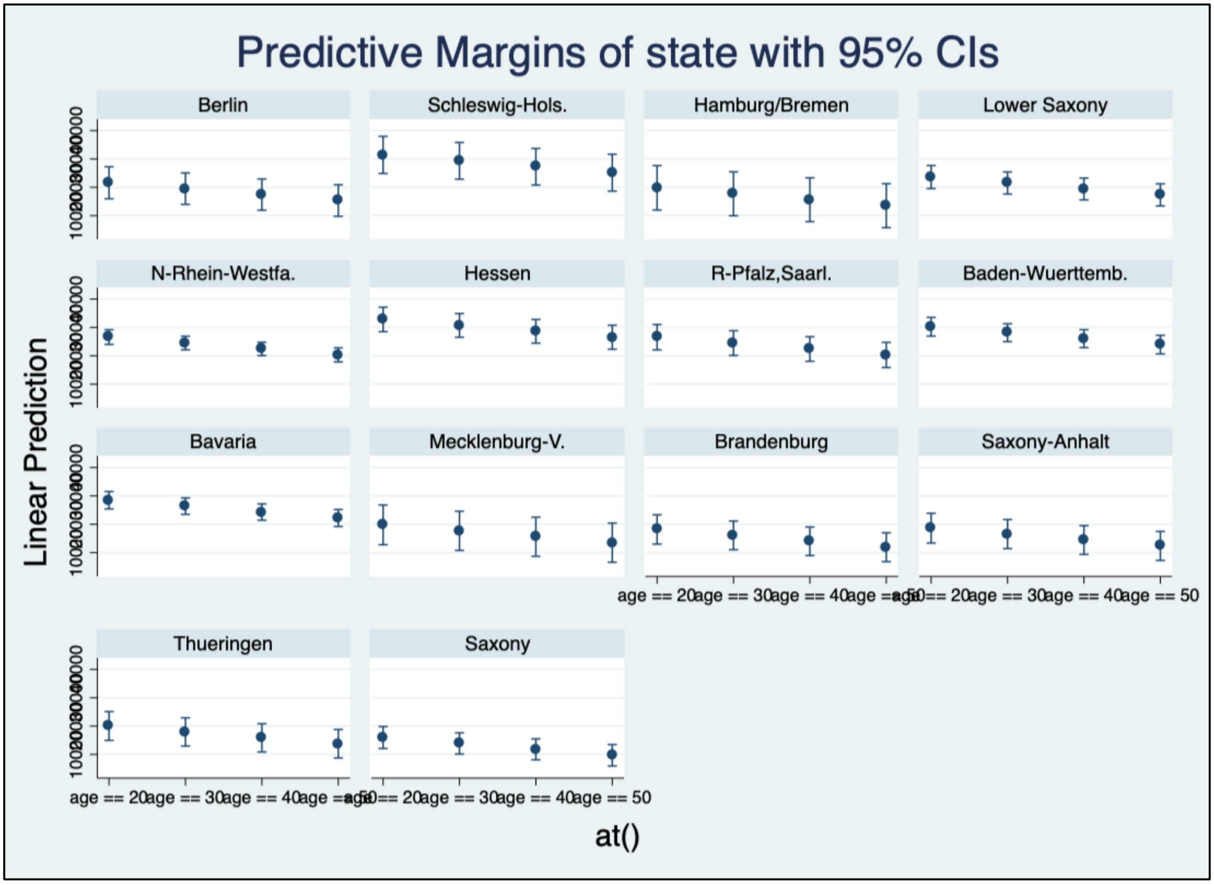


Over

```
.regress  income yedu age i.Men
.margins Men, at(age=(20(1)80))
.marginsplot, by(Men) recastci(rline) recast(line)
```



```
. regress  income yedu age i.state
. margins state,  at(age == 20)  at(age == 30)  at(age == 40)  at(age == 50)
. marginsplot, by(state)  recast(scatter)
```



Stata Exercises. 3

Task 1. Re-draw your graphs from tasks 4-7, but improve them using the techniques for scatter and line outlined above.

Task 9. We want to examine the relationship between Globalisation (dr_ig) and Ethnic Fractionalisation (al_ethnic), controlling for alternative explanations (fi_ftradeint_pd wdi_fdiin wdi_fdiout) as well as a new binary dummy variable for Democracy.

- Step 1. Recode ht_regtype1 into a binary dummy variable where 1 = Democracy and 0 = Non-democracy. Call your new variable ‘Democracy’.
- Step 2. Predict marginal effects for ‘al_ethnic’ at increments of 0.01 and ranging from 0 to 1 for the two categories of your Dummy variable Democracy.
- Step 3. Visualise your results using the appropriate graphing format and include two graphs (side by side), comparing Democracy to Non Democracy.

Advanced Task: The Privacy Survey

For the advanced task this week you will be using the data collected through the **Privacy Attitudes / Privacy Behaviours Survey**. First some background. Researchers have recently pinpointed something call the 'privacy paradox'. In short, the privacy paradox means that people have very **risk adverse attitudes** about privacy issues but very **risk perverse behaviours**. This is particularly the case when it comes to internet privacy. For instance, most people would say they value the privacy of their online identity and personal information, but rarely do things to actually protect their privacy (using VPNs, reading 'terms of agreement', checking cookies, etc). Their attitudes do not match their behaviour.

For this assignment you will start with the 'raw' survey data in a xlsx file and work your way up to two major tests: (1) a paired t-test to test the privacy paradox hypothesis and (2) a multivariate regression analysis examining the determinants of a 'privacy index'.

STEPS

Preparing your data:

1. import **Privacydata.xlsx**.
2. encode the string variable 'gender'
3. create a new variable 'age' using dob (data of birth).
4. Rename the following
   a. doyouconsideryourselftobealight1 = SocialMediaUse
   b. doyoutendtotrustotherpeopleveryl = TrustOtherPeople
   c. doyoutendtobelievewhatyoureadint = BelieveNewsMedia

Creating Privacy Indices:

1. create two different indices. One for **PrivacyAttitudes** and one for **PrivacyBehaviours**. Use the data from the remaining survey questions to create these two indices. It is up to you which data to include in which index. For each survey question, a response of Yes = 1 and a response of No = 0. Also note that Attitudes are about things like 'concern' and 'worry' and behaviors are what people actually do. Be sure to standardize each index (sum up the components and then divide by the number of components).
2. Run a paired t-test to test the privacy paradox hypothesis. This hypothesis says that Privacy Attitudes are different from Privacy Behaviors. Hence, the null hypothesis is that there are no differences between the two, while the alternative hypothesis says that there is a (statistically significant) difference. Do you find support for the Privacy Paradox Hypothesis?
3. Create a new index called the PrivacyIndex, which combines the PrivacyAttitudes and PrivacyBehaviours indices. Besure to standardize your new PrivacyIndex (sum up the components and then divide by the number of components).
4. Using your new PrivacyIndex as your DV, run a multivariate regression analysis where your IVs are (1) SocialMediaUse (2) TrustOtherPeople (3) BelieveNewsMedia, (4) Age and (5) Gender. Note that Gender is a binary dummy variable.
5. Report the results of your regression analysis and interpret regression coefficients, p-values, and adjusted r-squared.

6. Use pre- and post-estimation checks for multicollinearity. Are there any problematic IVs?
7. Using semi-partial correlations, which IV has the most explanatory power?
8. Predict marginal effects for all of your IVs and visualize using ‘marginsplot.’