How much did the woman with the highest weekly earnings earn?

```
summ earn
summ earn, det (det IS FOR DETAILED)
```

What is the number of children that has the highest share of mothers possessing a university degree ("degree +")?

```
tab nchild quals, row (row USED FOR ROW%)
```

How many countries have **all regions with non-missing values on X**?

```
sort code
by code: egen tot_nonmis=count(X)
by code: gen tot=_N
by code: gen sh_nonmis=tot_nonmis/tot
tab code if sh_nonmis==1
```

List variables for for Argentina:

```
list x1 x2 x3 … if code=="ARG"
```

Notice there are **11 observations less when we move to multivariate: list** these regions

```
list region if x1!=. & x2!=. & x3==.
```

We can also sort observation for two variables! (sort x1 x2)

~ RESIDUALS ~

Analyze the **residuals** of your regression. Are they **homoskedastic**? (**scatterplot**)

```
reg Y X1

predict res, res
predict yhat, xb

scatter res yhat, yline(0) msize(vsmall)
drop res yhat [if need to do it again]
```

Adjust for the **heteroskedasticity** (**robust** formula) of the residuals, if necessary.

```
reg Y X1, robust
```

~ GOODNESS OF FIT ~

Run regression and see **goodness of fit**:

```
reg violentcrime prop_black
*You can either exploit the information from the table…
*R-sq
di ESS/0.TSS
di 1-RSS/TSS
*…or you can also calculate it entirely manually (building yourself the RSS and the TSS)
egen avg_y=mean(y)
gen diff=violentcrime-avg_y
gen diff_sq=diff^2
egen tss=sum(diff_sq)

predict res, res
gen res_sq=res^2
egen rss=sum(res_sq)

gen aux=rss/tss
gen r2=1-aux
summ r2

*Adj R-sq
di 1-(n-1/n-k)*(RSS/TSS)      k is number of regressors! (including b0)
```

****

Omitted variable bias.

```
reg Y X1        (save b1univ)
reg Y X1 X2     (save b1multiv and b2multiv)
```

What is the bias on X1?

```
di b1univ-b1multiv
```

How can we retrieve the bias to doublecheck? We miss one element

```
reg X2 X1        (save b1univ-3)
di b2multiv*b1univ-3
```

Did we expect this direction of the omitted variable bias? Discuss (INTUITIVELY!)

```
corr Y X1 X2
```

Remember:

- When a variable that should be a number is saved as a string something is wrong
- When you import data from Excel save the Excel file in the comma-separated values (.csv) format and then use the ***insheet*** command
- You can show the content of a textfile in Stata with the ***type*** command
- You can show and change the current directory in Stata with the command ***cd***
- Missing values in Stata are denoted by "."
- Variable names in Stata cannot be numbers
- When you have a sequence of variables with similar names and only differing in referring to a different year, it is more efficient to use the *forvalues* loop

*** Asterisk**: you can use the asterisk for variable names in commands. For example, describe c* displays information on the variables with names beginning with the letter c.

STRINGS

Nonnumeric data are recorded as strings, typically enclosed in double quotes such as "UK". They can be a combination of alphabetic and numeric characters. They are stored as str# where # is an integer between 1 and 30 specifying the number of characters in the string. Two useful commands related with string variables are ***destring***, which converts string data to integer data and ***tostring***, which does the reverse.