

# Geneva Academics in Management and Economics

## *THE DO'S AND DON'TS OF A DO-FILE*

tips on how to do empirical work with Stata

Emmanuel Milet (emmanuel.milet@unige.ch)

15<sup>th</sup> Nov. 2016

# The do's and don'ts of a do-file

Why bother with a learning how to write a do-file?

- **THE** most important tool of your empirical analysis.
- You may share it with co-authors
- You will return to it several month after submitting your paper (and you'll possibly do this multiple times too)
- You may be asked by an editor to make your code and data publicly available

# The do's and don'ts of a do-file

Some important steps:

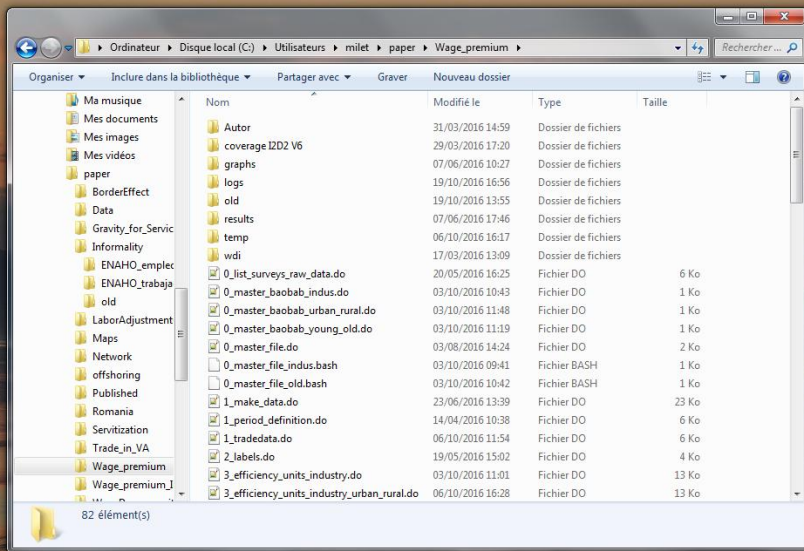
- 1 Folder
- 2 Master do-file
- 3 Logs
- 4 Names and label variables
- 5 Regression output
- 6 Various commands

# The do's and don'ts of a do-file: FOLDER

You are going to generate TONS of output while doing your empirical analysis. To make things *easier* and *clearer*, I recommend to create a folder for each type of output:

- logs
- graphs
- regression results
- tables and other output for the stylized facts
- manuscript
- an “old” folder where you can store do-files and other output that you do not need anymore

Your data and do-file can be stored in the root folder.



# The do's and don'ts of a do-file: MASTER DO FILE

A master do-file is a do-file which launches your other do-files in a specific order.

- It allows you to organize your do-files, and add comments to them.
- It makes sure that your final results (in your paper) are produced correctly.
- It allows you to declare a directory path, as well as other options that you may use across all your do-files.

```

1      * -----
2      * Wage Premium paper: master file
3      * -----
4      *
5      clear*
6      set more off
7      global sysdate=c(current_date)
8      global path "C:\Users\milet\paper\Wage_premium" // Directory Unige
9      *global path=c(pwd) // Directory Baobab
10     cd $path
11     *
12     *
13     *
14     *
15     * Data Creation
16     do 1_make_data.do // loops over the raw data, extract variables, and make the final dataset
17     do 1_period_definition.do // selects the surveys that we need to define the various periods
18     do 2_labels.do // puts the labels in the final dataset
19     do 1_tradedata.do // gets us the imports of capital, R&D intensive and R&D un-intensive goods
20     *
21     *
22     * Compute the prices and quantities in terms of efficiency units
23     do 3_efficiency_units.do // Get the supply and the wages in efficiency units (needed for the descriptive
24     statistics)
25     do 3_efficiency_units_industry.do // efficiency units at the industry level (to get the elasticities of substitution)
26     *do 3_efficiency_units_young_old.do // efficiency units for young and old workers
27     do 3_panel_data.do // creates the dataset at the country*level with skill premium and skill supply
28     *
29     * Stylized Facts
30     do 4_share_of_labor_income.do // get the share of labor income of total income for each country
31     do 4_table_list_IncGroup_WrldRegion.do // get the list of countries
32     do 4_table_years_period.do // get the years used for each country in each period
33     do 4_wage_distribution_percentiles.do // Evolution of the 10th, 50th, 90th percentiles. All workers, women only, men only
34     do 4_wage_distribution_education_level.do // Evolution of the composition-adjusted wages by education level
35     do 4_skillpremium.do // Evolution of the composition-adjusted wages by education level
36     *
37     * Elasticities of substitution
38     do 5_elasticities.do // get the elasticities of substitution between skilled and unskilled workers
39     *
40     *
41     *
42     ** end of file

```

# The do's and don'ts of a do-file: LOGS

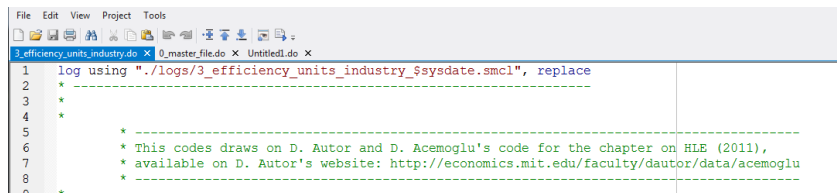
Log files are used to keep track of EVERYTHING your do-file does, and more:

- It saves what Stata erases in the result window.
- You should ALWAYS start your do-file by opening a log file, and end by closing it.
- Your log files should be stored in a specific log folder.
- Give the SAME name to your do-files and your logs.
- You can also add the date of the day in the log's name, so that you do not erase them each time you run your do-file.



# The do's and don'ts of a do-file: LOGS

Example:



```
File Edit View Project Tools
3_efficiency_units_industry.do x 0_master_file.do x Untitled1.do x
1 log using "./logs/3_efficiency_units_industry_$sysdate.smcl", replace
2 * -----
3 *
4 *
5 *
6 * This codes draws on D. Autor and D. Acemoglu's code for the chapter on HLE (2011),
7 * available on D. Autor's website: http://economics.mit.edu/faculty/dautor/data/acemoglu
8 * -----
9 *
```

# The do's and don'ts of a do-file: NAMES & LABELS

- 1 Give explicit names to your variables.
- 2 Adopt a naming rule and stick to it! For instance:
  - log variables can be named: `ln_`*varname*
  - dummy variables created from variables: *varname*\_dum
- 3 Attach labels to the variables.
- 4 Ideally, you should have a do-file with all the labels.
- 5 Labels will show up in any output that you produce.

```

3_panel_data.do X 3_efficiency_units_industry.do X 0_master.file.do X Untitled1.do X
106 gen rdoecd_incapoecd=lnm_rd_high_oecd-lnm_cap_high_oecd // share of R&D imports in in
107 *
108 * Log of FDI inflows
109 gen lnfdi=ln(fdi_inflow)
110 *
111 * put labels
112 label var relwage "Ln skill premium"
113 label var relsupply "Ln relative supply"
114 label var capttotal_intotal "Share of capital imports in total imports"
115 label var caphigh_inhigh "Share of capital imports in imports from high-income
116 label var capoecd_inoecd "Share of capital imports in imports from OECD countries"
117 label var rdttotal_incaptotal "Share of R&D imports in capital imports"
118 label var rdhigh_incaphigh "Share of R&D imports in imports from high-income countries"
119 label var rdoecd_incapoecd "Share of R&D imports in imports from OECD countries"
120 label var lnfdi "Ln FDI inflow"
121 label var year "Year"
122 label var coode "Country code"
123 label var skilled_worker "1=skilled worker"
124 label var lnwage_p "Ln wage"
125 label var supply "# workers"
126 label var lnsupply "Ln # workers"
127 label var nworker_y "# workers (sum of weights from the survey)"
128 label var fdi_inflow "FDI inflow"
129 label var lnm_all "Ln aggregate imports"
130 label var lnm_high "Ln imports from high income countries"
131 label var lnm_high_oecd "Ln imports from OECD countries"
132 label var lnm_cap_all "Ln capital imports"
133 label var lnm_cap_high "Ln capital imports from high income countries"
134 label var lnm_cap_high_oecd "Ln capital imports from OECD countries"
135 label var lnm_rd_all "Ln R&D imports"
136 label var lnm_rd_high "Ln R&D imports from high income countries"
137 label var lnm_rd_high_oecd "Ln R&D imports from OECD countries"
138 label var lnm_notrd_all "Ln capital non-R&D imports"
139 label var lnm_notrd_high "Ln capital non-R&D imports from high income countries"
140 label var lnm_notrd_high_oecd "Ln capital non-R&D imports from OECD countries"
141 *
142 *
143 order coode iso3 world_region inc_group_last skilled_worker supply lnsupply relwage r
144 compress

```

Filter variables here

Name	Label
landlocked	1 # landlocked
continent	Continent
year	Year
industry	
nworker_y	# workers (sum of weights from the survey)
m_all	Total imports (K USD)
m_high	Total imports from High-income countries (K USD)
m_high_oecd	Total imports from High-income OECD countries (K USD)
m_cap_all	Total capital imports (K USD)
m_cap_high	Capital imports from High-income countries (K USD)
m_cap_high_oecd	Capital imports from High-income OECD countries (K USD)
m_rd_all	Total R&D imports (K USD)
m_rd_high	R&D imports from High-income countries (K USD)
m_rd_high_oecd	R&D imports from High-income OECD countries (K USD)
m_notrd_all	Total R&D imports (K USD)
m_notrd_high	NOT R&D imports from High-income countries (K USD)
m_notrd_high_oecd	NOT R&D imports from High-income OECD countries (K USD)
fdi_inflow	FDI inflow
lnwage_p	Ln wage
lnm_all	Ln aggregate imports
lnm_high	Ln imports from high income countries
lnm_high_oecd	Ln imports from OECD countries
lnm_cap_all	Ln capital imports
lnm_cap_high	Ln capital imports from high income countries
lnm_cap_high_oecd	Ln capital imports from OECD countries
lnm_rd_all	Ln R&D imports
lnm_rd_high	Ln R&D imports from high income countries

Properties

# The do's and don'ts of a do-file: REGRESSION OUTPUT

Regression results are probably the most important output from your do-files.

It is important that they look nice!

- The “combo” `esttab/estout` commands is a good way to produce and export regression results.
- There is also `outreg2` (personally not a fan)

```
. do "C:\Users\milet\AppData\Local\Temp\STD00000000.tmp"
```

```
. esttab, compress nogaps label drop(c_dum*) starlevels(* 0.1 ** 0.05 *** 0.01) b(%5.3f) t(%5.3f)
```

	(1)	(2)
	c_y_beta	c_y_beta
Ln FDI inflow	0.002 (0.191)	0.005 (0.389)
Ln aggregate i~a	-0.046*** (-3.241)	-0.066*** (-3.635)
Share of capit~a	0.011 (0.382)	0.068 (1.384)
Share of R&D i~	0.065 (1.137)	0.036 (0.463)
Constant	1.571*** (7.655)	1.218*** (4.991)
Observations	238	232

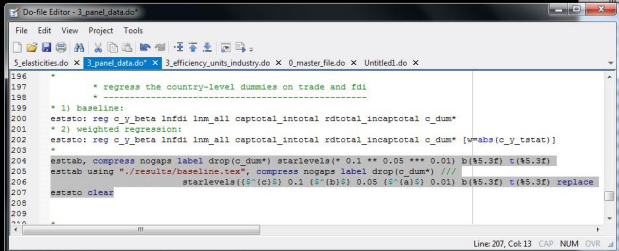
```
t statistics in parentheses  
* p<0.1, ** p<0.05, *** p<0.01
```

```
> . esttab using "./results/baseline.tex", compress nogaps label drop(c_dum*) ///
```

```
(output written to ./results/baseline.tex)
```

```
. eststo clear
```

```
. end of do-file
```



The screenshot shows a Do-file Editor window titled "Do-file Editor - 3\_panel\_data.do". The window contains the following Stata commands and output:

```
196 *  
197 * regress the country-level dummies on trade and fdi  
198 *  
199 * 1) baseline:  
200 eststo: reg c_y_beta lnfdi ln_all capttotal_intotal rdtotall_incapttotal c_dum*  
201 * 2) weighted regression:  
202 eststo: reg c_y_beta lnfdi ln_all capttotal_intotal rdtotall_incapttotal c_dum* [w=abs(c_y_tstat)]  
203 *  
204 esttab, compress nogaps label drop(c_dum*) starlevels(* 0.1 ** 0.05 *** 0.01) b(%5.3f) t(%5.3f)  
205 esttab using "./results/baseline.tex", compress nogaps label drop(c_dum*) ///  
206 starlevels({$^c[c]} 0.1 {$^b[b]} 0.05 {$^a[a]} 0.01) b(%5.3f) t(%5.3f) replace  
207  
208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220  
221  
222  
223  
224  
225  
226  
227  
228  
229  
230  
231  
232  
233  
234  
235  
236  
237  
238  
239  
240  
241  
242  
243  
244  
245  
246  
247  
248  
249  
250  
251  
252  
253  
254  
255  
256  
257  
258  
259  
260  
261  
262  
263  
264  
265  
266  
267  
268  
269  
270  
271  
272  
273  
274  
275  
276  
277  
278  
279  
280  
281  
282  
283  
284  
285  
286  
287  
288  
289  
290  
291  
292  
293  
294  
295  
296  
297  
298  
299  
300  
301  
302  
303  
304  
305  
306  
307  
308  
309  
310  
311  
312  
313  
314  
315  
316  
317  
318  
319  
320  
321  
322  
323  
324  
325  
326  
327  
328  
329  
330  
331  
332  
333  
334  
335  
336  
337  
338  
339  
340  
341  
342  
343  
344  
345  
346  
347  
348  
349  
350  
351  
352  
353  
354  
355  
356  
357  
358  
359  
360  
361  
362  
363  
364  
365  
366  
367  
368  
369  
370  
371  
372  
373  
374  
375  
376  
377  
378  
379  
380  
381  
382  
383  
384  
385  
386  
387  
388  
389  
390  
391  
392  
393  
394  
395  
396  
397  
398  
399  
400  
401  
402  
403  
404  
405  
406  
407  
408  
409  
410  
411  
412  
413  
414  
415  
416  
417  
418  
419  
420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000
```

# The do's and don'ts of a do-file: REGRESSION OUTPUT

```
{
\def\sym#1{\ifmmode^{#1}\else\(^{#1}\)\fi}
\begin{tabular}{l*{2}{c}}
\hline\hline
&&\multicolumn{1}{c}{(1)}&&\multicolumn{1}{c}{(2)}\\
&&\multicolumn{1}{c}{Dem. shifters}&&\multicolumn{1}{c}{Dem. shifters}
\hline
Ln FDI inflow & 0.002 & 0.005 & \\
& (0.191) & (0.389) & \\
Ln aggregate imports& 0.046{\mathstrut} & 0.066{\mathstrut} & \\
& (3.241) & (3.635) & \\
Share of capital imports in total imports& 0.011 & & 0.068 \\
\\
& (0.382) & (1.384) & \\
Share of R&D imports in capital imports& 0.065 & & 0.036 \\
\\
& (1.137) & (0.463) & \\
Constant & 1.571{\mathstrut} & 1.218{\mathstrut} & \\
& (7.655) & (4.991) & \\
\hline
Observations & 238 & 232 & \\
\hline\hline
\multicolumn{3}{l}{\footnotesize \textit{statistics in parentheses}}\\
\multicolumn{3}{l}{\footnotesize {\mathstrut} p<0.1, {\mathstrut} p<0.05, {\mathstrut} p<0.01}
\end{tabular}
}
```

# The do's and don'ts of a do-file: REGRESSION OUTPUT

	(1)	(2)
	Dem. shifters	Dem. shifters
Ln FDI inflow	0.002 (0.191)	0.005 (0.389)
Ln aggregate imports	-0.046 <sup>a</sup> (-3.241)	-0.066 <sup>a</sup> (-3.635)
Share of capital imports in total imports	0.011 (0.382)	0.068 (1.384)
Share of R&D imports in capital imports	0.065 (1.137)	0.036 (0.463)
Constant	1.571 <sup>a</sup> (7.655)	1.218 <sup>a</sup> (4.991)
Observations	238	232

*t* statistics in parentheses

<sup>c</sup>  $p < 0.1$ , <sup>b</sup>  $p < 0.05$ , <sup>a</sup>  $p < 0.01$

# The do's and don'ts of a do-file: USEFUL COMMANDS

The **COLLAPSE** command:

- It reduces the dimensionality of your dataset and calculates many statistics (count, median, mean, standard deviation, percentiles, first obs. ...) base on certain dimensions (useful when using panel data, or data with more than 2 dimensions:  $\text{firm} \times \text{product} \times \text{destination} \times \text{year}$  for instance)
- BEWARE: Weight normalization impacts only the sum, count, sd, semean, and sebinomial statistics
- A weighted average cannot be obtained directly from collapse, despite the fact that you can ask for the mean of a variable, and specify weights.



# The do's and don'ts of a do-file: USEFUL COMMANDS

The **TAG** command:

- It creates a dummy variable taking the value 1 for each occurrence of a variable.
- It is useful when you have nested dimensions (say individual/household/county/region), and want to quickly get statistics at a specific level.

# The do's and don'ts of a do-file: USEFUL COMMANDS

The **TAG** command:

- It creates a dummy variable taking the value 1 for each occurrence of a variable.
- It is useful when you have nested dimensions (say individual/household/county/region), and want to quickly get statistics at a specific level.

The **GROUP** command:

- Creates a variable taking integer values from 1 to  $N$  for each occurrence of a given variable or list of variables (i.e. an occurrence is therefore a combination of various variables).
- This is handy to create fixed-effect variables.
- Turns a string variable into a numeric variable (the command **encode** does this too, but only for 1 variable).

# The do's and don'ts of a do-file: USEFUL COMMANDS

The **LEVELSOF** command:

- Especially useful for loops
- IT lists all occurrences of a variable, and stores the list into a local variable.
- You can then loop over the elements of this list.
- The nice feature is that Stata does not create any variable, the elements of the list (i.e. the occurrences) are *local* elements.

```

. levelsof ccode, local(ccode_local)
`ARG' `BGD' `BLZ' `BOL' `BRA' `CHL'

. foreach c of local ccode_local{
  2. *
  . dis "`c'"
  3. }
ARG
BGD
BLZ
BOL
BRA
CHL

.
end of do-file

. do "C:\Users\milet\AppData\Local\Temp\S9
. levelsof year, local(year_local)
1993 1994 1995 1996 1997 1998 1999 2000

. foreach y of local year_local{
  2. *
  . dis `y'
  3. }
1993
1994
1995
1996
1997
1998
1999
2000

.
end of do-file

```

```

Do-file Editor - 0_list_surveys_raw_data.do
File Edit View Project Tools
0_master_file.do x 0_list_surveys_raw_data.do x
116 levelsof ccode, local(ccode_local)
117 foreach c of local ccode_local{
118     *
119     dis "`c'"
120 }
121 *
122 *
123 *
124 *
125 *
126 *
127 do "C:\Users\milet\AppData\Local\Temp\S9
128 *
129 levelsof year, local(year_local)
130 1993 1994 1995 1996 1997 1998 1999 2000
131 *
132 *
133 foreach y of local year_local{
134     *
135     dis `y'
136 }
137 *
138 *
139 *
140 levelsof year, local(year_local)
141 foreach y of local year_local{
142     *
143     dis `y'
144 }
145 *
146 *
147 *
Line: 136, Col: 1 CAP NUM OVR

```

The **#delimit** command:

- This is a line breaker.
- Essentially the same as using three forward slash bars: `///`
- It is more convenient though (personal opinion).

```

levelsof ccode, local(ccode_local)
foreach c of local ccode_local{
    preserve
    keep if ccode=="`c'"
        *
            * 1a) make the graph
            * -----
#delimit ;
twoway (scatter relsupply year, msymbol(Dh) mcolor(emerald))
      (line relsupply year, lcolor(emerald))
      (scatter relwage year, msymbol(Oh) mcolor(dkorange) yaxis(2))
      (line relwage year, lcolor(dkorange) yaxis(2)),
      scheme(sicolor) xtitle("") ytitle("") ytitle("", axis(2))
      legend(order(2 "Relative supply index (left axis)" 3 "Skill premium (right axis)"
      region(lpattern(blank)));
#delimit cr
graph export "../graphs/skillpremium_`c'.pdf", as(pdf) replace
*
*

```

- That's it for today.
- You can find a written version fo all this on my webpage:  
<http://emmanuelmilet.weebly.com/>
- Thank you for your attention