Create Latex table with xtable and stargazer

(Jing Li, Miami University)

1. This note introduces two R packages of **xtable** and **stargazer** that can generate latex or html codes of reporting statistics and regression in well-formatted tables. If you have never heard of latex, you may check out this

   https://www.overleaf.com/learn/latex/Learn_LaTeX_in_30_minutes

2. For illustration, I use the GaltonFamilies data in the HistData package

```
> library(HistData)
> data(GaltonFamilies)
> attach(GaltonFamilies)
> head(GaltonFamilies)
  family father mother midparentHeight children childNum gender childHeight
1    001   78.5   67.0           75.43        4        1   male        73.2
2    001   78.5   67.0           75.43        4        2 female        69.2
3    001   78.5   67.0           75.43        4        3 female        69.0
4    001   78.5   67.0           75.43        4        4 female        69.0
5    002   75.5   66.5           73.66        4        1   male        73.5
6    002   75.5   66.5           73.66        4        2   male        72.5
> str(GaltonFamilies)
'data.frame':   934 obs. of  8 variables:
$ family         : Factor w/ 205 levels "001","002","003",..: 1 1 1 1 2 2 2 2 3 3 .
$ father         : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
$ mother         : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
$ midparentHeight: num  75.4 75.4 75.4 75.4 73.7 ...
$ children       : int  4 4 4 4 4 4 4 4 2 2 ...
$ childNum       : int  1 2 3 4 1 2 3 4 1 2 ...
$ gender         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 2 1 ...
$ childHeight    : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
```

   Notice that

   (a) family is factor with 205 levels, one level for each family.

(b) gender is factor with level 2 being for male and 1 for female. It is NOT a dummy variable

3. We can obtain descriptive statistics with function **describe** in **psych** package

```
> library(psych)
> m0 = describe(GaltonFamilies,skew=F,ranges=F,quant=c(.25,.75))
> m0
                 vars   n   mean    sd   se Q0.25  Q0.75
family*             1 934 105.90 56.79 1.86 59.25 156.00
father              2 934  69.20  2.48 0.08 68.00  71.00
mother              3 934  64.09  2.29 0.07 63.00  65.88
midparentHeight     4 934  69.21  1.80 0.06 68.14  70.14
children            5 934   6.17  2.73 0.09  4.00   8.00
childNum            6 934   3.59  2.36 0.08  2.00   5.00
gender*             7 934   1.51  0.50 0.02  1.00   2.00
childHeight         8 934  66.75  3.58 0.12 64.00  69.70
```

R adds asterisk to factors, because the descriptive statistics of those variables are *misleading*.

4. Next I can generate the latex codes for a table using **xtable** package

```
> library(xtable)
> table = xtable(m0)
> print(table, include.rownames = T)
```

The **print** function displays latex codes that you can *copy and paste* into **latex editor** such as TeXstudio, TexShop, and Overleaf

5. I use TexShop, and the first latex table looks like

	vars	n	mean	sd	se	Q0.25	Q0.75
family*	1	934.00	105.90	56.79	1.86	59.25	156.00
father	2	934.00	69.20	2.48	0.08	68.00	71.00
mother	3	934.00	64.09	2.29	0.07	63.00	65.88
midparentHeight	4	934.00	69.21	1.80	0.06	68.14	70.14
children	5	934.00	6.17	2.73	0.09	4.00	8.00
childNum	6	934.00	3.59	2.36	0.08	2.00	5.00
gender*	7	934.00	1.51	0.50	0.02	1.00	2.00
childHeight	8	934.00	66.75	3.58	0.12	64.00	69.70

where I shade the cells you need to pay attention to. Shading is done in latex editor, not in R.

6. There are multiple ways to improve this table. For instance, I can drop family (since it is a factor), and I can re-create a dummy variable gender

```
> GaltonFamilies$gender = as.numeric(GaltonFamilies$gender)-1
> m0b = describe(GaltonFamilies[,-1],skew=F,ranges=F,quant=c(.25,.75))
> table = xtable(m0b)
> print(table, include.rownames = T)
```

The revised table looks like

	vars	n	mean	sd	se	Q0.25	Q0.75
father	1	934.00	69.20	2.48	0.08	68.00	71.00
mother	2	934.00	64.09	2.29	0.07	63.00	65.88
midparentHeight	3	934.00	69.21	1.80	0.06	68.14	70.14
children	4	934.00	6.17	2.73	0.09	4.00	8.00
childNum	5	934.00	3.59	2.36	0.08	2.00	5.00
gender	6	934.00	0.51	0.50	0.02	0.00	1.00
childHeight	7	934.00	66.75	3.58	0.12	64.00	69.70

Since gender is dummy variable, its mean 0.51 is sample proportion of males. You can make this table even better by dropping the useless vars column, and remove .00 for the sample size n column (since it is integer) *in your latex editor*.

7. Next I follow *specific-to-general* modeling principle, and try four specifications of OLS regressions

```
> m1 = lm(childHeight~father)
> m2 = lm(childHeight~father+mother)
> m3 = lm(childHeight~father+mother+gender)
> m4 = lm(childHeight~father+mother+gender+childNum)
> library(sandwich)
> library(lmtest)
> m4b = coeftest(m4, vcov = vcovCL, cluster = family)
```

   Notice that I use **coeftest** in order to obtain *cluster-robust standard error* for m4.

8. To generate the latex codes for a table, I use **stargazer** package

```
> library(stargazer)
> stargazer(m1, m2, m3, m4, m4b,
            type = "latex",
            title = "Regression Results",
            align = TRUE,
            column.labels = c("Model 1", "Model 2", "Model 3", "Model 4", "Model 4r"
            omit.stat = c("f","rsq","ser"),
            dep.var.caption = "Dependent Variable: y")
```

(Continued on next page)

9. The raw table looks like

Table 1: Regression Results

	Dependent Variable: y				
	childHeight				
	OLS				*coefficient test*
	Model 1	Model 2	Model 3	Model 4	Model 4r
	(1)	(2)	(3)	(4)	(5)
father	0.385***	0.368***	0.393***	0.354***	0.354***
	(0.046)	(0.045)	(0.029)	(0.027)	(0.039)
mother		0.291***	0.318***	0.306***	0.306***
		(0.049)	(0.031)	(0.029)	(0.049)
gender			5.215***	4.097***	4.097***
			(0.142)	(0.163)	(0.203)
childNum				−0.406***	−0.406***
				(0.035)	(0.049)
Constant	40.139***	22.643***	16.521***	21.992***	21.992***
	(3.160)	(4.262)	(2.727)	(2.590)	(4.311)
Observations	934	934	934	934	
Adjusted R^2	0.070	0.103	0.634	0.681	
Note:					*p<0.1; **p<0.05; ***p<0.01

(Continued on next page)

Again, *I can edit the table in the latex editor.* The revised table looks like

Table 2: Regression Results

	Dependent Variable: childHeight				
	(1)	(2)	(3)	(4)	(5)
father	0.385***	0.368***	0.393***	0.354***	0.354***
	(0.046)	(0.045)	(0.029)	(0.027)	(0.039)
mother		0.291***	0.318***	0.306***	0.306***
		(0.049)	(0.031)	(0.029)	(0.049)
gender			5.215***	4.097***	4.097***
			(0.142)	(0.163)	(0.203)
childNum				−0.406***	−0.406***
				(0.035)	(0.049)
Constant	40.139***	22.643***	16.521***	21.992***	21.992***
	(3.160)	(4.262)	(2.727)	(2.590)	(4.311)
n	934	934	934	934	934
Adj R^2	0.070	0.103	0.634	0.681	0.681
Note: significance					*p<0.1; **p<0.05; ***p<0.01

Can you detect the changes I have made?