Problem Set 5 Solution

17.842 Quantitative Research Methods
TA. Jiyoon Kim

Question 1. The data already has a weighting variable under the name of "weight." I believe that the variable use all different socio-economic or racial background to correct the sampling frame bias. For the following conditional density and regression problems, I picked question 10 (global warming concern), question 14b (policy response to global warming) and question 19 (religiosity). Their sample statistics follows.

```
. sum q10 q14b q19

    Variable |     Obs        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         q10 |    1205    2.633195   1.273246         -1          5
        q14b |    1205     .853112   3.142914         -2          7
         q19 |    1205    1.880498   .7110843         -1          3

. sum q10 q14b q19 [aw=weight]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         q10 |    1205   1205.1800    2.683956   1.299613         -1          5
        q14b |    1205   1205.1800    .8744669   3.104373         -2          7
         q19 |    1205   1205.1800    1.907383   .7080763         -1          3
```

Now, I created a table with joint density of q10 and q14b, and q10 and q19.

```
. tab2 q10 q19

-> tabulation of q10 by q19

           |                    Q19
       Q10 |        -1          1          2          3 |     Total
-----------+--------------------------------------------+----------
        -1 |         0          0          2          1 |         3
         1 |         1         49        113         48 |       211
         2 |         3        106        258         73 |       440
         3 |         3         95        158         38 |       294
         4 |         1         24         44         13 |        82
         5 |         3         45         92         35 |       175
-----------+--------------------------------------------+----------
     Total |        11        319        667        208 |      1205

. tab2 q14b q10

-> tabulation of q14b by q10

           |                         Q10
      Q14B |        -1          1          2          3          4 |     Total
-----------+---------------------------------------------------------+----------
        -2 |         1        112        221        142         47 |       606
        -1 |         0          0          1          2          0 |        15
         1 |         0          1          5          7          2 |        26
         2 |         0         27         62         39          5 |       147
         3 |         0         11         51         19          3 |        93
         4 |         0          7         24         24          4 |        63
         5 |         2         38         61         40          3 |       165
         6 |         0         15         15          9          2 |        51
         7 |         0          0          0         12         16 |        39
-----------+---------------------------------------------------------+----------
     Total |         3        211        440        294         82 |      1205
```

Q14B	Q10 5	Total
-2	83	606
-1	12	15
1	11	26
2	14	147
3	9	93
4	4	63
5	21	165
6	10	51
7	11	39
Total	175	1205

These tables only show frequencies. By dividing the sample size of 1205, you can have a probability tables.

Joint Density Table for q10 and q19

	Q19				
Q10	**-1**	**1**	**2**	**3**	**Total**
-1	0	0	0.002	0.001	0.002
1	0.001	0.041	0.094	0.040	0.175
2	0.002	0.088	0.214	0.061	0.365
3	0.002	0.079	0.131	0.032	0.244
4	0.001	0.020	0.037	0.011	0.068
5	0.002	0.037	0.076	0.029	0.145
Total	0.009	0.265	0.554	0.173	1

Joint Density Table for q10 and q14b

	Q10						
Q14B	**-1**	**1**	**2**	**3**	**4**	**5**	**Total**
-2	0.001	0.093	0.183	0.118	0.039	0.069	0.503
-1	0.000	0.000	0.001	0.002	0.000	0.010	0.012
1	0.000	0.001	0.004	0.006	0.002	0.009	0.022
2	0.000	0.022	0.051	0.032	0.004	0.012	0.122
3	0.000	0.009	0.042	0.016	0.002	0.007	0.077
4	0.000	0.006	0.020	0.020	0.003	0.003	0.052
5	0.002	0.032	0.051	0.033	0.002	0.017	0.137
6	0.000	0.012	0.012	0.007	0.002	0.008	0.042
7	0.000	0.000	0.000	0.010	0.013	0.009	0.032
Total	0.002	0.175	0.365	0.244	0.068	0.145	1.000

The conditional probabilities can be obtained by using the formula of $Pr(A|B) = Pr(A \& B)/Pr(B)$. For example, the conditional probability of policy response given different concerns about global warming is

	Q19		
Q10	1	2	3
-1	0	0.002996	0.004797
1	**0.153449**	0.169271	0.230254
2	0.33195	0.386476	0.350179
3	0.297503	0.236679	0.182285
4	0.075159	0.065911	0.062361
5	0.140922	0.137813	0.167894

	Q10					
Q14B	-1	1	2	3	4	5
-2	0.414938	0.53112	0.502473	0.48296	0.57359	0.475032
-1	0	0	0.002274	0.006802	0	0.068679
1	0	0.004742	0.011368	0.023808	0.024408	0.062956
2	0	0.128038	**0.140965**	0.132644	0.06102	0.080126
3	0	0.052164	0.115955	0.064621	0.036612	0.05151
4	0	0.033195	0.054567	0.081627	0.048816	0.022893
5	0.829876	0.180202	0.138692	0.136045	0.036612	0.120189
6	0	0.071132	0.034104	0.03061	0.024408	0.057233
7	0	0	0	0.040814	0.195265	0.062956

From the first table, we see that the conditional probability of the level of concern about global warming given a particular level of religiosity. For example, if the person is very religious (q19 = 1), the probability that the person says there is a serious global warming problem is .15 which is least among religiosity category. In the second table, it is a conditional probability of policies given the level of concern about global warming. This says that if a person believes that "there is enough evidence that global warming is taking place" (q10-2), the probability that the person thinks that "we should invest R&D"(q14b-2) is .141. The conditional probability table shows that as the concern of global warming increases (from 4 to 1), at least less probability for answering "do nothing(q14b-1).

In the regression result, you can get a little different outcome. From the conditional probability tables, we have some confidence that the more a person is religious, s/he does not really think that global warming is an imminent problem. However,

```
. reg q10 q19

      Source |       SS       df       MS              Number of obs =    1205
-------------+------------------------------           F(  1,  1203) =    3.51
       Model |  5.68303187     1  5.68303187           Prob > F      =  0.0611
    Residual |  1946.18917  1203  1.61777986           R-squared     =  0.0029
-------------+------------------------------           Adj R-squared =  0.0021
       Total |   1951.8722  1204  1.62115631           Root MSE      =  1.2719

------------------------------------------------------------------------------
         q10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         q19 |  -.0966175   .0515496    -1.87   0.061    -.1977546    .0045196
       _cons |   2.814884   .1036326    27.16   0.000     2.611563    3.018205
------------------------------------------------------------------------------
```

The regression result does not guarantee any statistically significant causation relationship between concern about global warming by religiosity. The same story goes for policy responses and global warming concern.

```
. reg q14b q10

      Source |       SS       df       MS              Number of obs =    1205
-------------+------------------------------           F(  1,  1203) =    0.11
       Model |  1.04095052     1  1.04095052           Prob > F      =  0.7456
    Residual |  11891.9599  1203  9.88525343           R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0007
       Total |  11893.0008  1204  9.87790767           Root MSE      =  3.1441

------------------------------------------------------------------------------
        q14b |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         q10 |   .0230935   .0711653     0.32   0.746    -.1165284    .1627154
       _cons |   .7923024   .2081329     3.81   0.000     .3839585    1.200646
------------------------------------------------------------------------------
```

Coding problem may have caused these results. As you have seen in the file, there are a huge number of "no opinion" or "I don't know" answers. Also, many of variables are not coded with ordinal concerns. Therefore, I transformed codings of concern about global warming question. "No opinion" and "refused" were coded as missing variables (not always recommended. There are enormous literatures about missing variables and how they can cause problems… anyway,) and saw the causal relationship between religiosity and global warming concern.

```
. reg q10 q19

      Source |       SS       df       MS              Number of obs =    1019
-------------+------------------------------           F(  1,  1017) =    8.84
       Model |  6.61127946     1  6.61127946           Prob > F      =  0.0030
    Residual |  760.440732  1017  .747729334           R-squared     =  0.0086
-------------+------------------------------           Adj R-squared =  0.0076
       Total |  767.052012  1018  .753489206           Root MSE      =  .86471

------------------------------------------------------------------------------
         q10 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         q19 |  -.1231698   .0414222    -2.97   0.003    -.2044526   -.0418869
       _cons |    2.47248   .0832298    29.71   0.000     2.309158    2.635801
------------------------------------------------------------------------------
```

The result is different from the first regression of q10 on q19. Now, we have a negative coefficient which is statistically significant. Therefore, as the person becomes more religious, s/he is less concerned about global warming problem.

Question 2. Please see attached do-file for the process I created variables. Summary statistics are :

Variable	Obs	Mean	Std. Dev.	Min	Max
x	100000	0.004265	1.002622	-4.51892	4.448323
y	100000	1.006364	1.076699	-3.46255	5.711002

Summary statistics for truncated data sets given upper 5 percent and lower 5 percent of each variable follow:

```
. sum y if x1 !=.

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |    5000    .1667309     1.01062  -3.127756   3.890636

. sum y if x2 !=.

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           y |    5000    1.843936    1.003058  -1.397682   5.711002


. sum x if y1 !=.

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |    5000   -.7615851    .9547275  -4.518918    2.23154

. sum x if y2 !=.

    Variable |     Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
           x |    5000    .7779928    .9445251   -2.47037   4.448323
```

As you see, once you have truncated data sets, you have now different means from original datasets.

Finally, if we regress y on x, we have the result of:

```
. reg y x

      Source |       SS       df       MS              Number of obs =  100000
-------------+------------------------------           F(  1, 99998) =16399.52
       Model |   16333.219     1   16333.219           Prob > F      =  0.0000
    Residual |  99593.7378 99998  .995957297           R-squared     =  0.1409
-------------+------------------------------           Adj R-squared =  0.1409
       Total |  115926.957 99999  1.15928116           Root MSE      =  .99798

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .4030887   .0031476   128.06   0.000     .3969194    .4092581
       _cons |   1.004645   .0031559   318.34   0.000     .9984597    1.010831
------------------------------------------------------------------------------
```

The coefficient of x is .4 and very close to our artificial coefficient for the creation of y variable. (yeah, of course.) The standard error is very small, but still have some value. That's because we inserted error term at the end of y variable equation which gave some noise in y and x relations.

The formula for correlation coefficient is Cov(x,y)/St.dv(x)St.dv(x), while the regression coefficient formula (in bivariate case) is Cov(x,y)/Var(x). Since the variance of x and y are very similar (almost same as one. Why? Think about the variable property of Var(a + bX) = b^2Var(X). Here, b is .4, therefore, the squared term is only .16, which results in .16 x 1 = .16. But, we have an error term at the end of y equation. Error term is also standard normal with 0

mean and 1 variance. Therefore, the variance of y is about .16 + 1 = .16, and standard deviation is 1.077. Still, standard deviations of two variables are pretty similar and we expect the correlation coefficient and regression coefficient will be similar, too. And they are.

```
. corr y x
(obs=100000)

             |        y        x
-------------+------------------
           y |   1.0000
           x |   0.3754   1.0000
```