```
                                                              ___  ____  ____  ____  ____(R)
                                                             /__    /   ____/   /   ____/
                                                            ___/   /   /___/   /   /___/
                                                              Statistics/Data Analysis

                                                                User: Jiahao Ye
                                                             Project: LAB2
-----------------------------------------------------------------------------------------------------
      name:  <unnamed>
       log:  D:\Econ4G03LAB2\output.smcl
  log type:  smcl
 opened on:   1 Oct 2020, 12:35:05

  1 .  set more off

  2 .
  3 .  use ${DtaDir}LFS-71M0001-E-2020-January_F1, clear

  4 .
  5 .
end of do-file

  6 . do "C:\Users\LOCAL_~3\Temp\STD00000000.tmp"

  7 . /*Generate a table for sex*/
  8 . tab SEX, m

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
       Male |     48,398       48.80       48.80
     Female |     50,777       51.20      100.00
------------+-----------------------------------
      Total |     99,175      100.00

  9 . tab SEX, m nolabel

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |     48,398       48.80       48.80
          2 |     50,777       51.20      100.00
------------+-----------------------------------
      Total |     99,175      100.00

 10 . label list SEX
SEX:
           1 Male
           2 Female

 11 .
 12 .
 13 . gen byte female=.  /* "Byte" stands for a float variable */
(99175 missing values generated)

 14 .  replace female=0 if SEX==1
(48398 real changes made)

 15 .  replace female=1 if SEX==2
(50777 real changes made)

 16 .
 17 . tab female SEX, missing /*Replacing variable for sex,and generate a table.*/

           |   Sex of respondent
    female |      Male     Female |     Total
-----------+----------------------+----------
         0 |    48,398          0 |    48,398
         1 |         0     50,777 |    50,777
-----------+----------------------+----------
     Total |    48,398     50,777 |    99,175
```

```
18 .
19 . /*Question ii):*/
20 . /*Generate indicator variables for four regions: east, qc, on, west */
21 . tab PROV, nolabel
```

Province	Freq.	Percent	Cum.
10	3,722	3.75	3.75
11	2,609	2.63	6.38
12	5,191	5.23	11.62
13	5,058	5.10	16.72
24	17,883	18.03	34.75
35	28,353	28.59	63.34
46	7,800	7.86	71.20
47	7,071	7.13	78.33
48	9,561	9.64	87.97
59	11,927	12.03	100.00
Total	99,175	100.00	

```
22 . label list PROV
   PROV:
             10 Newfoundland and Labrador
             11 Prince Edward Island
             12 Nova Scotia
             13 New Brunswick
             24 Quebec
             35 Ontario
             46 Manitoba
             47 Saskatchewan
             48 Alberta
             59 British Columbia

23 .
24 . /*Here, I change the province(east west QC ON) into a categorical variable */
25 . label list PROV
   PROV:
             10 Newfoundland and Labrador
             11 Prince Edward Island
             12 Nova Scotia
             13 New Brunswick
             24 Quebec
             35 Ontario
             46 Manitoba
             47 Saskatchewan
             48 Alberta
             59 British Columbia

26 . gen east=1 if PROV <=14
   (82595 missing values generated)

27 . replace east=0 if PROV >14
   (82595 real changes made)

28 . tab east
```

east	Freq.	Percent	Cum.
0	82,595	83.28	83.28
1	16,580	16.72	100.00
Total	99,175	100.00	

```
29 .
30 . gen west=1 if PROV >=46
   (62816 missing values generated)

31 . replace west=0 if PROV <46
   (62816 real changes made)

32 . tab west
```

west	Freq.	Percent	Cum.
0	62,816	63.34	63.34
1	36,359	36.66	100.00
Total	99,175	100.00	

```
33 .
34 . gen on=1 if PROV == 35
   (70822 missing values generated)

35 . replace on=0 if PROV != 35
   (70822 real changes made)

36 . tab on
```

on	Freq.	Percent	Cum.
0	70,822	71.41	71.41
1	28,353	28.59	100.00
Total	99,175	100.00	

```
37 .
38 . gen qc=1 if PROV == 24
   (81292 missing values generated)

39 . replace qc=0 if PROV != 24
   (81292 real changes made)

40 . tab qc
```

qc	Freq.	Percent	Cum.
0	81,292	81.97	81.97
1	17,883	18.03	100.00
Total	99,175	100.00	

```
41 .
42 . /*Question iii):create a new age variable*/
43 .
44 . tab AGE_12
```

Five-year age group of respondent	Freq.	Percent	Cum.
15 to 19 years	6,608	6.66	6.66
20 to 24 years	6,309	6.36	13.02
25 to 29 years	6,980	7.04	20.06
30 to 34 years	7,594	7.66	27.72
35 to 39 years	7,912	7.98	35.70
40 to 44 years	7,630	7.69	43.39
45 to 49 years	7,545	7.61	51.00
50 to 54 years	7,986	8.05	59.05
55 to 59 years	9,215	9.29	68.34
60 to 64 years	8,616	8.69	77.03
65 to 69 years	7,674	7.74	84.77
70 and over	15,106	15.23	100.00
Total	99,175	100.00	

```
45 . label list AGE_12
   AGE_12:
              1 15 to 19 years
              2 20 to 24 years
              3 25 to 29 years
              4 30 to 34 years
              5 35 to 39 years
              6 40 to 44 years
              7 45 to 49 years
              8 50 to 54 years
              9 55 to 59 years
             10 60 to 64 years
             11 65 to 69 years
             12 70 and over

46 .
47 . gen age=17 if AGE_12==1
   (92567 missing values generated)

48 . replace age=22 if AGE_12==2
   (6309 real changes made)

49 . replace age=27 if AGE_12==3
   (6980 real changes made)

50 . replace age=32 if AGE_12==4
   (7594 real changes made)

51 . replace age=37 if AGE_12==5
   (7912 real changes made)

52 . replace age=42 if AGE_12==6
   (7630 real changes made)

53 . replace age=47 if AGE_12==7
   (7545 real changes made)

54 . replace age=52 if AGE_12==8
   (7986 real changes made)

55 . replace age=57 if AGE_12==9
   (9215 real changes made)

56 . replace age=62 if AGE_12==10
   (8616 real changes made)

57 . replace age=67 if AGE_12==11
   (7674 real changes made)

58 . replace age=72 if AGE_12==12
   (15106 real changes made)

59 .
60 . tab age
```

age	Freq.	Percent	Cum.
17	6,608	6.66	6.66
22	6,309	6.36	13.02
27	6,980	7.04	20.06
32	7,594	7.66	27.72
37	7,912	7.98	35.70
42	7,630	7.69	43.39
47	7,545	7.61	51.00
52	7,986	8.05	59.05
57	9,215	9.29	68.34
62	8,616	8.69	77.03
67	7,674	7.74	84.77
72	15,106	15.23	100.00
Total	99,175	100.00	

```
61 .
62 . /*Question iv use LFSSTAT): Exam "HRLYEARN", create varianles=0 if missing values
   >                                                                        variables=1 i
   >                                                                        variabels="mi
   > */
63 . tab LFSSTAT
```

Labour force status	Freq.	Percent	Cum.
Employed, at work	53,557	54.00	54.00
Employed, absent from work	4,660	4.70	58.70
Unemployed	3,962	3.99	62.70
Not in labour force	36,996	37.30	100.00
Total	99,175	100.00	

```
64 . label list LFSSTAT
   LFSSTAT:
              1 Employed, at work
              2 Employed, absent from work
              3 Unemployed
              4 Not in labour force

65 . gen status=0 if LFSSTAT==4  | LFSSTAT==2       /*missing value, either not in labor force or
   (57519 missing values generated)

66 . replace status=1 if LFSSTAT==1   /*Non-missing value */
   (53557 real changes made)

67 . replace status=. if LFSSTAT==3   /*unemployed*/
   (0 real changes made)

68 .
69 . tab status
```

status	Freq.	Percent	Cum.
0	41,656	43.75	43.75
1	53,557	56.25	100.00
Total	95,213	100.00	

```
70 .
71 . /*In this LAB, we will be focusing on non-missing samples, which is status = 1. */
72 .
73 . /*Question 2): Run regression analysis*/
74 . /* i)reg HRLYEARN age female */
75 . regress HRLYEARN age female
```

Source	SS	df	MS
Model	569503.468	2	284751.734
Residual	8996755.24	49261	182.634442
Total	9566258.71	49263	194.187498

Number of obs	= 49264
F(2, 49261)	= 1559.13
Prob > F	= 0.0000
R-squared	= 0.0595
Adj R-squared	= 0.0595
Root MSE	= 13.514

HRLYEARN	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.2045397	.00438	46.70	0.000	.1959549	.2131245
female	-3.690772	.1217903	-30.30	0.000	-3.929482	-3.452061
_cons	21.26753	.2017506	105.41	0.000	20.87209	21.66296

```
76 .
77 . /*Explanation: Regression line: HRLYEARN = 21.26753-3.690772*female+0.2045397*age
   > Equation shows female has less HRLYEARN than male does(When female=1,negative slope, which de
   > holding age constant).
   > HRLYEARN increases, on average, by 0.2045397 for unit increase in age group (5-year increase
78 .
79 . /* ii)repeat above, but with robust (reweighted OLS, allowing heteroskedasticity) this time *
80 . regress HRLYEARN age female, robust

   Linear regression                                      Number of obs =    49264
                                                          F(  2, 49261) = 1627.57
                                                          Prob > F      =  0.0000
                                                          R-squared     =  0.0595
                                                          Root MSE      =  13.514

   ------------------------------------------------------------------------------
               |               Robust
      HRLYEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   ------------+----------------------------------------------------------------
           age |   .2045397   .0041641    49.12   0.000      .196378    .2127014
        female |  -3.690772   .1219684   -30.26   0.000    -3.929831   -3.451712
         _cons |   21.26753    .178266   119.30   0.000     20.91812    21.61693
   ------------------------------------------------------------------------------

81 .
82 . /*Robust OLS: Very similar to above, which means outliers do not have significant impact */
83 .
84 . /* iii)Regression Y=ln(HRLYEARN), X=Age,female */
85 . gen lnearn = ln(HRLYEARN)
   (49911 missing values generated)

86 . regress lnearn age female

         Source |       SS       df       MS              Number of obs =    49264
   -------------+------------------------------           F(  2, 49261) = 1848.26
          Model |  748.578845     2  374.289423           Prob > F      =  0.0000
       Residual |  9975.82495 49261  .202509591           R-squared     =  0.0698
   -------------+------------------------------           Adj R-squared =  0.0698
          Total |  10724.4038 49263  .217696929           Root MSE      =  .45001

   ------------------------------------------------------------------------------
         lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -------------+----------------------------------------------------------------
            age |   .0075549   .0001458    51.80   0.000     .0072691    .0078408
         female |  -.1276835   .0040555   -31.48   0.000    -.1356323   -.1197346
          _cons |   2.968137   .0067181   441.81   0.000     2.954969    2.981304
   ------------------------------------------------------------------------------

87 .
88 . /*age coefficient=0.0075549, female coefficient=-0.12768
   > An unit increase in age will have a 0.0075549 increase in ln(HRLYEARN).*/
89 .
90 . /* iv)Regression Y=HRLYEARN, X=ln(age) female */
91 . gen lnage = ln(age)

92 . regress HRLYEARN lnage female

         Source |       SS       df       MS              Number of obs =    49264
   -------------+------------------------------           F(  2, 49261) = 2193.71
          Model |  782336.527     2  391168.264           Prob > F      =  0.0000
       Residual |  8783922.18 49261  178.313923           R-squared     =  0.0818
   -------------+------------------------------           Adj R-squared =  0.0817
          Total |  9566258.71 49263  194.187498           Root MSE      =  13.353

   ------------------------------------------------------------------------------
       HRLYEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
   -------------+----------------------------------------------------------------
          lnage |   9.566883   .1634185    58.54   0.000     9.246581    9.887185
         female |  -3.694531   .1203398   -30.70   0.000    -3.930398   -3.458663
          _cons |  -5.282225   .6049176    -8.73   0.000    -6.467871   -4.096579
   ------------------------------------------------------------------------------
```

```
 93 .
 94 . /* v) Regression Y=ln(HRLYEARN), x=ln(age) female */
 95 . regres lnearn lnage female

      Source |       SS       df       MS              Number of obs =   49264
-------------+------------------------------           F(  2, 49261) = 2741.23
       Model |  1074.03141     2  537.015705           Prob > F      =  0.0000
    Residual |  9650.37238 49261  .195902892           R-squared     =  0.1001
-------------+------------------------------           Adj R-squared =  0.1001
       Total |  10724.4038 49263  .217696929           Root MSE      =  .44261

------------------------------------------------------------------------------
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lnage |   .3607236   .0054166    66.60   0.000      .350107    .3713403
      female |  -.1277958   .0039888   -32.04   0.000    -.1356138   -.1199778
       _cons |   1.960524   .0200504    97.78   0.000     1.921224    1.999823
------------------------------------------------------------------------------

 96 .
 97 . /*Explanation: While holding female constant, a unit increase in ln(age) will have
   > a 0.3607236 increase in ln(HRLYEARN).
   > The R^2 (Explanatory power) increased, as I applied logorithmic. Log function has a
   > advantage of transferring skewed data into linear relationship, which will give us better
   > result for linear regression analysis. */
 98 .
 99 . /* vi)Factor variables, Three types of variables: Categ. variables, indicator variables, cont
100 . regress lnearn c.age i.female c.age#i.female, robust

Linear regression                                      Number of obs =   49264
                                                       F(  3, 49260) = 1188.23
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.0710
                                                       Root MSE      =  .44973

------------------------------------------------------------------------------
             |               Robust
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0087097   .0002181    39.93   0.000     .0082821    .0091372
    1.female |  -.0317036   .0124616    -2.54   0.011    -.0561284   -.0072787
             |
female#c.age |
           1 |  -.0023131   .0002979    -7.77   0.000     -.002897   -.0017293
             |
       _cons |   2.920114   .0091378   319.56   0.000     2.902204    2.938024
------------------------------------------------------------------------------

101 .
102 . /*"#"stands for interaction variables "##"means a factorial of interaction.
   > Here, we are trying to model it like: HRLYEARN=b0+b1*age+b2*female+b3*age*female(interaction
   > Results here shows that, as age for women increases, HRLYEARN tends to decrease.
   > But for men, it is less of an issue  */
103 .
104 . /*  vii) */
105 . regress lnearn c.age c.age#c.age female,robust

Linear regression                                      Number of obs =   49264
                                                       F(  3, 49260) = 5314.23
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1789
                                                       Root MSE      =   .4228
```

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0725511	.0007419	97.79	0.000	.0710969	.0740052
c.age#c.age	-.0007754	9.15e-06	-84.69	0.000	-.0007933	-.0007574
female	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
_cons	1.758419	.0137591	127.80	0.000	1.731451	1.785387

```
106 . margins, dydx(*)at(age=(25(5)60))

Average marginal effects                        Number of obs     =     49264
Model VCE    : Robust

Expression   : Linear prediction, predict()
dy/dx w.r.t. : age female

1._at        : age             =          25

2._at        : age             =          30

3._at        : age             =          35

4._at        : age             =          40

5._at        : age             =          45

6._at        : age             =          50

7._at        : age             =          55

8._at        : age             =          60
```

	dy/dx	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
age						
_at						
1	.0337827	.0003006	112.38	0.000	.0331935	.0343719
2	.026029	.0002205	118.02	0.000	.0255968	.0264613
3	.0182754	.0001539	118.76	0.000	.0179738	.018577
4	.0105217	.0001244	84.57	0.000	.0102778	.0107656
5	.002768	.0001551	17.85	0.000	.0024641	.0030719
6	-.0049856	.0002222	-22.44	0.000	-.0054212	-.0045501
7	-.0127393	.0003024	-42.12	0.000	-.0133321	-.0121466
8	-.020493	.0003877	-52.86	0.000	-.0212529	-.0197331
female						
_at						
1	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
2	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
3	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
4	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
5	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
6	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
7	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449
8	-.132411	.0038092	-34.76	0.000	-.139877	-.1249449

```
107 . marginsplot

      Variables that uniquely identify margins: age _deriv

108 . /* c.age#c.age means age squared. Regression line:
    > HRLYEARN = 0.0725511*age - 0.0007754*age^2 - 0.132411*female + 1.758419,
    > take the partial derivative w.r.t age*/
109 .
110 . margins, at(age=(25(5)65)) /*This commands calculate the exact prediction value */

    Predictive margins                              Number of obs     =     49264
    Model VCE    : Robust

    Expression   : Linear prediction, predict()

    1._at        : age             =          25

    2._at        : age             =          30

    3._at        : age             =          35

    4._at        : age             =          40

    5._at        : age             =          45

    6._at        : age             =          50

    7._at        : age             =          55

    8._at        : age             =          60

    9._at        : age             =          65

    ------------------------------------------------------------------------------
                 |            Delta-method
                 |     Margin   Std. Err.      t    P>|t|     [95% Conf. Interval]
    -------------+----------------------------------------------------------------
             _at |
              1  |   3.020418   .0024066  1255.06   0.000     3.015701    3.025135
              2  |   3.169947     .00234  1354.69   0.000     3.165361    3.174533
              3  |   3.280708   .0025835  1269.88   0.000     3.275644    3.285772
              4  |   3.352701   .0027416  1222.92   0.000     3.347327    3.358074
              5  |   3.385925   .0027177  1245.87   0.000     3.380598    3.391252
              6  |   3.380381   .0026318  1284.44   0.000     3.375223    3.385539
              7  |   3.336068   .0028349  1176.79   0.000     3.330512    3.341625
              8  |   3.252988   .0037247   873.36   0.000     3.245687    3.260288
              9  |   3.131139   .0053713   582.94   0.000     3.120611    3.141666
    ------------------------------------------------------------------------------

111 . marginsplot

      Variables that uniquely identify margins: age

112 .
113 . /* viii) */
114 . margins, at(age=(25(5)65))by(female)

    Predictive margins                              Number of obs     =     49264
    Model VCE    : Robust

    Expression   : Linear prediction, predict()
    over         : female

    1._at        : 0.female
                       age             =          25
                   1.female
                       age             =          25
```

```
2._at        : 0.female
                   age          =          30
               1.female
                   age          =          30

3._at        : 0.female
                   age          =          35
               1.female
                   age          =          35

4._at        : 0.female
                   age          =          40
               1.female
                   age          =          40

5._at        : 0.female
                   age          =          45
               1.female
                   age          =          45

6._at        : 0.female
                   age          =          50
               1.female
                   age          =          50

7._at        : 0.female
                   age          =          55
               1.female
                   age          =          55

8._at        : 0.female
                   age          =          60
               1.female
                   age          =          60

9._at        : 0.female
                   age          =          65
               1.female
                   age          =          65
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at#female						
1 0	3.087591	.0031353	984.78	0.000	3.081446	3.093736
1 1	2.95518	.0030031	984.05	0.000	2.949294	2.961066
2 0	3.23712	.0030565	1059.11	0.000	3.231129	3.243111
2 1	3.104709	.0029781	1042.50	0.000	3.098872	3.110546
3 0	3.347881	.0032282	1037.07	0.000	3.341554	3.354208
3 1	3.21547	.0031913	1007.58	0.000	3.209215	3.221725
4 0	3.419874	.0033461	1022.06	0.000	3.413315	3.426432
4 1	3.287463	.0033303	987.13	0.000	3.280935	3.29399
5 0	3.453098	.0033244	1038.72	0.000	3.446582	3.459614
5 1	3.320687	.0033129	1002.36	0.000	3.314194	3.32718
6 0	3.447554	.0032603	1057.42	0.000	3.441164	3.453944
6 1	3.315143	.003237	1024.13	0.000	3.308798	3.321488
7 0	3.403242	.0034395	989.45	0.000	3.3965	3.409983
7 1	3.270831	.0033913	964.48	0.000	3.264184	3.277478
8 0	3.320161	.0042202	786.73	0.000	3.311889	3.328433
8 1	3.18775	.0041471	768.66	0.000	3.179621	3.195878
9 0	3.198312	.0057431	556.90	0.000	3.187055	3.209568
9 1	3.065901	.0056556	542.10	0.000	3.054816	3.076986

```
115 . marginsplot

      Variables that uniquely identify margins: age female

116 . /*From the result, we can show that, for both men&women, the margin starts to decline
    > at around age group 5~6.*/
117 .
118 . /*Then, run another regression, for DIFFERENT INTERCEPT & DIFFERENT SLOPE
    > for men&women w.r.t how earnings change with age.
    > Approach:   1)HRLYEARN = b1*age + b2*age*female +b3
    >                       2)HRLYEARN = b1*age + b2*age^2+ b3*age*female +b4
    >         refer to textbook page 259, fig8.8 */
119 .
120 . regress HRLYEARN c.age c.age#i.female, robust

    Linear regression                                      Number of obs =    49264
                                                           F(  2, 49261) = 1522.48
                                                           Prob > F      =  0.0000
                                                           R-squared     =  0.0612
                                                           Root MSE      =  13.503

    ------------------------------------------------------------------------------
                 |               Robust
        HRLYEARN |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
    -------------+----------------------------------------------------------------
             age |   .2485544   .0045477    54.66   0.000     .2396409    .2574679
                 |
    female#c.age |
              1  |  -.0881511   .0029879   -29.50   0.000    -.0940075   -.0822948
                 |
           _cons |   19.42031   .1659044   117.06   0.000     19.09513    19.74548
    ------------------------------------------------------------------------------

121 . margins, at(age=(25(5)65))by(female)

    Predictive margins                                Number of obs   =      49264
    Model VCE    : Robust

    Expression   : Linear prediction, predict()
    over         : female

    1._at        : 0.female
                       age             =          25
                   1.female
                       age             =          25

    2._at        : 0.female
                       age             =          30
                   1.female
                       age             =          30

    3._at        : 0.female
                       age             =          35
                   1.female
                       age             =          35

    4._at        : 0.female
                       age             =          40
                   1.female
                       age             =          40

    5._at        : 0.female
                       age             =          45
                   1.female
                       age             =          45

    6._at        : 0.female
                       age             =          50
                   1.female
                       age             =          50
```

```
7._at        : 0.female
                   age              =          55
               1.female
                   age              =          55

8._at        : 0.female
                   age              =          60
               1.female
                   age              =          60

9._at        : 0.female
                   age              =          65
               1.female
                   age              =          65
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at#female						
1 0	25.63417	.0868602	295.12	0.000	25.46392	25.80442
1 1	23.43039	.0856394	273.59	0.000	23.26254	23.59824
2 0	26.87694	.0815535	329.56	0.000	26.71709	27.03679
2 1	24.23241	.077802	311.46	0.000	24.07991	24.3849
3 0	28.11971	.0824095	341.22	0.000	27.95819	28.28124
3 1	25.03442	.0754403	331.84	0.000	24.88656	25.18229
4 0	29.36248	.0892509	328.99	0.000	29.18755	29.53742
4 1	25.83644	.0790466	326.85	0.000	25.68151	25.99137
5 0	30.60526	.1008673	303.42	0.000	30.40756	30.80296
5 1	26.63846	.0878892	303.09	0.000	26.46619	26.81072
6 0	31.84803	.1158307	274.95	0.000	31.621	32.07506
6 1	27.44047	.1005967	272.78	0.000	27.2433	27.63764
7 0	33.0908	.1330164	248.77	0.000	32.83009	33.35151
7 1	28.24249	.1159047	243.67	0.000	28.01531	28.46966
8 0	34.33357	.1516708	226.37	0.000	34.0363	34.63085
8 1	29.0445	.1329178	218.51	0.000	28.78398	29.30503
9 0	35.57634	.1713148	207.67	0.000	35.24057	35.91212
9 1	29.84652	.1510609	197.58	0.000	29.55044	30.1426

```
122 . marginsplot

      Variables that uniquely identify margins: age female

123 .
124 . regress HRLYEARN c.age c.age##c.age c.age#i.female
    note: age omitted because of collinearity
```

Source	SS	df	MS
Model	1341264.67	3	447088.224
Residual	8224994.04	49260	166.971052
Total	9566258.71	49263	194.187498

Number of obs	=	49264
F(3, 49260)	=	2677.64
Prob > F	=	0.0000
R-squared	=	0.1402
Adj R-squared	=	0.1402
Root MSE	=	12.922

HRLYEARN	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.903573	.0249796	76.20	0.000	1.854613	1.952533
age	0	(omitted)				
c.age#c.age	-.0197163	.000293	-67.30	0.000	-.0202905	-.019142
female#c.age						
1	-.09275	.002662	-34.84	0.000	-.0979675	-.0875326
_cons	-11.39969	.493259	-23.11	0.000	-12.36648	-10.43289

```
125 . margins, at(age=(25(5)65))by(female)

    Predictive margins                              Number of obs   =      49264
    Model VCE    : OLS

    Expression   : Linear prediction, predict()
    over         : female

    1._at        : 0.female
                       age             =          25
                   1.female
                       age             =          25

    2._at        : 0.female
                       age             =          30
                   1.female
                       age             =          30

    3._at        : 0.female
                       age             =          35
                   1.female
                       age             =          35

    4._at        : 0.female
                       age             =          40
                   1.female
                       age             =          40

    5._at        : 0.female
                       age             =          45
                   1.female
                       age             =          45

    6._at        : 0.female
                       age             =          50
                   1.female
                       age             =          50

    7._at        : 0.female
                       age             =          55
                   1.female
                       age             =          55

    8._at        : 0.female
                       age             =          60
                   1.female
                       age             =          60

    9._at        : 0.female
                       age             =          65
                   1.female
                       age             =          65
```

	Margin	Delta-method Std. Err.	t	P>|t|	[95% Conf.	Interval]
_at#female						
1 0	23.86698	.1001691	238.27	0.000	23.67065	24.06331
1 1	21.54823	.0998627	215.78	0.000	21.3525	21.74396
2 0	27.96288	.0874138	319.89	0.000	27.79154	28.13421
2 1	25.18037	.0861247	292.37	0.000	25.01157	25.34918
3 0	31.07296	.0911292	340.98	0.000	30.89435	31.25157
3 1	27.82671	.0890978	312.32	0.000	27.65208	28.00134
4 0	33.19723	.0979049	339.08	0.000	33.00534	33.38913
4 1	29.48723	.0955347	308.65	0.000	29.29998	29.67448
5 0	34.33569	.101649	337.79	0.000	34.13646	34.53492
5 1	30.16194	.0993063	303.73	0.000	29.9673	30.35658
6 0	34.48833	.1033548	333.69	0.000	34.28576	34.69091
6 1	29.85083	.1015344	294.00	0.000	29.65182	30.04984
7 0	33.65517	.1098092	306.49	0.000	33.43994	33.8704
7 1	28.55392	.1091926	261.50	0.000	28.3399	28.76793

```
         8 0  |   31.83619   .1310385   242.95   0.000     31.57935    32.09303
         8 1  |   26.27119   .1320762   198.91   0.000     26.01232    26.53006
         9 0  |    29.0314   .1732618   167.56   0.000      28.6918    29.37099
         9 1  |   23.00264   .1757867   130.86   0.000      22.6581    23.34719
------------------------------------------------------------------------------

126 . marginsplot

      Variables that uniquely identify margins: age female

127 .
128 . /*As we can see from plottin the graph, the wage gap enlarges b/w men&women,
    > as age increases. From the above two regression analysis result, the coefficients
    > in front of age##female are both negative, which indicates firmly, that there
    > exists a gender inequality in terms of workplace earning.*/
129 .
130 .
131 . /*Q3: Run the following regression */
132 .
133 .
134 . regress lnearn age female east qc west, robust

Linear regression                               Number of obs =     49264
                                                F(  5, 49258) =    920.17
                                                Prob > F      =    0.0000
                                                R-squared     =    0.0864
                                                Root MSE      =    .44598

------------------------------------------------------------------------------
             |               Robust
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0076787   .0001478    51.95   0.000      .007389    .0079684
      female |  -.1249803   .0040202   -31.09   0.000    -.1328599   -.1171006
        east |   -.173054   .0064204   -26.95   0.000    -.1856381   -.1604699
          qc |  -.0595095   .0058914   -10.10   0.000    -.0710567   -.0479624
        west |  -.0058784   .0050138    -1.17   0.241    -.0157055    .0039488
       _cons |   3.000941   .0072677   412.91   0.000     2.986696    3.015185
------------------------------------------------------------------------------

135 . /*i) ON is being omitted because of collinearity. If ON and QC both exist in the model,
    > STATA will omit QC since ON & QC are highly corelated.  */
136 .
137 . /*ii) Whats the difference b/w men and female, conditonal on other variables?*/
138 . regress lnearn age female, robust

Linear regression                               Number of obs =     49264
                                                F(  2, 49261) =   1820.69
                                                Prob > F      =    0.0000
                                                R-squared     =    0.0698
                                                Root MSE      =    .45001

------------------------------------------------------------------------------
             |               Robust
      lnearn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .0075549    .000149    50.69   0.000     .0072628     .007847
      female |  -.1276835   .0040574   -31.47   0.000    -.1356361   -.1197308
       _cons |   2.968137   .0066289   447.76   0.000     2.955144     2.98113
------------------------------------------------------------------------------
```

```
139 .
140 . /*As can be see from the result, -.1290945 is the difference */
141 .
142 . /*iii)Repeat above, without conditional this time */
143 . regress lnearn age female east qc west, robust

Linear regression                               Number of obs =   49264
                                                F(  5, 49258) =  920.17
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0864
                                                Root MSE      =  .44598
```

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.0076787	.0001478	51.95	0.000	.007389	.0079684
female	-.1249803	.0040202	-31.09	0.000	-.1328599	-.1171006
east	-.173054	.0064204	-26.95	0.000	-.1856381	-.1604699
qc	-.0595095	.0058914	-10.10	0.000	-.0710567	-.0479624
west	-.0058784	.0050138	-1.17	0.241	-.0157055	.0039488
_cons	3.000941	.0072677	412.91	0.000	2.986696	3.015185

```
144 . /*Without conditioning, it wil be -.12498 */
145 .
146 . /*iv)Compare east and west */
147 . regress lnearn west, robust

Linear regression                               Number of obs =   49264
                                                F(  1, 49262) =  141.69
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0029
                                                Root MSE      =  .46591
```

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
west	.0517451	.0043471	11.90	0.000	.0432248	.0602654
_cons	3.197281	.0026444	1209.08	0.000	3.192098	3.202464

```
148 . regress lnearn east, robust

Linear regression                               Number of obs =   49264
                                                F(  1, 49262) =  672.24
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0133
                                                Root MSE      =  .46346
```

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
east	-.1507873	.0058157	-25.93	0.000	-.1621862	-.1393885
_cons	3.239501	.0022675	1428.65	0.000	3.235057	3.243945

```
149 . /*As a ran the above two regressions, results show that "west" has a positive coeff
    > while "east" has a negative coeff. Conditonal on other variables, in general, "west"
    > has a higher HRLYEARN than "east" */
```

```
150 .
151 . /*v)Compare east and Ontario */
152 . regress lnearn on, robust
```

Linear regression

Number of obs	=	**49264**
F(1, 49262)	=	**150.42**
Prob > F	=	**0.0000**
R-squared	=	**0.0030**
Root MSE	=	**.46588**

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
on	**.0568566**	**.0046358**	**12.26**	**0.000**	**.0477704**	**.0659428**
_cons	**3.200587**	**.0024873**	**1286.79**	**0.000**	**3.195712**	**3.205462**

```
153 . regress lnearn east, robust
```

Linear regression

Number of obs	=	**49264**
F(1, 49262)	=	**672.24**
Prob > F	=	**0.0000**
R-squared	=	**0.0133**
Root MSE	=	**.46346**

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
east	**-.1507873**	**.0058157**	**-25.93**	**0.000**	**-.1621862**	**-.1393885**
_cons	**3.239501**	**.0022675**	**1428.65**	**0.000**	**3.235057**	**3.243945**

```
154 .
155 . /*vi)change qc to on instead */
156 . regress lnearn age female east on west, robust
```

Linear regression

Number of obs	=	**49264**
F(5, 49258)	=	**920.17**
Prob > F	=	**0.0000**
R-squared	=	**0.0864**
Root MSE	=	**.44598**

lnearn	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	**.0076787**	**.0001478**	**51.95**	**0.000**	**.007389**	**.0079684**
female	**-.1249803**	**.0040202**	**-31.09**	**0.000**	**-.1328599**	**-.1171006**
east	**-.1135445**	**.0068888**	**-16.48**	**0.000**	**-.1270467**	**-.1000423**
on	**.0595095**	**.0058914**	**10.10**	**0.000**	**.0479624**	**.0710567**
west	**.0536312**	**.0055987**	**9.58**	**0.000**	**.0426576**	**.0646048**
_cons	**2.941431**	**.007601**	**386.98**	**0.000**	**2.926533**	**2.956329**

```
157 . /*The results show that, coeff on east is -.1135. If I put "on" instead of "qc",
    > STATA will omit none of the result, which shows that Ontario earning is not highly
    > correlated with other area's earning. */
158 .
```

```
159 . /*Section#2 Omitted variable bias */
160 . clear

161 . capture drop _all

162 . set obs 10000
    obs was 0, now 10000

163 . set seed 2000

164 . more

165 .
166 . gen schl = round(10 + 10*runiform(), 1) /*Generate a variable called schl */

167 . gen detrm = runiform()*10

168 .
169 . more

170 .
171 . tab schl
```

schl	Freq.	Percent	Cum.
10	514	5.14	5.14
11	1,006	10.06	15.20
12	953	9.53	24.73
13	1,043	10.43	35.16
14	1,054	10.54	45.70
15	954	9.54	55.24
16	948	9.48	64.72
17	1,055	10.55	75.27
18	1,008	10.08	85.35
19	984	9.84	95.19
20	481	4.81	100.00
Total	10,000	100.00	

```
172 . sum detrm, d

                            detrm
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .0973219       .0000191
 5%     .4872551        .000447
10%     .9971442       .0009003       Obs               10000
25%     2.504195       .0010381       Sum of Wgt.       10000

50%     4.985089                      Mean           5.015467
                        Largest       Std. Dev.       2.89706
75%     7.551486       9.997013
90%      9.00843       9.997205       Variance       8.392955
95%     9.538775       9.997805       Skewness        -.00307
99%     9.925169       9.999406       Kurtosis       1.802995

173 .
174 . gen wage = 5 + 0.1*schl + 0.2*detrm + invnorm(runiform())

175 . /*generate function called wage */
```

```
176 .
177 . list in 1/10

     +--------------------------------+
     | schl     detrm          wage   |
     |--------------------------------|
  1. |   11   1.031281      6.34586   |
  2. |   10   8.695457     9.396244   |
  3. |   11   4.699914     6.620628   |
  4. |   11   .3991777     4.479694   |
  5. |   10   6.511797     9.842613   |
     |--------------------------------|
  6. |   11   9.959957      8.58425   |
  7. |   15   4.428554     7.546886   |
  8. |   18   4.721651     8.374249   |
  9. |   12   6.707707     7.594052   |
 10. |   14   .4922629     6.409592   |
     +--------------------------------+

178 . sum wage, d

                            wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     4.777252       3.202093
 5%     5.523408       3.318614
10%     5.957373       3.411666       Obs               10000
25%      6.69338       3.464624       Sum of Wgt.       10000

50%     7.496738                      Mean           7.513497
                        Largest       Std. Dev.      1.203672
75%     8.331977       11.33636
90%     9.055006       11.35213       Variance       1.448826
95%     9.512343       11.36492       Skewness       .0322359
99%      10.3262       11.86834       Kurtosis       2.898942

179 .
180 . corr wage schl detrm  /*Check correlation*/
(obs=10000)

             |     wage     schl    detrm
-------------+---------------------------
        wage |   1.0000
        schl |   0.2501   1.0000
       detrm |   0.4815   0.0091   1.0000


181 . regress wage schl detrm

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) = 2063.73
       Model |  4233.34566     2  2116.67283           Prob > F      =  0.0000
    Residual |  10253.4634  9997  1.02565403           R-squared     =  0.2922
-------------+------------------------------           Adj R-squared =  0.2921
       Total |   14486.809  9999  1.44882579           Root MSE      =  1.0127

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        schl |   .1015889   .0034788    29.20   0.000     .0947697    .1084082
       detrm |   .1991217   .0034961    56.96   0.000     .1922687    .2059748
       _cons |   4.992702   .0557689    89.52   0.000     4.883384    5.102021
------------------------------------------------------------------------------
```

```
182 . /*Determination is not easy to observe, so we omit detrm */
183 . regress wage schl

      Source |       SS           df       MS      Number of obs   =    10000
-------------+----------------------------------   F(1, 9998)      =   667.14
       Model |  906.195669         1  906.195669   Prob > F        =   0.0000
    Residual |  13580.6134     9,998    1.358333   R-squared       =   0.0626
-------------+----------------------------------   Adj R-squared   =   0.0625
       Total |   14486.809     9,999  1.44882579   Root MSE        =   1.1655

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        schl |   .1034018   .0040033    25.83   0.000     .0955545    .1112491
       _cons |   5.964228   .0611035    97.61   0.000     5.844453    6.084003
------------------------------------------------------------------------------

184 . /*Coeff on schooling becomes larger. And, as we see from the corr calculation above,
    >  corr(wage,detrm)=0.4815, corr(detrm,schl)=0.0091 (Which is closed to zero correlation),
    >  therefore there are no omitted variable bias, corr()!=0 not satisfied*/
185 .
186 .  /* 5)Coded by me,
    >  create a new variable schl2, with variable 1,random number drawn from distribution
    >                                                                                2, al
    >                          variable wage2, with variable 1,schl2
    >                                                                                2, de
    >                                                                                3, in
187 . gen schl2 = round(10 + 10*runiform(), 1) + detrm

188 . list in 1/10

     +-------------------------------------------+
     | schl      detrm       wage       schl2   |
     |-------------------------------------------|
  1. |   11   1.031281    6.34586    12.03128   |
  2. |   10   8.695457   9.396244    25.69546   |
  3. |   11   4.699914   6.620628    15.69991   |
  4. |   11   .3991777   4.479694    11.39918   |
  5. |   10   6.511797   9.842613     24.5118   |
     |-------------------------------------------|
  6. |   11   9.959957    8.58425    26.95996   |
  7. |   15   4.428554   7.546886    20.42855   |
  8. |   18   4.721651   8.374249    21.72165   |
  9. |   12   6.707707   7.594052    20.70771   |
 10. |   14   .4922629   6.409592    17.49226   |
     +-------------------------------------------+

189 .
190 .
191 . gen wage2 = schl2 + detrm + invnorm(runiform())

192 . sum wage2, d

                            wage2
-------------------------------------------------------------
      Percentiles      Smallest
 1%     11.67679       8.849176
 5%     14.30382       8.862227
10%     16.22426       9.409176       Obs              10000
25%     19.91388         9.4366       Sum of Wgt.      10000

50%     24.98862                      Mean          25.02529
                        Largest       Std. Dev.     6.596099
75%     30.03234        41.0708
90%      33.9467       41.30936       Variance      43.50852
95%     35.76719       41.55243       Skewness       .016538
99%     38.51094       41.72418       Kurtosis      2.230368
```

```
193 .
194 . corr wage2 schl2 detrm  /*Check correlation*/
    (obs=10000)

                 |   wage2    schl2    detrm
    -------------+---------------------------
           wage2 |  1.0000
           schl2 |  0.9389   1.0000
           detrm |  0.8833   0.7086   1.0000


195 . /*result shows that these three variables are highly correlated,
    > with corr(wage2,schl2)=0.938, corr(wage2,detrm)=0.8839, corr(schl2,detrm)=0.7070*/
196 .
197 . regress wage2 schl2 detrm

          Source |       SS       df       MS              Number of obs =   10000
    -------------+------------------------------           F(  2,  9997) =       .
           Model |  425044.404     2  212522.202           Prob > F      =  0.0000
        Residual |  9997.31888  9997   1.0000319           R-squared     =  0.9770
    -------------+------------------------------           Adj R-squared =  0.9770
           Total |  435041.723  9999  43.5085232           Root MSE      =       1

    ------------------------------------------------------------------------------
           wage2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
    -------------+----------------------------------------------------------------
           schl2 |   1.003955   .0034316   292.56   0.000     .9972284    1.010682
           detrm |   .9969656   .0048923   203.78   0.000     .9873758    1.006555
           _cons |   -.061455    .055031    -1.12   0.264    -.1693269    .0464168
    ------------------------------------------------------------------------------

198 . /*Estimated regression line: wage2 = 0.9972522*schl + 1.00*detrm + 0.044 */
199 .
200 . regress wage2 schl2

          Source |       SS       df       MS              Number of obs =   10000
    -------------+------------------------------           F(  1,  9998) =74415.76
           Model |  383515.208     1  383515.208           Prob > F      =  0.0000
        Residual |  51526.5154  9998  5.15368228           R-squared     =  0.8816
    -------------+------------------------------           Adj R-squared =  0.8815
           Total |  435041.723  9999  43.5085232           Root MSE      =  2.2702

    ------------------------------------------------------------------------------
           wage2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
    -------------+----------------------------------------------------------------
           schl2 |   1.499488   .0054968   272.79   0.000     1.488714    1.510263
           _cons |  -4.975527   .1122953   -44.31   0.000    -5.195648   -4.755405
    ------------------------------------------------------------------------------

201 . /*wage2 = 1.49416*schl2 - 4.8838 */
202 .
203 . /*Omitted variable bias: check the following two conditions:
    >         1) A determinant of Y
    >         2) Correalted with another covariate , cov(X,Z)!=0
    >         In this case, detrm is indeed an omitted variable, and a1 is a biased estimator
    >         a1 --> true a1 + biase term */
204 .
205 .
206 . /* 6)Another coded by me:
    >         create something so that: schl&detrm are correlated,
    >                                                             but NO CAUSAL RELATIONSHIP
    > */
```

```
207 . matrix C  = (1,.25\.25, 1) /*Backward slash to seperate rows, this is a 2x2 matrix*/

208 . drawnorm y1 y2, n($obs) corr(C)

209 . /*Here, create a matrix with corr(y1,y2) = 0.25 */
210 .
211 . gen schl3 = round(10 + y1*2,1)

212 . gen detrm3 = 20 + 5*y2

213 . drop y1 y2

214 .
215 . gen wage3 = schl3 + detrm3 + invnorm(runiform())

216 .
217 . regress wage3 schl3 detrm3

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  2,  9997) =       .
       Model |  344537.985     2  172268.992           Prob > F      =  0.0000
    Residual |  10026.8542  9997  1.00298631           R-squared     =  0.9717
-------------+------------------------------           Adj R-squared =  0.9717
       Total |  354564.839  9999  35.4600299           Root MSE      =  1.0015

------------------------------------------------------------------------------
       wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       schl3 |   1.006409    .005091   197.68   0.000     .9964298    1.016389
      detrm3 |   1.001966   .0020523   488.22   0.000     .9979434    1.005989
       _cons |  -.0890153   .0581162    -1.53   0.126    -.2029348    .0249043
------------------------------------------------------------------------------

218 . /* wage3 = 1.0061*schl3 + 1.001966*detrm3-0.0890153 */
219 . regress wage3 schl3

      Source |       SS       df       MS              Number of obs =   10000
-------------+------------------------------           F(  1,  9998) = 4233.02
       Model |  105465.377     1  105465.377           Prob > F      =  0.0000
    Residual |  249099.462  9998  24.9149292           R-squared     =  0.2975
-------------+------------------------------           Adj R-squared =  0.2974
       Total |  354564.839  9999  35.4600299           Root MSE      =  4.9915

------------------------------------------------------------------------------
       wage3 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       schl3 |   1.602663    .024633    65.06   0.000     1.554378    1.650949
       _cons |   13.90172    .251993    55.17   0.000     13.40776    14.39568
------------------------------------------------------------------------------

220 . /*wage3 = 1.602*schl3 + 13.918 */
221 .
222 . corr wage3 schl3 detrm3
(obs=10000)

             |    wage3    schl3   detrm3
-------------+---------------------------
       wage3 |   1.0000
       schl3 |   0.5454   1.0000
      detrm3 |   0.9280   0.2399   1.0000
```

```
223 . /*corr(wage3,schl3)=0.555, corr(wage3,detrm3)=0.9259, corr(schl3,detrm3)=0.2461.
    > In this case, determination has a higher correlation than schooling does */
224 .
    end of do-file

225 . log close
          name:  <unnamed>
           log:  D:\Econ4G03LAB2\output.smcl
      log type:  smcl
     closed on:   1 Oct 2020, 12:36:48
```