UL Summer School: Regression session 2

Brendan Halpin, Sociology
2023 Summer School

Outline

Session 2

Session 2

Outline

Outline

- Multiple regression: more than 1 explanatory variable
- Estimate net effects of each variable, controlling for the others
- Very important class of statistical model
- Begin by considering 3-way relationships in the abstract
- Then consider the mechanics of multiple regression

Session 2

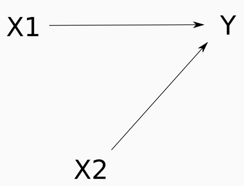
Multidimensional causality

Multidimensional causality

- Regression analysis never proves causal relationships, but it "thinks" in causal terms
- To use it we need to understand causal relationships: what process generates the data we see, and what can regression tell us about it.
- Start by considering the relationship between variables and patterns of association

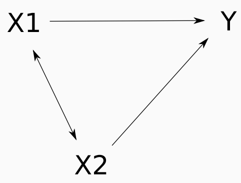
3-variable pictures

- Let's consider patterns of causality and association between three variables, X1 and X2, and Y
- If X1 and X2 are not correlated with each other, their separate effects on Y more or less just add up



Correlated X variables

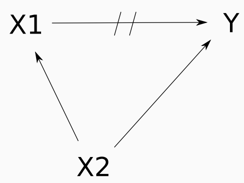
- But if X1 and X2 are correlated, things can get funny:



- In particular, if we measure the effect of one X without taking account of the other we will likely over-estimate it

Spurious association

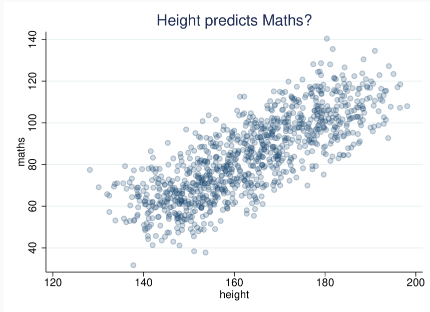
- X1 may have an association with Y, implying a causal relationship
- But if X2 affects both X1 and Y the relationship between X1 and Y may be spurious



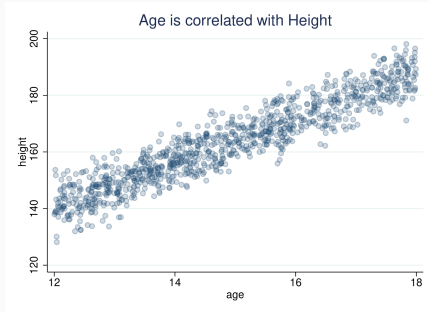
Spurious association: Maths and height

- (Artificial) example: students in secondary school are given a standardised maths test
- And their height is measured
- A strong correlation between height (X1) and test score (Y): a causal relationship?

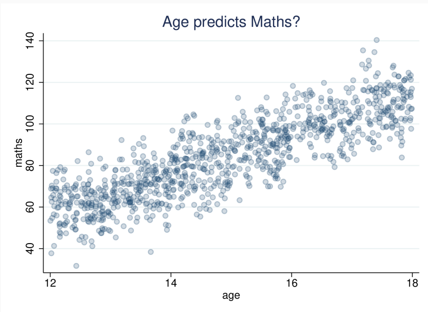
Maths and height: X1 -> Y?



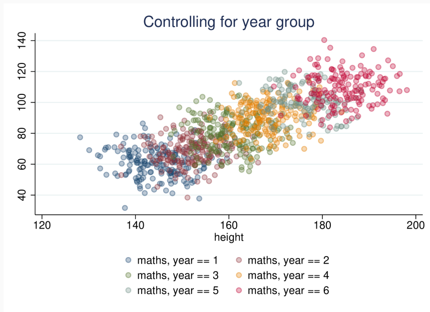
Maths and height: X1 <-> X2



Maths and height: X2 -> Y

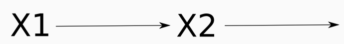


Maths and height: control for year group



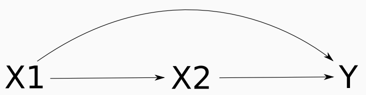
Indirect effects

- Where there is a time-order (X1 before X2), we may see direct and indirect effects
- X1 may affect X2, which affects Y, but not affect Y directly
- Thus there is association between X1 and Y without a direct causal effect



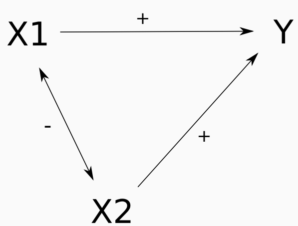
Direct and indirect effects

- However, it is possible for both direct and indirect effects to be present at the same time



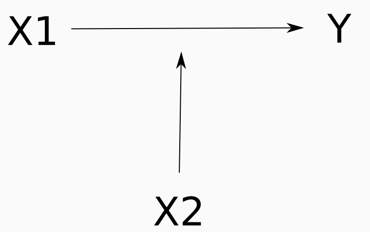
Suppression

- Where X1 and X2 have positive effects on Y, but a negative correlation, or different effects on Y with a positive correlation, the association between X1 and Y may be supressed
- That is, it may be invisible if we don't take account of X2



Interactions

- An interaction effect is where the effect of one variable on Y changes depending on the value of another



Session 2

Multiple regression

Multiple explanatory variables

- Regression analysis can be extended to the case where there is more than one explanatory variable – multiple regression
- This allows us to estimate the net simultaneous effect of many variables, and thus to begin to disentangle more complex relationships
- Interpretation is relatively easy: each variable gets its own slope coefficient, standard error and significance
- The slope coefficient is the effect on the dependent variable of a 1 unit change in the explanatory variable, *while taking account of the other variables*

Unpicking multiple effects

- We will see how regression can be used to throw light on the 3-variable problems we have described above
  - Over-estimation of X1's effect
  - Spurious X1
  - X1 with an indirect (mediated) effect
  - Under-estimation of X1's effect (suppression)
  - X1's effect differing according to the values of X2 (interaction)

Example: Over-estimation

- Example: income may be affected by gender, and also by work hours: competing explanations – one or the other, or both could have effects
- We can fit bivariate regressions:

$$Income = a + b \times WorkTime$$

or

$$Income = a + b \times Female$$

- We can also fit a single multivariate regression

$$Income = a + b \times WorkTime + c \times Female$$

Aside: Dichotomous variables

- We deal with gender in a special way: this is a *binary* or *dichotomous* variable – has two values
- We turn it into a yes/no or 0/1 variable – *e.g.*, female or not
- If we put this in as an explanatory variable a *one-unit change in the explanatory variable* is the difference between being male and female
- Thus the *c* coefficient we get in the $Income = a + b \times WorkTime + c \times Female$ regression is the net change in predicted income of females, once you take account of work hours.
- The *b* coefficient is then the net effect of a unit change in work hours, once you take gender into account.

3-variable Logic

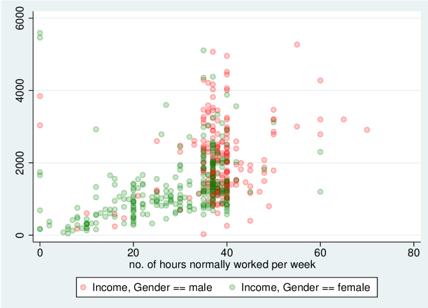
- X1 (hours) is correlated with income (higher H, higher I)
- X2 (gender) affects income (females lower)
- Hours and gender are strongly associated (females lower)

Income, hours and gender

```
. corr Income Gender Hours
(obs=506)
```

	Income	Gender	Hours
Income	1.0000		
Gender	-0.3280	1.0000	
Hours	0.3638	-0.4360	1.0000

Income, hours and gender



T-test: Income by gender

```
. ttest Income, by(Gender)
Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
male	437	1618.348	59.11677	1235.809	1502.159	1734.537
female	531	992.1805	40.82127	940.6625	911.9892	1072.372
combined	968	1274.861	36.23219	1127.281	1203.759	1345.964
diff		626.1674	70.00484		488.7883	763.5465

```
    diff = mean(male) - mean(female)                          t =   8.9446
Ho: diff = 0                                 degrees of freedom =      966

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Regression: Just hours

```
. reg Income Hours

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(1, 504)       =     76.86
       Model |  86947928.8         1  86947928.8   Prob > F        =    0.0000
    Residual |   570128215       504  1131206.78   R-squared       =    0.1323
-------------+----------------------------------   Adj R-squared   =    0.1306
       Total |   657076144       505  1301140.88   Root MSE        =    1063.6

------------------------------------------------------------------------------
      Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Hours |   37.82204   4.314061     8.77   0.000     29.34628     46.2978
       _cons |   449.7435   150.1722     2.99   0.003      154.703    744.7841
------------------------------------------------------------------------------
```


Regression: Hours and binary gender

```
. reg Income Hours i.Gender

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(2, 503)       =     50.70
       Model |   110236231         2  55118115.6   Prob > F        =    0.0000
    Residual |   546839912       503  1087156.88   R-squared       =    0.1678
-------------+----------------------------------   Adj R-squared   =    0.1645
       Total |   657076144       505  1301140.88   Root MSE        =    1042.7

------------------------------------------------------------------------------
      Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Hours |   28.33857   4.699451     6.03   0.000      19.1056    37.57155
             |
      Gender |
     female  |  -478.4214   103.3684    -4.63   0.000    -681.5084   -275.3344
       _cons |   1022.139   192.2717     5.32   0.000     644.3844    1399.893
------------------------------------------------------------------------------
```


3-var logic

- The gender gap reduces (but not to zero) if you control for hours
- The effect of hours controlling for gender falls

Spurious relationship

- Sometimes controlling for X2 makes the effect of X1 entirely disappear
- X1 -> Y is a "spurious" relationship

Maths and height by regression

```
. reg maths height

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =   1706.40
       Model |  235991.871         1  235991.871   Prob > F        =    0.0000
    Residual |  138021.727       998  138.298324   R-squared       =    0.6310
-------------+----------------------------------   Adj R-squared   =    0.6306
       Total |  374013.599       999  374.387987   Root MSE        =     11.76

------------------------------------------------------------------------------
       maths | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      height |   1.058213   .0256173    41.31   0.000     1.007943    1.108483
       _cons |  -89.11602   4.200327   -21.22   0.000     -97.3585   -80.87353
------------------------------------------------------------------------------
```


Spurious relationship: controlled for

```
. reg maths height age

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =   1273.62
       Model |   268802.74         2   134401.37   Prob > F        =    0.0000
    Residual |  105210.858       997  105.527441   R-squared       =    0.7187
-------------+----------------------------------   Adj R-squared   =    0.7181
       Total |  374013.599       999  374.387987   Root MSE        =    10.273

------------------------------------------------------------------------------
       maths | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
      height |  -.0067167   .0644065    -0.10   0.917    -.1331045    .1196711
         age |   9.579467   .5432693    17.63   0.000     8.513385    10.64555
       _cons |  -57.88381   4.074241   -14.21   0.000    -65.87888   -49.88874
------------------------------------------------------------------------------
```


Regression controls for linear effects

- We have seen this spurious relationship debunked visually
  - by separating into 6 year groups (subsetting the sample)
- Regression does it by attributing an effect to age
- Accounting for age strips the effect of height
- Regression can be more efficient than subsetting the sample
  - if the effect is linear, additive.

Regression: Direct and indirect 1

```
. reg ownscore fatherscore

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =     53.50
       Model |  13269.3853         1  13269.3853   Prob > F        =    0.0000
    Residual |  247525.861       998  248.021905   R-squared       =    0.0509
-------------+----------------------------------   Adj R-squared   =    0.0499
       Total |  260795.247       999  261.056303   Root MSE        =    15.749

------------------------------------------------------------------------------
    ownscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 fatherscore |   .2370829    .032413     7.31   0.000     .1734773    .3006884
       _cons |   37.90861   1.672157    22.67   0.000     34.62726    41.18996
------------------------------------------------------------------------------
```


Regression: Direct and indirect 2

```
. reg education fatherscore

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    111.01
       Model |  311.104929         1  311.104929   Prob > F        =    0.0000
    Residual |  2797.00607       998  2.80261129   R-squared       =    0.1001
-------------+----------------------------------   Adj R-squared   =    0.0992
       Total |    3108.111       999  3.11122222   Root MSE        =    1.6741

------------------------------------------------------------------------------
   education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
 fatherscore |   .0363018   .0034455    10.54   0.000     .0295405    .0430631
       _cons |   1.295213   .1777516     7.29   0.000     .9464035    1.644023
------------------------------------------------------------------------------
```

Regression: Direct and indirect 3

```
. reg ownscore education

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =    447.54
       Model |  80742.8091         1  80742.8091   Prob > F        =    0.0000
    Residual |  180052.437       998  180.413264   R-squared       =    0.3096
-------------+----------------------------------   Adj R-squared   =    0.3089
       Total |  260795.247       999  261.056303   Root MSE        =    13.432

------------------------------------------------------------------------------
    ownscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   5.096871   .2409273    21.16   0.000     4.624089    5.569653
       _cons |   33.87079   .8556481    39.58   0.000     32.19171    35.54986
------------------------------------------------------------------------------
```


Regression: Direct and indirect 4

```
. reg ownscore education fatherscore

      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(2, 997)       =    226.41
       Model |  81453.7212         2  40726.8606   Prob > F        =    0.0000
    Residual |  179341.525       997  179.881169   R-squared       =    0.3123
-------------+----------------------------------   Adj R-squared   =    0.3109
       Total |  260795.247       999  261.056303   Root MSE        =    13.412

------------------------------------------------------------------------------
    ownscore |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   4.937369   .2535982    19.47   0.000     4.439722    5.435017
 fatherscore |   .0578475   .0290984     1.99   0.047     .0007463    .1149486
       _cons |   31.51367   1.461439    21.56   0.000     28.64582    34.38152
------------------------------------------------------------------------------
```


Interaction

- Where the effect of X1 changes across values of X2, we have "interaction"

Regression: for men only

```
. reg Income Hours if Gender==1

      Source |       SS           df       MS      Number of obs   =       232
-------------+----------------------------------   F(1, 230)       =      5.36
       Model |  8009519.02         1  8009519.02   Prob > F        =    0.0215
    Residual |   343845612       230  1494980.92   R-squared       =    0.0228
-------------+----------------------------------   Adj R-squared   =    0.0185
       Total |   351855131       231  1523182.38   Root MSE        =    1222.7

------------------------------------------------------------------------------
      Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Hours |   24.61855   10.63597     2.31   0.022     3.662162    45.57495
       _cons |   1164.366   414.4901     2.81   0.005     347.6826    1981.049
------------------------------------------------------------------------------
```


Regression: for women only

```
. reg Income Hours if Gender==2

      Source |       SS           df       MS      Number of obs   =       274
-------------+----------------------------------   F(1, 272)       =     42.63
       Model |  31772944.2         1  31772944.2   Prob > F        =    0.0000
    Residual |   202744304       272  745383.469   R-squared       =    0.1355
-------------+----------------------------------   Adj R-squared   =    0.1323
       Total |   234517248       273  859037.537   Root MSE        =    863.36

------------------------------------------------------------------------------
      Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Hours |   29.70376   4.549594     6.53   0.000     20.74687    38.66065
       _cons |   504.6153   140.3614     3.60   0.000     228.2824    780.9482
------------------------------------------------------------------------------
```


Regression: interaction

```
. reg Income c.Hours##i.Gender

      Source |       SS           df       MS      Number of obs   =       506
-------------+----------------------------------   F(3, 502)       =     33.82
       Model |   110486228         3  36828742.8   Prob > F        =    0.0000
    Residual |   546589915       502  1088824.53   R-squared       =    0.1681
-------------+----------------------------------   Adj R-squared   =    0.1632
       Total |   657076144       505  1301140.88   Root MSE        =    1043.5

--------------------------------------------------------------------------------
        Income |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
         Hours |   24.61855   9.076915     2.71   0.007     6.785132    42.45198
               |
        Gender |
       female  |  -659.7502   392.3082    -1.68   0.093    -1430.518    111.0181
               |
Gender#c.Hours |
       female  |   5.085207   10.61255     0.48   0.632    -15.76529     25.9357
               |
         _cons |   1164.366   353.7327     3.29   0.001     469.3865    1859.345
--------------------------------------------------------------------------------
```