# SO5041 Unit 16: Regression

Brendan Halpin, Sociology

Autumn 2023/4

# SO5041 Unit 16

## Outline

# Outline

- Moving on from the idea of correlation: relating two (or more) interval/ratio variables
- Identifying the line that best summarises the scatterplot
- Directional: X predicts Y
- Predictive: Given the relationship between X and Y, knowing X helps us predict Y better
- Bivariate regression: one X variable predicting one Y
- Multiple regression: multiple X variables predicting one Y
- Reading: Agresti Ch 9

# SO5041 Unit 16

## Bivariate Linear Regression

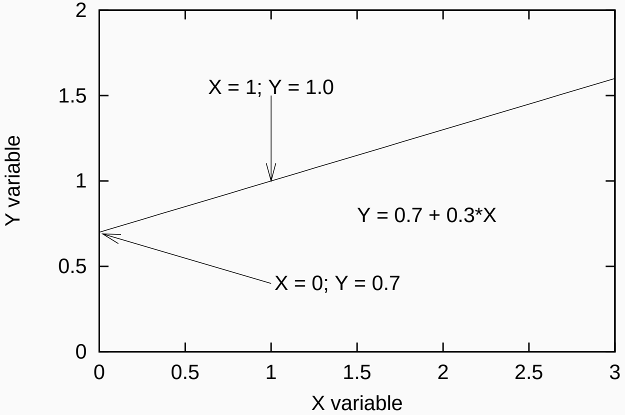
# Bivariate regression analysis

- Regression Analysis: Fitting the “best” line through the scatter
- Very closely related to correlation, but treats one variable as dependent and the other(s) as explanatory, while correlation is asymmetric

sociology
UNIVERSITY OF LIMERICK

# Some geometry: equation of a line

$$Y = a + bX$$



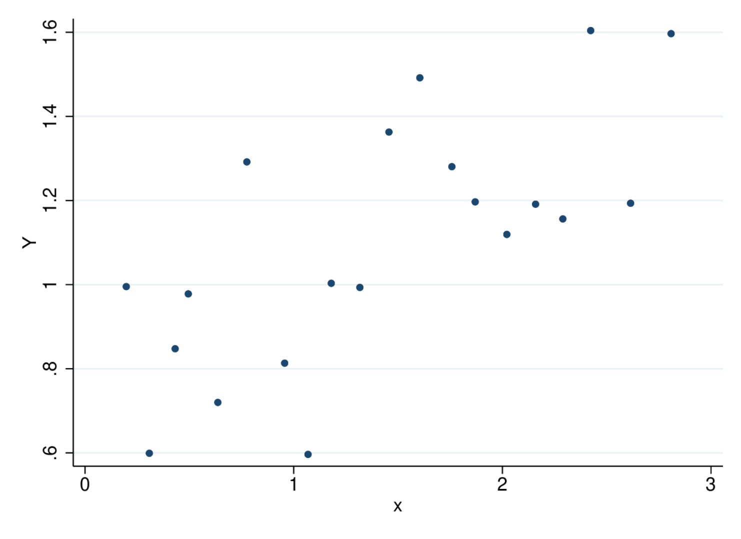
# Applet

https://teaching.sociology.ul.ie/apps/abx/

# Predictive in intent

- Asymmetric: use X to predict Y
- Find the best *a* and *b* to summarise the data scatter:
  - 'Best' is defined as minimising the squared deviations between the observed data-points and the fitted line, hence often called 'least-squares' regression
  - Deviations are the vertical distance between the line and the observed data points.
  - Very similar logic to the mean (minimise variance).

# A simple example: scatterplot



# Regression in Stata

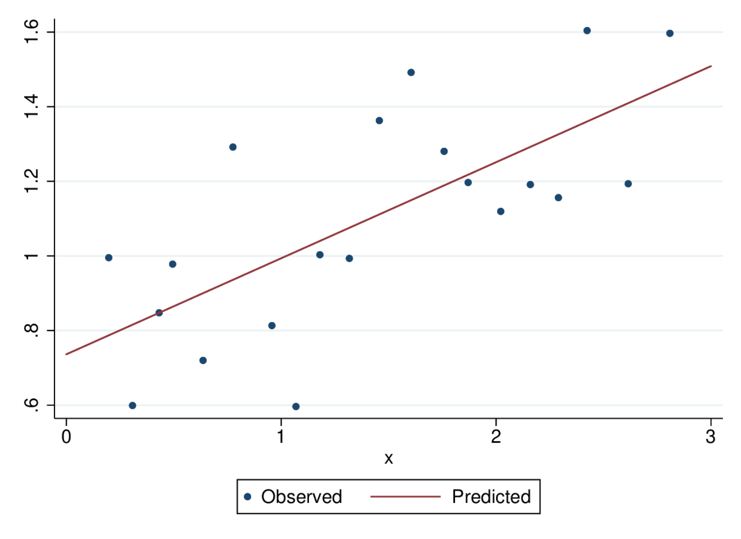
```
. reg y x

      Source |       SS           df       MS      Number of obs   =        20
-------------+----------------------------------   F(1, 18)        =     17.51
       Model |  .820567701         1  .820567701   Prob > F        =    0.0006
    Residual |  .843474028        18  .046859668   R-squared       =    0.4931
-------------+----------------------------------   Adj R-squared   =    0.4650
       Total |  1.66404173        19  .087581144   Root MSE        =    .21647

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .2574678   .0615269     4.18   0.001     .1282045    .3867311
       _cons |   .7363586   .0998061     7.38   0.000     .5266737    .9460435
------------------------------------------------------------------------------
```

# Regression equation

$$\hat{Y} = 0.7364 + X \times 0.2575$$

- To draw the line by hand, calculate two predicted values for Y at opposite sides of the graph
  - e.g., for X = 0, Y = a = 0.7364
  - for X = 3, Y = a + 3*b = 0.7364 + 3*0.2575 = 1.509
- Join them with a ruler!

# The line



# Predicted values

- The line gives a predicted value of $Y$ for each value of $X$:

$$\hat{Y} = a + bX$$

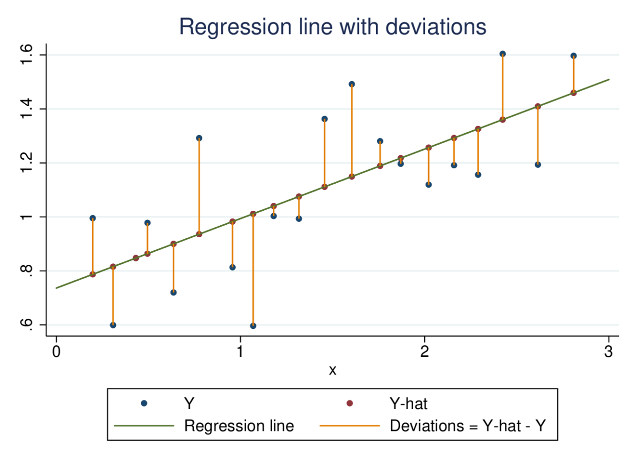
$$Y = \hat{Y} + e$$

$$Y = a + bX + e$$

  $e$ is the 'residual' or deviation.

- That is, knowing $X$ we "predict" or guess $Y$ as $a + bX$

# Deviations from the line



# Regression equation

- Regression equation: the estimate of $Y$, called $\hat{Y}$, depends on $X$:

$$\hat{Y} = a + bX$$

- The regression slope $b$ depends on $SXY$ and $SXX$, the intercept $a$ is calculated from $b$ and the mean values of $Y$ and $X$:

$$b = \frac{SXY}{SXX}$$

$$a = \bar{Y} - b\bar{X}$$

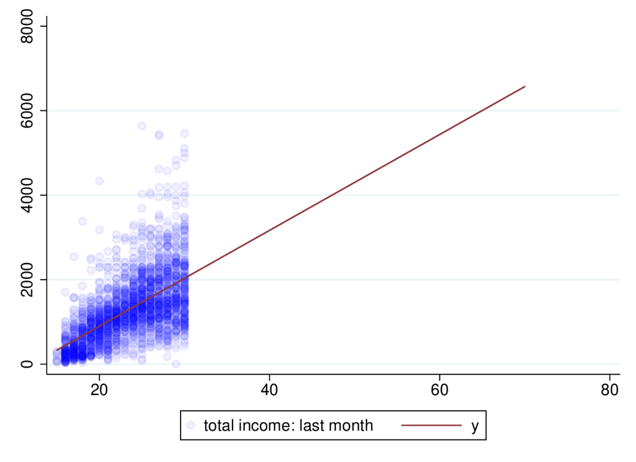
$$SXX = \sum (X_i - \bar{X})^2$$

$$SXY = \sum (X_i - \bar{X})(Y_i - \bar{Y})$$

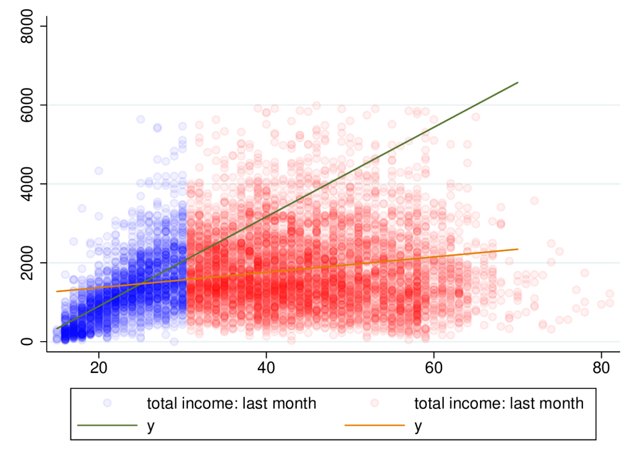
# Pitfalls

- Spurious relationships will fit just as well as real ones (e.g., if A affects B and A affects C, B and C will seem to be related and a regression line might fit well)
- Predicting outside the range of the data: the relationship we see only holds for the data we use, and it may well not hold for higher (or lower) values of X and Y
- Like correlation, non-linear relationships may be missed

# Predicting outside the range of the data: income and age for <30 years



# Predicting outside the range of the data: full age range



# Income and age: true relationship is non-linear

# Fit

- How well does it "fit"? We use $R^2$ to tell:
    - ranges from 0: no relationship at all
    - to 1: perfect relationship, all $Y$s are exactly equal to $a + bX$
    - values from 0.7 up indicate quite a good relationship
    - smaller values may indicate an interesting relationship
- In the case of bivariate regression (one independent variable), $R^2$ is the same as $r \times r$ (squared correlation coefficient).

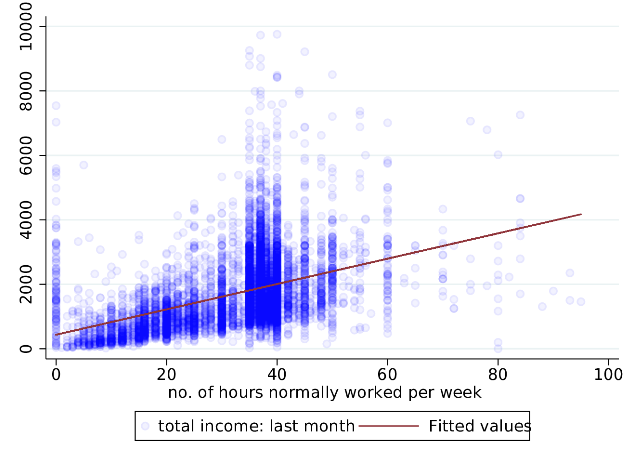
# Hypothesis testing

- If we regress Y on X we are asking whether X has a (statistical) effect on Y
- The null is therefore that X has no "effect" on Y
- This is equivalent to a slope, $\beta$, of zero: $\hat{Y} = \alpha + 0 \times X$
- In any sample from a population where this is true, b, the estimate of $\beta$ is likely to be close to, but not exactly zero
- We use its SE to create a t-stat: $t = \frac{\hat{\beta}}{\mathrm{SE}}$
- DF = n - k + 1 where k in number of X variables

# Regression in Stata:

```
. reg ofimn ojbhrs

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(1, 7943)      =   1398.95
       Model |  1.7000e+09         1  1.7000e+09   Prob > F        =    0.0000
    Residual |  9.6522e+09     7,943   1215179.2   R-squared       =    0.1497
-------------+----------------------------------   Adj R-squared   =    0.1496
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1102.4

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ojbhrs |   39.34202   1.051854    37.40   0.000     37.28011    41.40393
       _cons |   434.7389    36.8029    11.81   0.000     362.5955    506.8822
------------------------------------------------------------------------------
```

# Predicted regression line



# SO5041 Unit 16

## 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*

# Example

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

$$DWT = a + b \times PaidWork$$

or

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

- We can also fit a single multiple regression

$$DWT = a + b \times PaidWork + c \times Female$$

sociology
UNIVERSITY OF LIMERICK

# 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 $DWT = a + b \times PaidWork + c \times Female$ regression is the net change in predicted domestic work time for females, once you take account of paid work time.
- The *b* coefficient is then the net effect of a unit change in paid work time, once you take gender into account.

# Sex only predicting income

```
. reg ofimn i.osex

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(1, 7943)      =    606.72
       Model |  805586626         1   805586626    Prob > F        =    0.0000
    Residual |  1.0547e+10     7,943  1327780.12   R-squared       =    0.0710
-------------+----------------------------------   Adj R-squared   =    0.0708
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1152.3

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        osex |
     female |  -637.3352   25.87467   -24.63   0.000    -688.0563    -586.614
       _cons |   2062.275   18.64855   110.59   0.000     2025.719    2098.831
------------------------------------------------------------------------------
```

# Sex and job hours predicting income

```
. reg ofimn ojbhrs i.osex

      Source |       SS           df       MS      Number of obs   =     7,945
-------------+----------------------------------   F(2, 7942)      =    794.96
       Model |  1.8935e+09         2   946761687   Prob > F        =    0.0000
    Residual |  9.4586e+09     7,942  1190962.07   R-squared       =    0.1668
-------------+----------------------------------   Adj R-squared   =    0.1666
       Total |  1.1352e+10     7,944  1429021.17   Root MSE        =    1091.3

------------------------------------------------------------------------------
       ofimn |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      ojbhrs |   33.96065   1.123629    30.22   0.000     31.75804    36.16326
             |
        osex |
     female  |  -337.0889   26.44232   -12.75   0.000    -388.9228    -285.255
       _cons |   787.1759   45.73595    17.21   0.000     697.5214    876.8304
------------------------------------------------------------------------------
```

# Sex and hours