UL Summer School: Regression session 1

Brendan Halpin, Sociology
2023 Summer School

Outline

Session 1: Correlation and bivariate regression

Outline

Outline

- How to summarise association between pairs of interval/ratio variables
- Linear (straight line) association
- Correlation summarises the strength and direction of the relationship
- Bivariate regression treats one variable as a response, one as a predictor
- Estimates the "effect" of the predictor on the response
- Bivariate regression generalises easily to multiple predictors, "multiple regression"

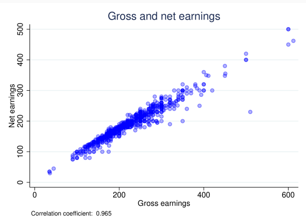
Session 1: Correlation and bivariate regression

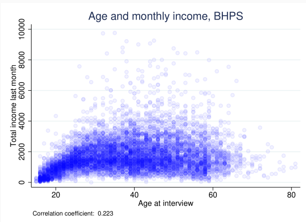
Correlation

Correlation

- We can visualise relationships between interval/ratio variables with scatterplots
- Correlation & regression seek to model the relationship as a straight line
  - with greater and lesser success

Scatterplots





Summarising association simply

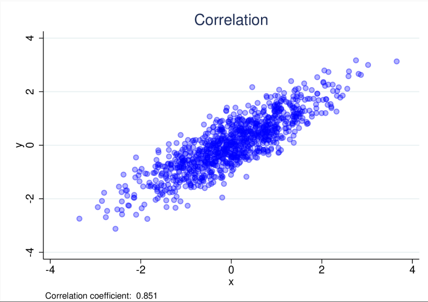
- We can see a lot of detail in a scatterplot, but sometimes we can summarise it in simple ways
- For instance the two variables may have a positive association: when one is high the other tends to be high, and vice versa
- Or a negative association: when one is high the other tends to be low

Correlation coefficient

- How well a straight line summarises the relationship
- Positive or negative
- Zero implies no relationship
- See `http://teaching.sociology.ul.ie:3838/so5041/corr`
- And `http://teaching.sociology.ul.ie:3838/apps/corrspread`
- Skip

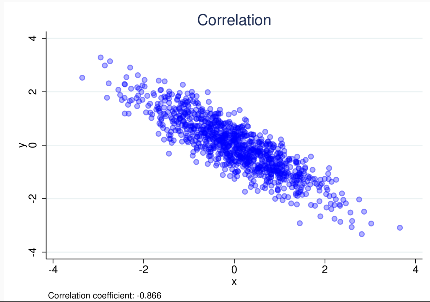
Strong positive

Figure 3: Fictional data displaying a strong positive linear relationship.



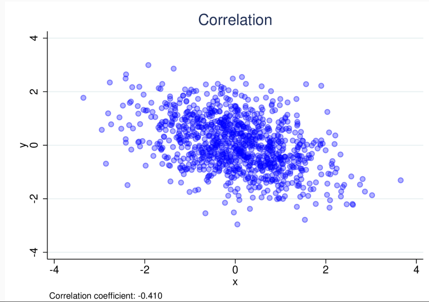
Strong negative

Figure 4: Fictional data displaying a strong negative linear relationship.



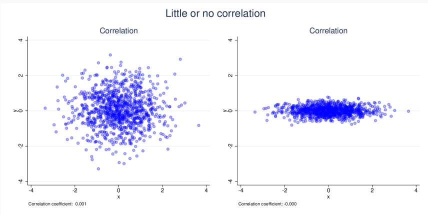
Weak negative

Figure 5: Fictional data displaying a weak negative linear relationship.



No relationship

Figure 6: Fictional data displaying absence of a relationship.



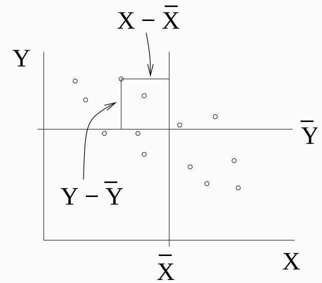
Practice app

- View scatterplot, guess correlation

http://teaching.sociology.ul.ie:3838/apps/corrgame

Calculating the correlation coefficient

- Combine X deviations ($X_i - \bar{X}$) and Y deviations ($Y_i - \bar{Y}$) – i.e., compares each point with the mean for X and the mean for Y
- With positive association, cases below average on X tend to be below average on Y (and above average on X tend to be above average on Y)
- With negative association, cases below average on X tend to be above average on Y and vice versa



Pearson Product-Moment Correlation Coefficient

- Pearson Product-Moment Correlation Coefficient (r)

$$r = \frac{SXY}{\sqrt{SXX.SYY}}$$

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

$$SYY = \Sigma(Y - \bar{Y})^2$$

$$SXY = \Sigma(X - \bar{X})(Y - \bar{Y})$$

- Range: $-1 \leq r \leq +1$
- r is a symmetric measure: $r_{xy} = r_{yx}$

Combining deviations

- With positive association below the mean, both $X_i - \bar{X}$ and $Y_i - \bar{Y}$ are negative, so $SXY = (X_i - \bar{X})(Y_i - \bar{Y})$ is positive
- With negative association, $X_i - \bar{X}$ and $Y_i - \bar{Y}$ tend to have opposite signs, so SXY is negative
- If no association, SXY is approximately zero
- Scaling by $\frac{1}{\sqrt{SXX.SYY}}$ makes its range $-1 \leq r \leq +1$

Robust to transformations

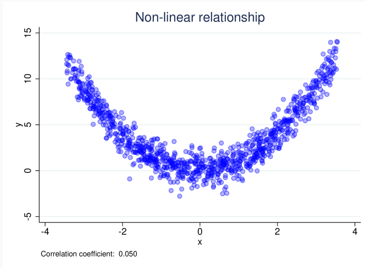
- Changing the scale of one variable (additively or multiplicatively) doesn't change the results
  - Corr(x, y) = Corr(x + c, y)
  - Corr(x, y) = Corr(x × c, y)
- => "Scale invariant": the correlation between ice-cream sales and temperature doesn't change if you switch between ℉ and ℃
- If $y = a + b \times x$:
  - $r_{xy} = r_{yx} = 1$
  - $r_{zx} = r_{zy}$
- Suits interval and ratio variables

Pitfalls

- Correlation is not causality (association is not proof of a causal relationship)
- Absence of correlation does not mean absence of relationship: non-linear relationships may exist

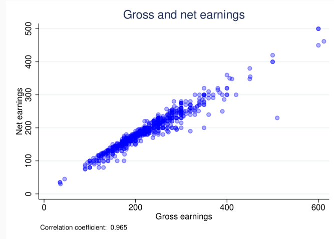
Non-linear!

A strong non-linear relationship and a near-zero correlation.



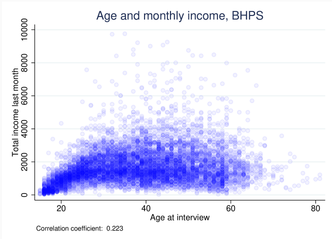
Correlations on data: School Leaver's Earnings

```
. corr grsearn netearn
(obs=756)

             |  grsearn  netearn
-------------+------------------
     grsearn |   1.0000
     netearn |   0.9650   1.0000
```



Correlations on data: BHPS

```
. corr ofimn oage
(obs=7,934)

             |    ofimn     oage
-------------+------------------
       ofimn |   1.0000
        oage |   0.2228   1.0000
```



Strong and weak

- There is a strong simple mechanism linking gross and net earnings, leading to a very high correlation of 0.965
- The relationship between age and income is much more complex, but is still real: thus a much lower correlation of 0.223

Hypothesis testing

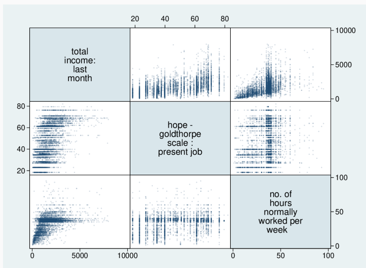
- Null hypothesis: no association, correlation = 0
- Test statistic: $\frac{r}{SE}$, normally distributed (SE not in output)
- P-value: Chance of getting a correlation this far from zero if null is true

```
. pwcorr ofimn oage, sig

             |    ofimn     oage
-------------+------------------
       ofimn |   1.0000
             |
             |
        oage |   0.2228   1.0000
             |   0.0000
             |
```


Estimating correlations in Stata

```
. pwcorr ofimn ojbhgs ojbhrs, sig

             |    ofimn   ojbhgs   ojbhrs
-------------+---------------------------
       ofimn |   1.0000
             |
             |
      ojbhgs |   0.4851   1.0000
             |   0.0000
             |
      ojbhrs |   0.4245   0.2489   1.0000
             |   0.0000   0.0000
             |
```


Viewing the same correlations



Practice apps summary

- http://teaching.sociology.ul.ie:3838/so5041/corr
- http://teaching.sociology.ul.ie:3838/apps/corrspread
- http://teaching.sociology.ul.ie:3838/apps/corrgame

Session 1: Correlation and bivariate regression

Bivariate regression

Correlation is limited

- Correlation summarises straight-line association between two interval/ratio variables
- Single statistic, from -1 (perfect negative) through 0 (no association) to 1 (perfect positive)
- How well a scatter is described by a line
- Regression analysis goes one further: find the line

Bivariate regression: relating 2 (or more) interval/ratio variables

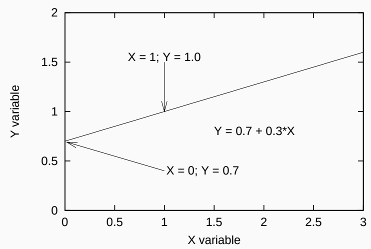
- Identifying the line that best summarises the scatterplot
- Directional: One "response" variable (Y), one (or more) "predictors" (X)
- Predictive: Given the relationship between X and Y, knowing X helps us predict Y better
- Reading: Agresti, *Statistics for the Social Sciences*, Ch 9 or any intro text

Bivariate vs multiple regression

- Bivariate regression: one X variable predicting one Y
- Multiple regression: multiple X variables predicting one Y
  - Estimate "net" effect of each X variables, "controlling for" the others
  - Very powerful general model, generalises easily from bivariate regression

Some geometry: equation of a line

$$Y = a + bX$$



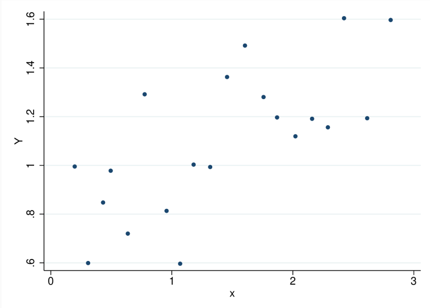
Applet

http://teaching.sociology.ul.ie:3838/apps/abx/

What's the "best" line?

- '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

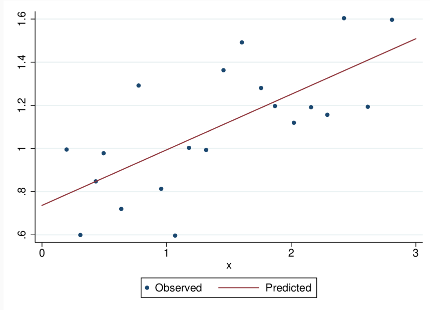
```
. reg y x

      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:

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

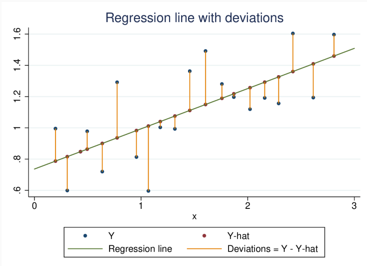
$$Y = \hat{Y} + e$$

$$Y = a + bX + e$$

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:

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

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

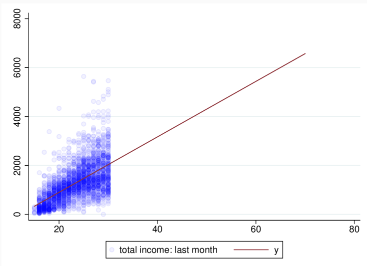
$$SXX = \sum (X_i - \bar{X})^2$$

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

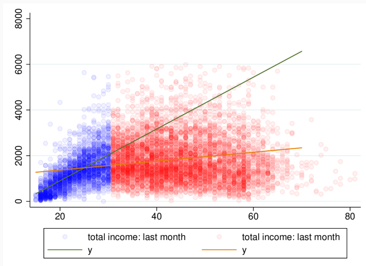
Pitfalls

- 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

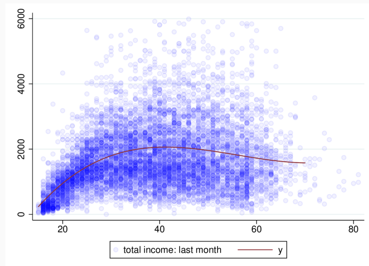
Predicting outside the range of the data: income and age for <30 years



Predicting outside the range of the data: full age range



Income and age: true relationship is non-linear



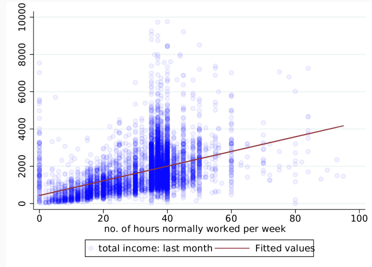
Fit: How well the straight line captures the relationship

- How well does it “fit”? We use R^2 to tell:
  - ranges from 0: no relationship at all
  - to 1: perfect relationship, all Ys 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).

Regression in Stata:

```
. reg income hours

      Source |       SS           df       MS      Number of obs   =       737
-------------+----------------------------------   F(1, 735)       =    104.91
       Model |  86809818.5         1  86809818.5   Prob > F        =    0.0000
    Residual |   608164307       735  827434.432   R-squared       =    0.1249
-------------+----------------------------------   Adj R-squared   =    0.1237
       Total |   694974126       736  944258.323   Root MSE        =    909.63

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       hours |   28.27485   2.760468    10.24   0.000     22.85551    33.69419
       _cons |   742.3841   97.00251     7.65   0.000      551.949    932.8191
------------------------------------------------------------------------------
```


Predicted regression line



Hypothesis testing

- Linear regression is asking whether Y is "affected by" X
- The interpretation of the b estimate is the effect of a 1-unit change in X on the predicted $\hat{Y}$
- If X has no effect, the true value of b is zero
- Can we reject the null hypothesis that b = 0?
  - Does the CI around b include zero?
  - Is the absolute value of b/SE greater than the critical value of t?

Example

- In the previous example, b=39.34, with an SE of 1.05
  - Confidence interval: $b \pm SE \times 1.96$: 37.28011 to 41.40393 <– excludes zero
  - t-stat: $\frac{b}{SE} = 37.4 >> 1.96$
- Conclusion: null hypothesis of no effect extremely unlikely to be true
- More formally: the pattern in this sample very unlikely to be observed if no effect in population

The key numbers to read

```
. 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
------------------------------------------------------------------------------
```


Practice apps

- Equation of a line: `http://teaching.sociology.ul.ie:3838/apps/abx`
- Reading regression output: `http://teaching.sociology.ul.ie:3838/bivar`

Session 1: Correlation and bivariate regression

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: income may be affected by gender, and also by work time: 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 multiple regression

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

sociology

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 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.

sociology

Sex and income: independent samples t-test

```
. ttest income, by(sex)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. err.   Std. dev.   [95% conf. interval]
---------+--------------------------------------------------------------------
    male |     346    1991.997    55.94547    1040.646    1881.959    2102.034
  female |     391    1394.113    40.95243    809.7818    1313.598    1474.628
---------+--------------------------------------------------------------------
Combined |     737    1674.802    35.79412    971.7296    1604.531    1745.073
---------+--------------------------------------------------------------------
    diff |            597.8836    68.29865                463.7999    731.9673
------------------------------------------------------------------------------
    diff = mean(male) - mean(female)                              t =   8.7540
H0: diff = 0                                     Degrees of freedom =      735

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

sociology

Sex only predicting income

```
. reg income i.sex

      Source |       SS           df       MS      Number of obs   =       737
-------------+----------------------------------   F(1, 735)       =     76.63
       Model |  65617342.3         1  65617342.3   Prob > F        =    0.0000
    Residual |   629356784       735  856267.733   R-squared       =    0.0944
-------------+----------------------------------   Adj R-squared   =    0.0932
       Total |   694974126       736  944258.323   Root MSE        =    925.35

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         sex |
      female |  -597.8836   68.29865    -8.75   0.000    -731.9673   -463.7999
       _cons |   1991.997   49.74698    40.04   0.000     1894.333     2089.66
------------------------------------------------------------------------------
```

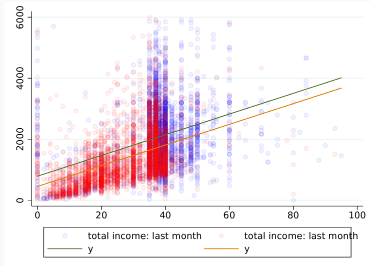
sociology

Sex and job hours predicting income

```
. reg income hours i.sex

      Source |       SS           df       MS      Number of obs   =       737
-------------+----------------------------------   F(2, 734)       =     70.91
       Model |   112534221         2  56267110.4   Prob > F        =    0.0000
    Residual |   582439905       734  793514.857   R-squared       =    0.1619
-------------+----------------------------------   Adj R-squared   =    0.1596
       Total |   694974126       736  944258.323   Root MSE        =    890.79

------------------------------------------------------------------------------
      income | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       hours |   22.29842   2.899927     7.69   0.000     16.60528    27.99156
             |
         sex |
      female |  -401.5815   70.53076    -5.69   0.000    -540.0475   -263.1154
       _cons |   1152.519   119.2162     9.67   0.000     918.4735    1386.564
------------------------------------------------------------------------------
```

sociology

Sex and hours



sociology