Multivariate descriptives

Johan A. Elkink
School of Politics & International Relations
University College Dublin

25 September 2017

1. Two scale variables
2. Categorical and scale variable
3. Two categorical variables

Correlation

Two variables are correlated when the variance in one variable coincides with variance in another variable.

- **Positive correlation**: high values in one variable coincide with high values in another variable.
- **Negative correlation**: high values in one variable coincide with low values in another variable.

Causation

Slightly simplified, for X to be a cause of Y, we generally require:

1. X to precede Y
2. X to correlate with Y (either positively or negatively)
3. no other factor to explain the correlation between X and Y (no **confounding factor**)

Causation: terminology

If X causes Y,

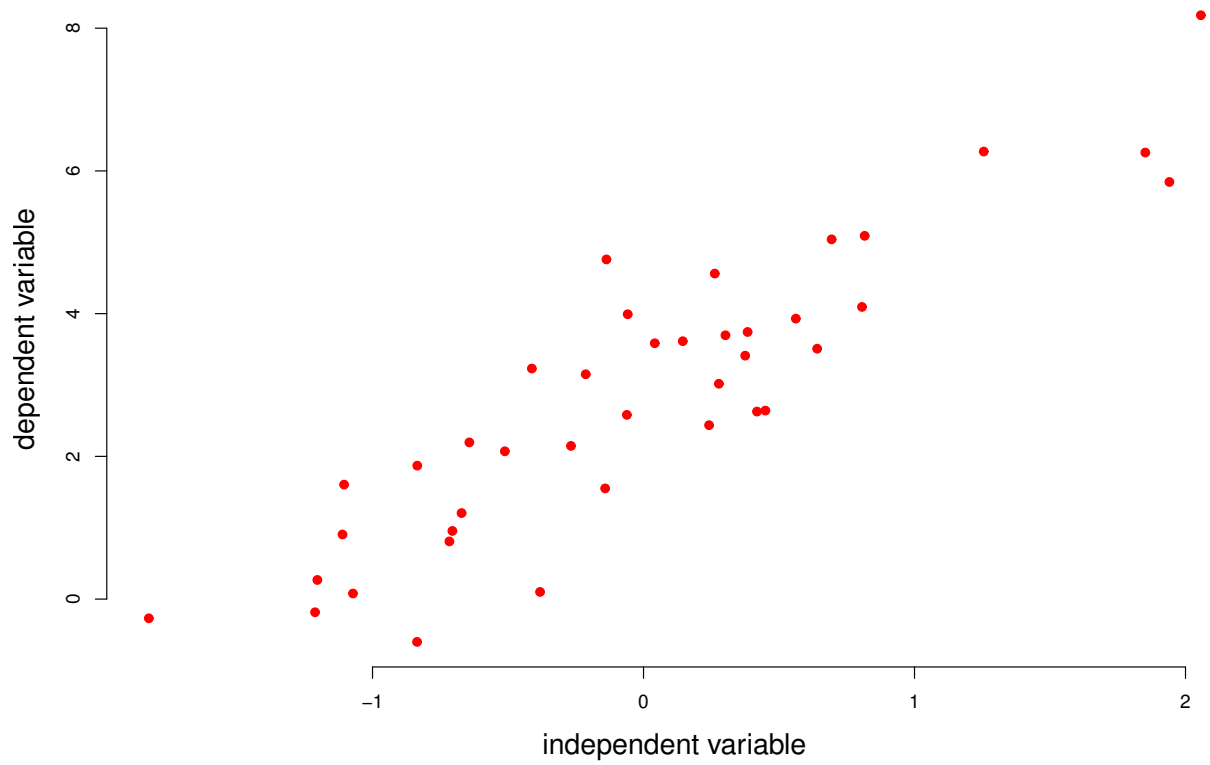
- Y is called the **dependent variable**, or **outcome variable**, or **response**, or . . . ;
- X is called the **independent variable**, or **explanatory variable**, or **factor**, or . . . .

In political science, most common (unfortunately) is the usage of the terms independent and dependent variables.

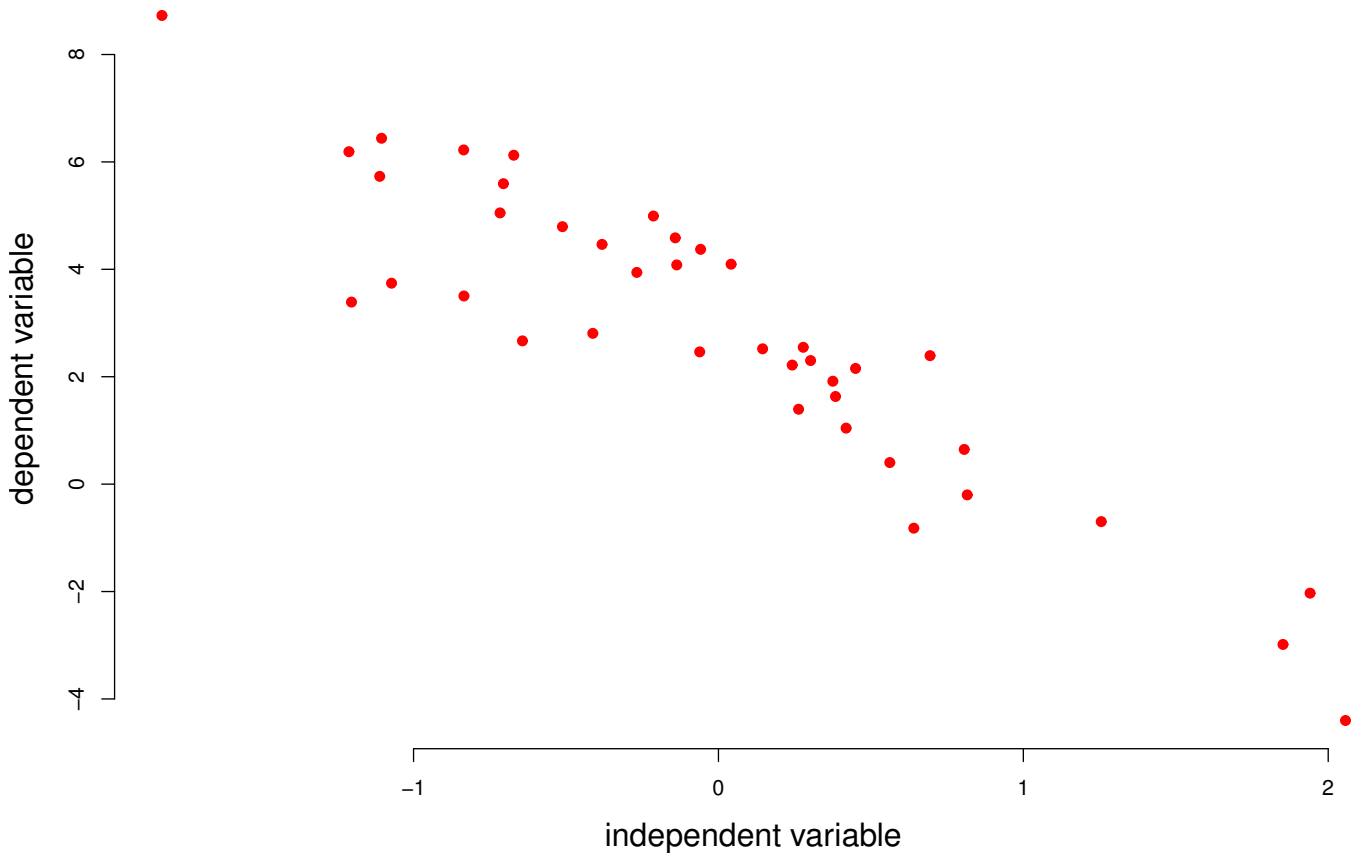
Outline

Scatter plot

The relationship between two continuous variables is most easily displayed using a scatter plot.



Scatter plot



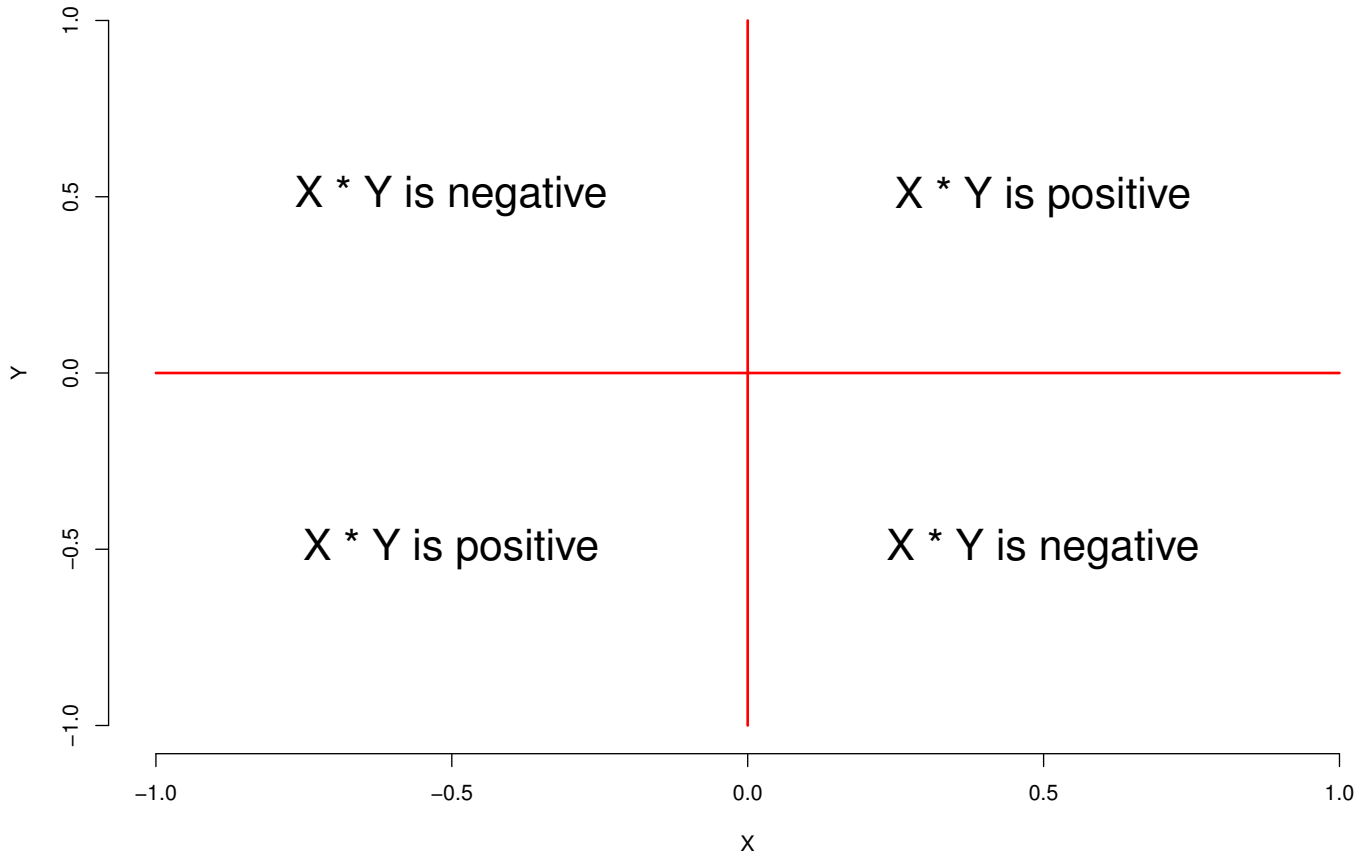
Covariance

$$Var(x) = s^2 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}$$

$$Cov(x, y) = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{N}$$

- Unbounded - size depends on variance in x and y
- Positive correlation: $Cov(x, y) > 0$
- Negative correlation: $Cov(x, y) < 0$

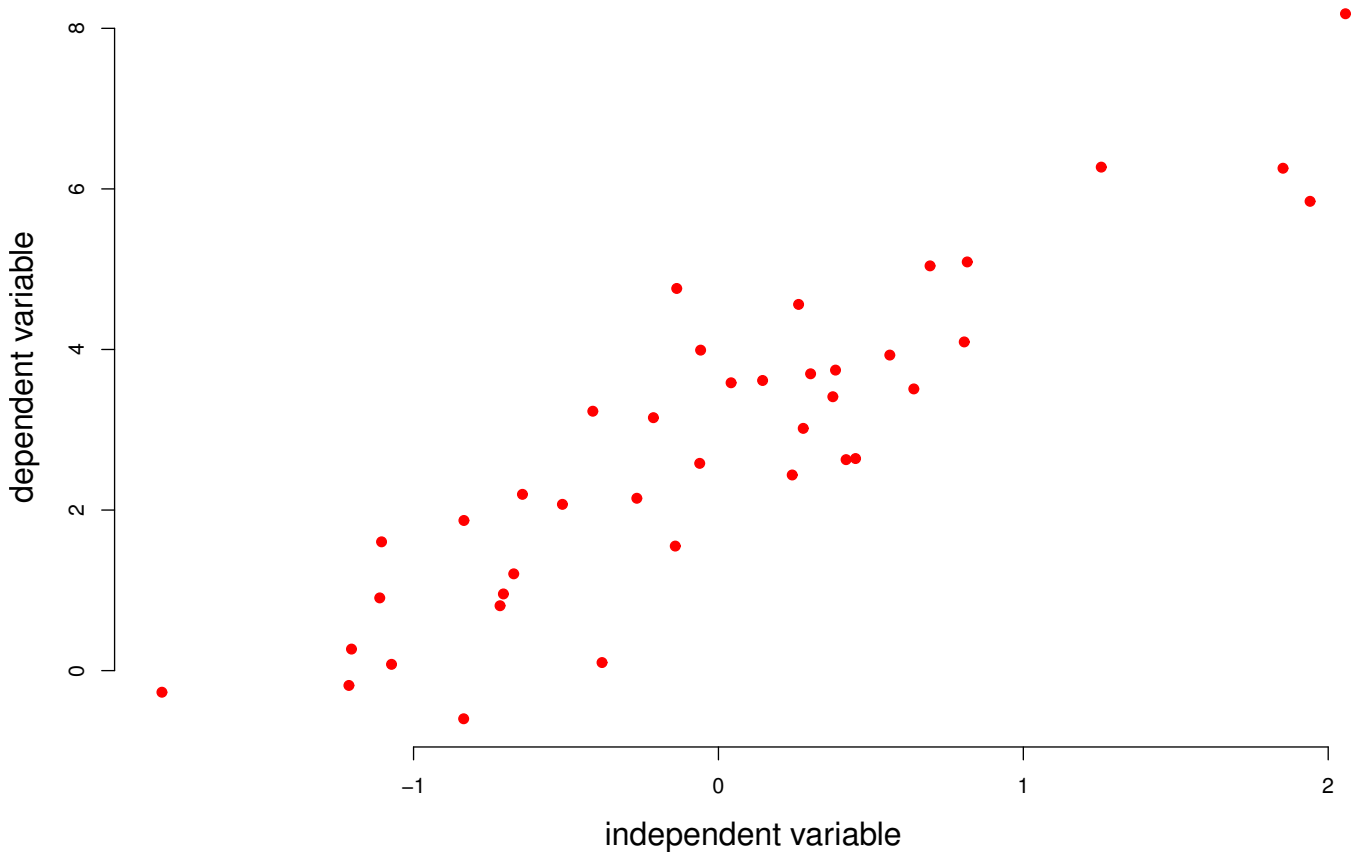
Covariance: intuition



Pearson correlation

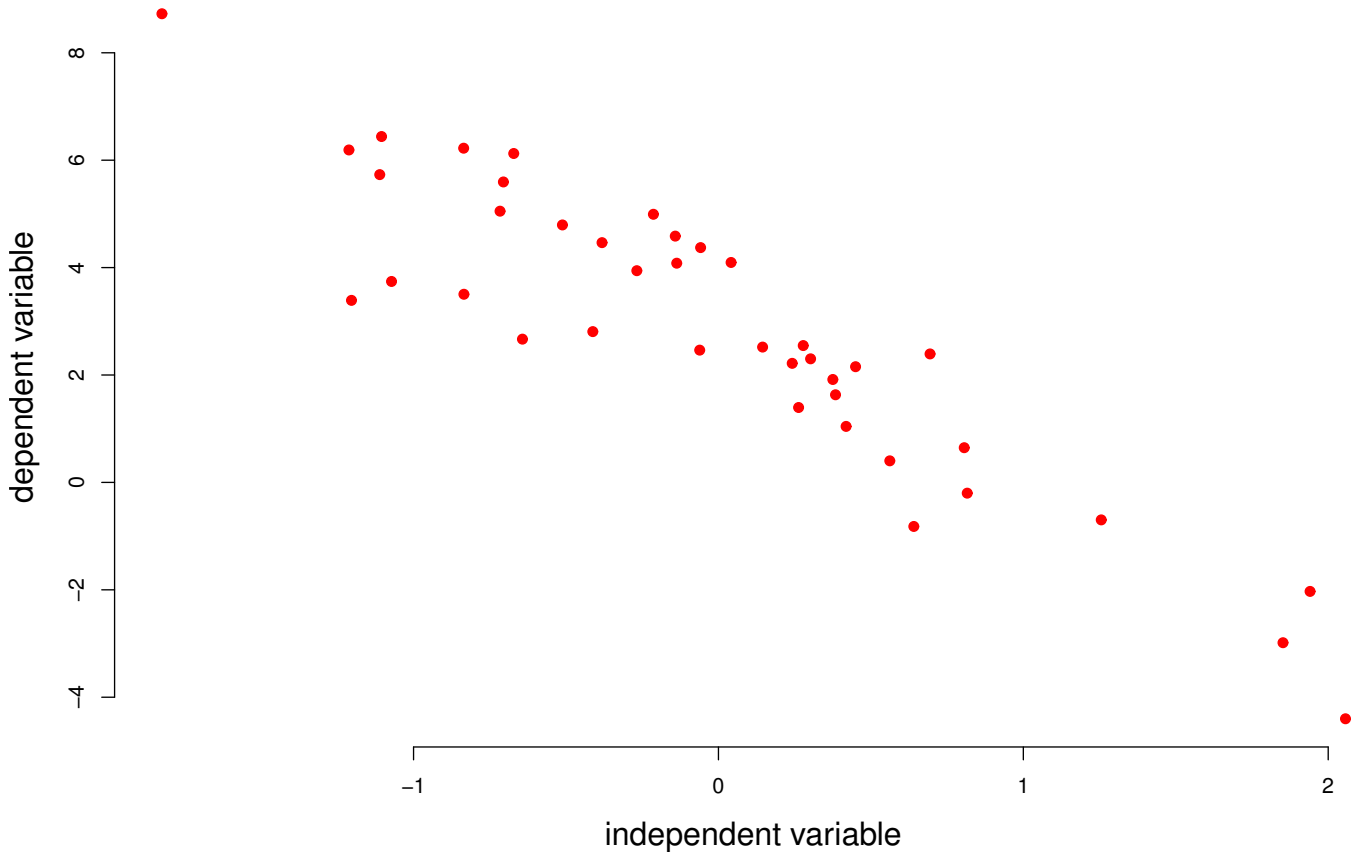
$$r_{xy} = \frac{Cov(x,y)}{s_x s_y} = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

- Bounded between -1 and 1
- Positive correlation: $r > 0$
- Negative correlation: $r < 0$



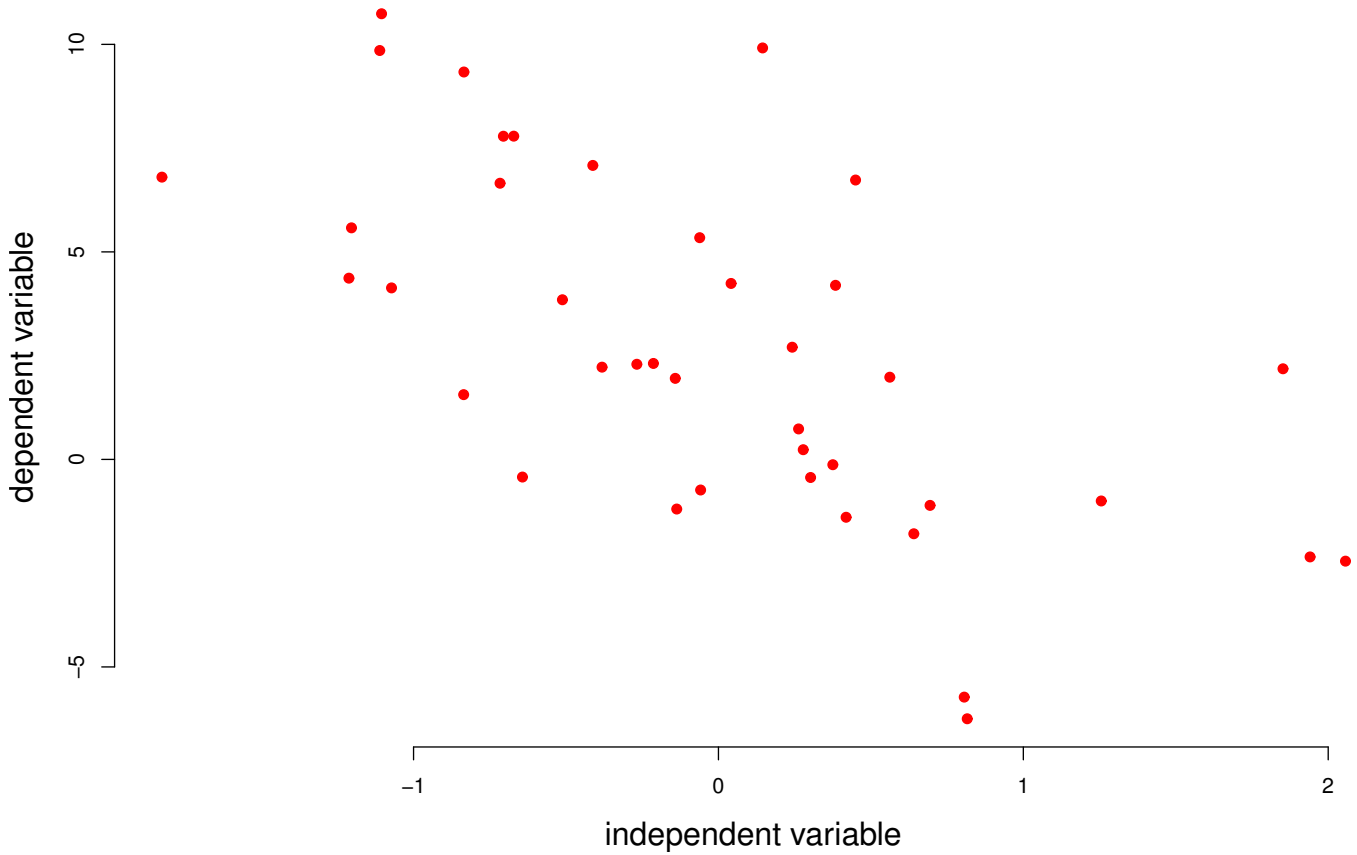
$Var(x) = 1.17$ $Var(y) = 4.95$

$Cov(x, y) = 2.15$ $Cor(x, y) = 0.89$



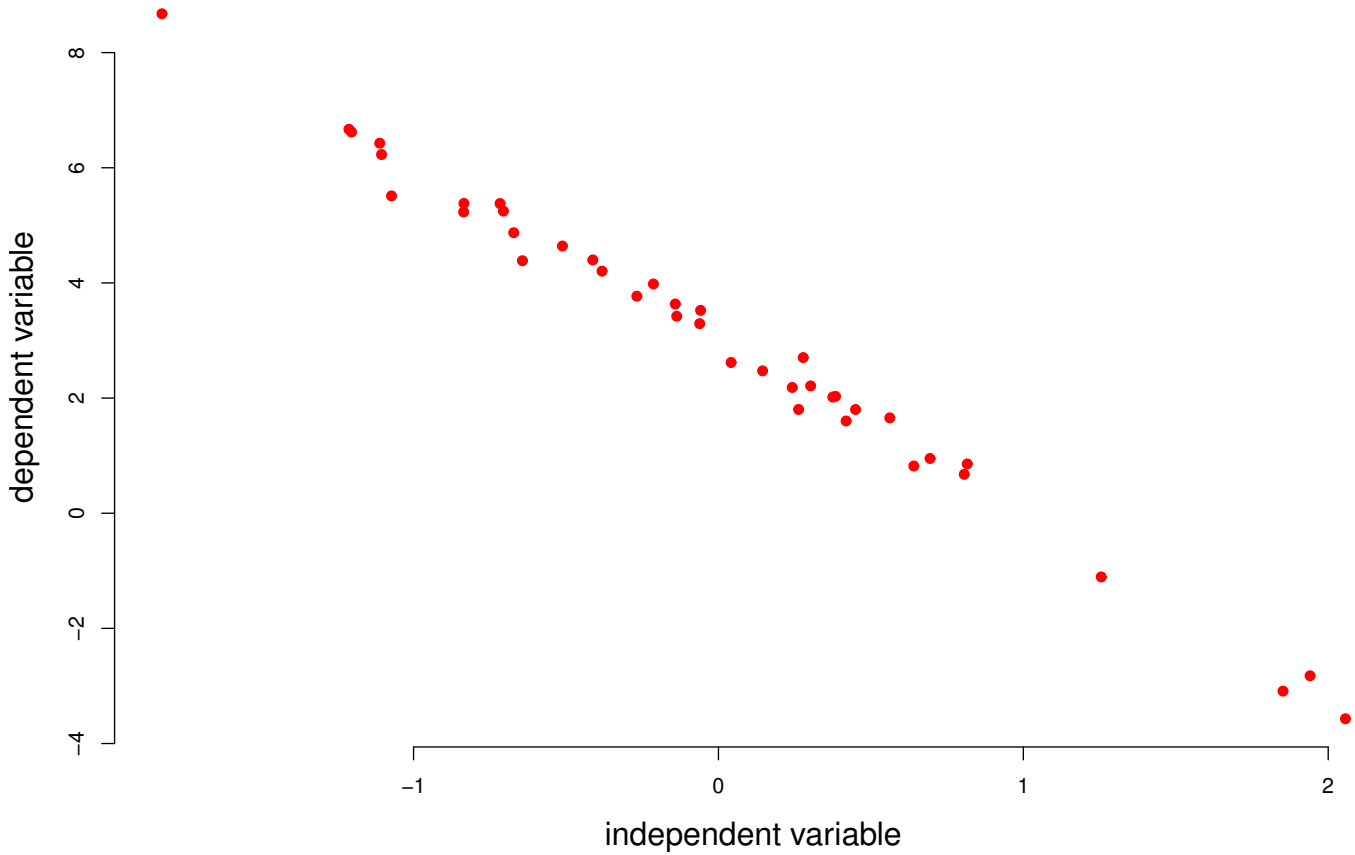
$Var(x) = 1.17$ $Var(y) = 10.60$

$Cov(x, y) = -3.35$ $Cor(x, y) = -0.95$



$Var(x) = 1.17$ $Var(y) = 26.80$

$Cov(x, y) = -4.21$ $Cor(x, y) = -0.75$



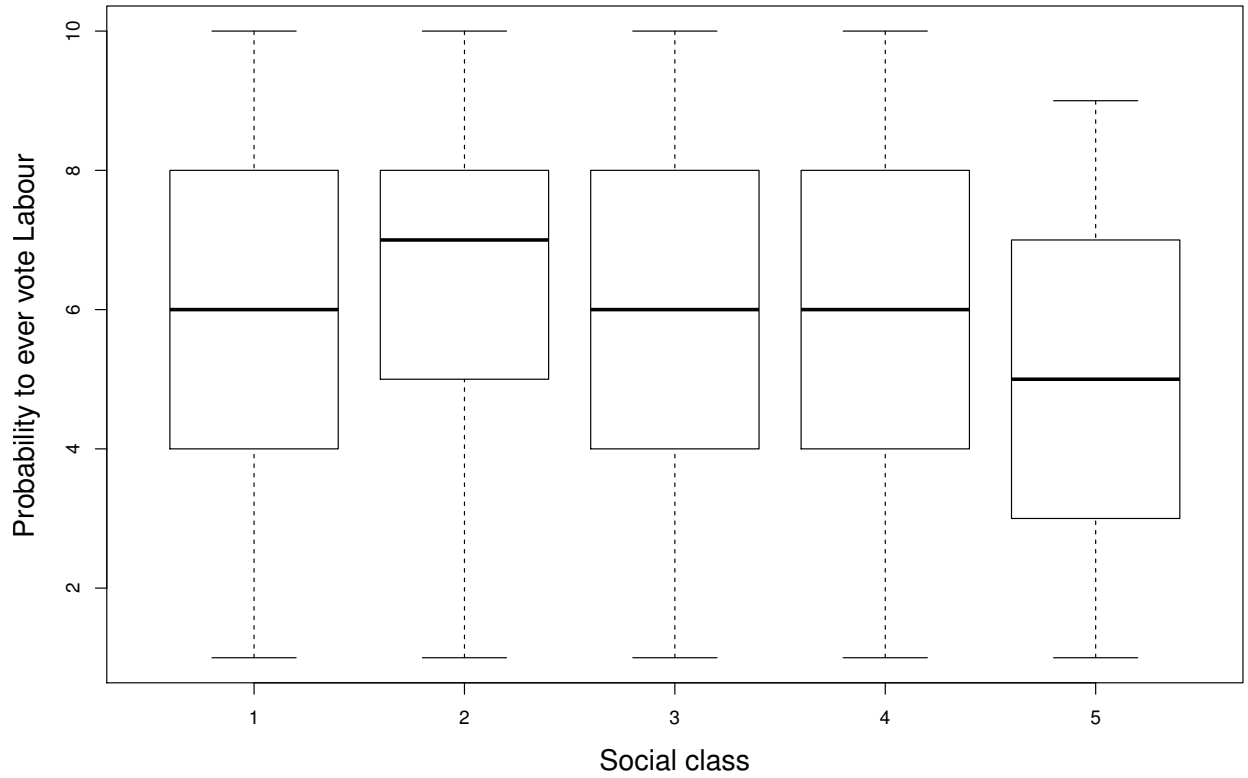
$Var(x) = 1.17$ $Var(y) = 10.90$

$Cov(x, y) = -3.58$ $Cor(x, y) = -0.99$

Outline

Box plots

One way to graphically look at the relation between a categorical and a scale variable is the use of multiple box plots next to each other.



Outline

Example: English parliament

Data from a vote in 1844 to reduce the working day for children to 10 hours.

party	count	perc.	vote	count	perc.
liberals	150	39%	yes	194	50.4%
conversatives	235	61%	no	191	49.6%
total	385			385	

(http://www.let.leidenuniv.nl/history/RES/stat/html/les9.html)

Example: English parliament

	yes	no	total
liberals	94	56	150
conservatives	100	135	235
total	194	191	385

Exercise: calculate row and column percentages.

Which are more appropriate?

(`http://www.let.leidenuniv.nl/history/RES/stat/html/les9.html`)