105demography

B-MAT-100

Relationships between variables

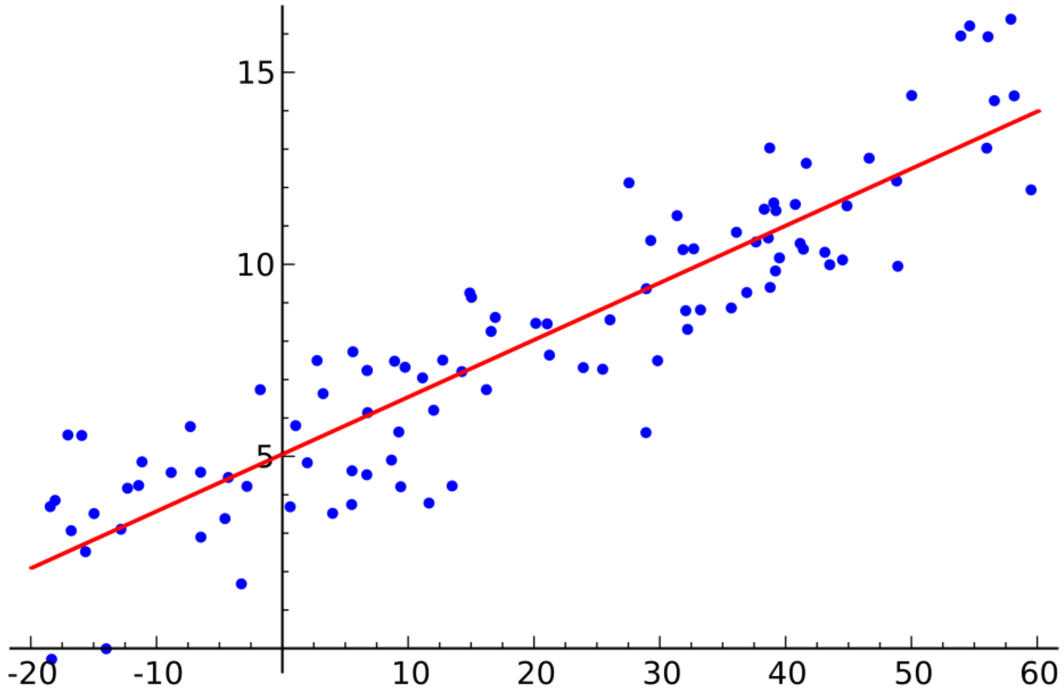
- There often is relationships between different variables
  - Example: radius and area of a circle, gas pressure and temperature…
- This relationships are expressed with an equation:

$$Y = f(X)$$

- X is the independent variable and Y is the dependent variable
- Y is explained by X
- What if your data come from observations and no exact relationship exists?

Regression analysis

- Estimation of the relationships between variables
- Given a set of N data points (X_i, Y_i), the goal is to find a function f such as
$$Y_i \approx f(X_i)$$
- This function is called a regression, or a fit



Correlation theory

- Correlation indicates how closely the data fits the regression
- If the fit is exact, there is a perfect correlation
  - Example: radius and area of a circle
- There is no correlation when the variables are independent
  - Example: two dice rolls
- If the variable are somewhat related, there is some correlation
  - Example: size and weight of an individual

Linear regression

- Simple model with a single independent variable
- The regression is a line:

$$f(X) = aX + b$$

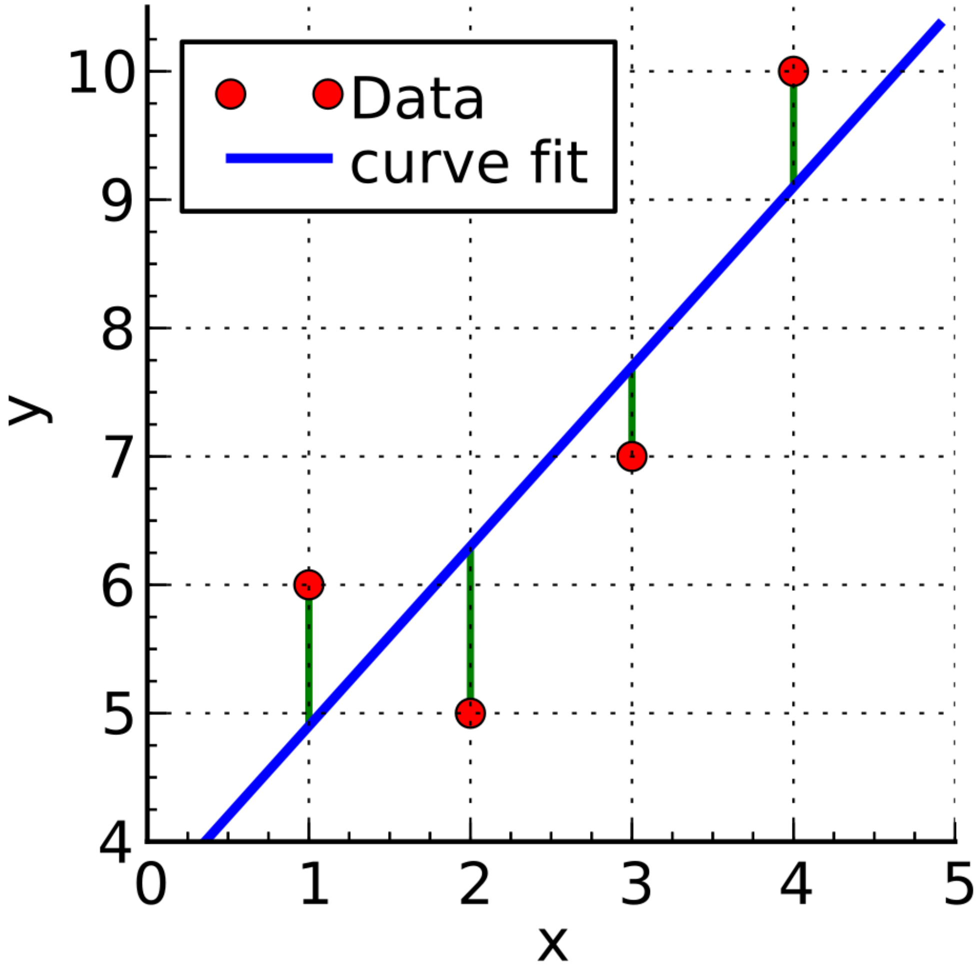
- The differences between the prediction of the fit $f(X_i) = \widehat{Y}_i$ and the actual observation Y_i are called the residuals ε_{i}

$$Y_i = \widehat{Y}_i + \varepsilon_{\mathrm{i}} = aX_i + b + \varepsilon_{\mathrm{i}}$$

- a and b are obtained by minimizing the sum of squared residuals. This is the method of least squares

Least squares (1/2)

- The goal is to find a and b to minimize
$$S = \sum_{i=1}^{N} \varepsilon_i^2$$
- Since $Y_i = aX_i + b + \varepsilon_i$ we can write
$$S = \sum_{i=1}^{N} (Y_i - aX_i - b)^2$$
- The minimum is found by setting the gradient to 0:
$$\frac{\partial S}{\partial a} = \frac{\partial S}{\partial b} = 0$$



Least squares (2/2)

- By expanding the derivates we get the following equations

$$\begin{cases} \sum_{i=1}^{N} Y_i = a \sum_{i=1}^{N} X_i + bN \\ \sum_{i=1}^{N} X_i Y_i = a \sum_{i=1}^{N} X_i^2 + b \sum_{i=1}^{N} X_i \end{cases}$$

- And then

$$a = \frac{N(\sum XY) - (\sum X)(\sum Y)}{N(\sum X^2) - (\sum X)^2}$$

$$b = \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N(\sum X^2) - (\sum X)^2}$$

Root-mean-squared deviation

- Quantifies the amount of dispersion of the data around the fit

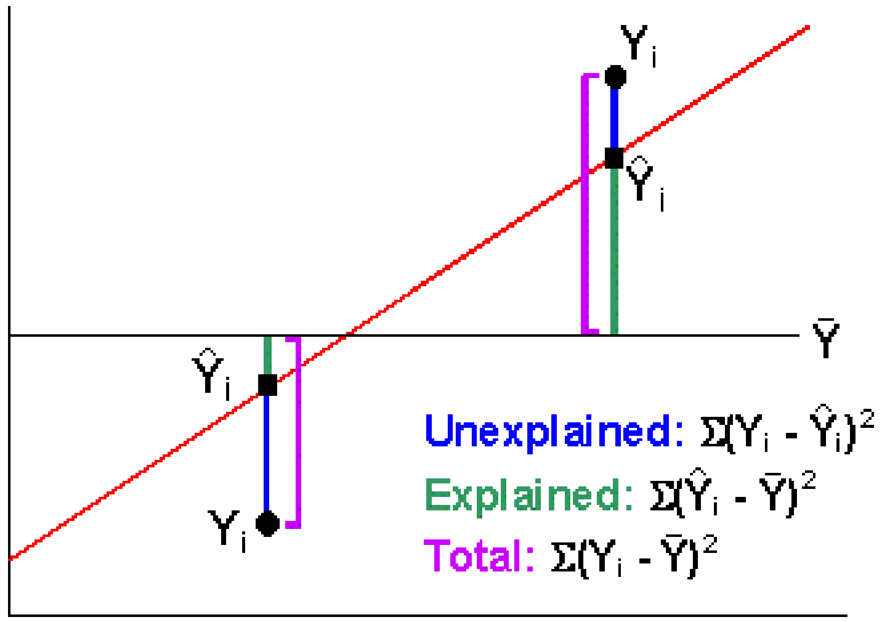
$$s_{Y,X} = \sqrt{\frac{\sum_{i=1}^{N} \varepsilon_i^2}{N}} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2}{N}}$$

- Similar properties as the standard deviation: if N is large enough, 68% of the data points are at a distance less than $s_{Y,X}$, 95% at less than $2s_{Y,X}$, and 99.7% at less than $3s_{Y,X}$.

Explained and unexplained variances

$$\sum_{i=1}^{N}(Y_i - \bar{Y})^2 = \sum_{i=1}^{N}\left(Y_i - \widehat{Y}_i\right)^2 + \sum_{i=1}^{N}\left(\widehat{Y}_i - \bar{Y}\right)^2$$

- $\sum(Y - \bar{Y})^2$ is the total variance of Y (TSS: Total sum of squares)
- $\sum\left(Y - \widehat{Y}_i\right)^2$ is the unexplained variance (RSS: Residual sum of squares)
- $\sum\left(\widehat{Y}_i - \bar{Y}\right)^2$ is the explained variance (ESS: Explained sum of squares)



Coefficient of correlation

- The correlation coefficient is the proportion of expected variance in the total variance:

$$r = \sqrt{\frac{ESS}{TSS}} = \sqrt{\frac{\sum\left(\widehat{Y}_i - \bar{Y}\right)^2}{\sum(Y_i - \bar{Y})^2}}$$

- If σ_Y is the standard deviation of Y, r can also be written

$$r = \sqrt{1 - \frac{RSS}{TSS}} = \sqrt{1 - \frac{\sum\left(Y_i - \widehat{Y}_i\right)^2}{\sum(Y_i - \bar{Y})^2}} = \sqrt{1 - \frac{s_{Y,X}^2}{\sigma_Y^2}}$$

- r measures the quality of the fit

Covariance

- If we suppose there is a linear relationship between two variables X and Y with standard deviations σ_X and σ_Y, we can define the covariance as:

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$
$$= \frac{\sum(X_i - \bar{X})\sum(Y_i - \bar{Y})}{N}$$

- In this case, the coefficient of correlation can now be expressed as:

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

105demography

- Goal: Use the linear least square regression to predict a country's population
- 105demography_data.csv: data file containing every country's population from 1960 to 2017
- Inputs: one or several country codes

105demography

- Y is the population
- X is the year
- Outputs
  - Coefficients of the fit $Y = a_X X + b_X$
  - Root-mean-square deviation of the fit
  - Population prediction in 2050 according to this fit
  - Coefficients of the fit $\mathrm{X} = a_Y Y + b_Y$
  - Root-mean-square deviation of the fit
  - Population prediction in 2050 according to this fit
  - Coefficient of correlation

Exercise: Sums

- Given a data set (X_i, Y_i), compute the following sums:
  - $\sum X$
  - $\sum Y$
  - $\sum X^2$
  - $\sum Y^2$
  - $\sum X.Y$
- It should then be easy to compute the coefficients of a linear fit

Exercise: root-mean-square deviation

- Given a data set (X_i, Y_i) and a linear fit, compute the root-mean-square deviation of the fit.

$$s_{Y,X} = \sqrt{\frac{\sum_{i=1}^{N}\left(Y_i - \widehat{Y_i}\right)^2}{N}}$$

Exercise: data parsing

- Given a country code, parse the data file 105demography_data.csv and store the corresponding population data for every year from 1960 to 2017.