Applied Regression

Ilaria Battiston [*]

Winter Semester 2019-2020

[*] All notes are collected with the aid of material provided by D. Ankerst, K. Selig and the book *Introduction to Statistics and Data Analysis* by R. Peck, C. Olsen, J. Devore. All images have been retrieved by slides present on the Moodle Applied Regression Course.

Contents

1 Exploratory data analytics

1.1 Variables and frequency

A *variable* is any characteristic whose value may change from one observation to another. Making observations on variables results in data, which can be classified as:

- Univariate dataset, with a single variable;
- Bivariate dataset, with two variables;
- Multivariate dataset, with two or more variables.

Data are called categorical, qualitative or nominal if the individual observations belong in one of several possible groups.

Categorical data are ordinal if those groups can be ordered (from the biggest to the smallest).

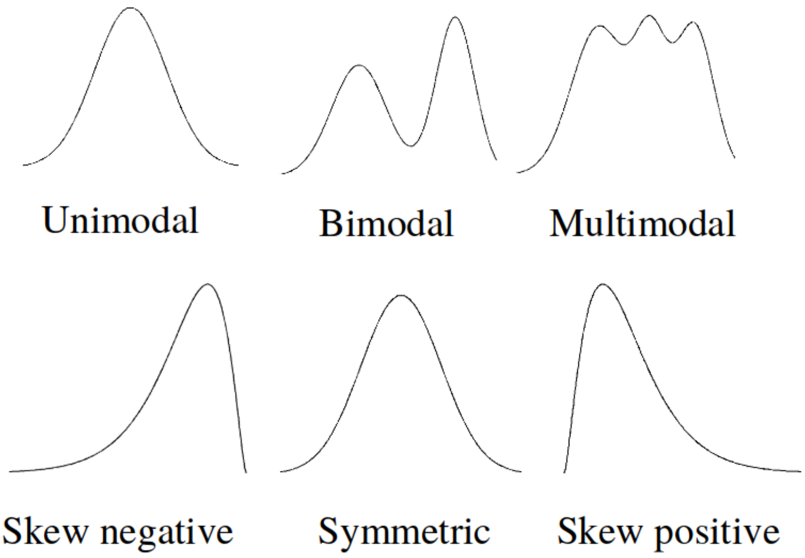
Data are numerical, quantitative or metric if the individual observations are real-valued numbers where numerical operations can be performed: they are classified in discrete and continue.

À frequency distribution for categorical data displays the possible categories along with their frequency or relative frequencies (number of times the category appears in the dataset). The relative frequency is the proportion between a frequency and the total observations.

The cumulative frequency table displays the proportion of values falling below the upper end of each interval, calculated summing the previous relative frequencies with the current one.

1.2 Charts

A bar chart is a graph of bars, with heights equal to either the frequencies or the relative frequencies. Its analog for continuous data (usually grouped in intervals)is the histogram.



A pie chart uses the familiar notion of a slice of pie to compare variable frequencies.

The dotplot is an alternative to the barplot used with continuous or discrete numerical data, showing each of the values along with their location. It is useful to highlight the most common values with their spread and outliers, allowing quick comparisons between categorical groups.

Steam and leaf plots work well with small datasets, showing consecutively increasing stems and one leaf for each number with the same stem.

1.3 Medians, quartiles

Those values can be obtained after a preprocessing, ordering the n observations from smallest to largest:

- Median:
  - If n is odd, the middle value;
  - If n is even, the mean of the two middle values;
- Lower quartile, median of the lower half of the data;
- Upper quartile, median of the upper half of the data;
- Interquartile range (iqr), calculated as the difference between upper and lower quartile.

If an observation is more than 1.5 iqr away from the closest end of the interval, it is defined an outlier. An outlier is extreme if it falls further than 3 iqr, otherwise mild.

A boxplot represents outliers with shaded or open circles, according to their type, and whiskers extending toward observations which are not outliers.

2 Mean and variance

The sample mean is calculated as:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n}$$

It doesn't always consist in an accurate representation of the dataset: it can be broadly affected by outliers. Visualising the data is useful for a deeper insight.

Outliers can be removed since they have a great impact while not being descriptive of the data. The trimmed mean is the mean of a subset of the ordered values, excluding the extreme ones (trimming percentage).

The range identifies variability of data, using the difference between the largest and smallest value. This is defined deviation, and it is summarized using sample variance and standard deviation.

$$s^2 = \frac{(x - \bar{x})^2}{n - 1}$$

$$s = \sqrt{s^2} = \sqrt{\frac{(x - \bar{x})^2}{n - 1}}$$

$s^2 \geq 0$ and becomes larger as data becomes further from the mean. Outliers have an extreme impact on variance and standard deviation.

The standardization operation allows to extract Z-score from data, explaining how many standard deviations the observation is far from the mean:

$$Z = \frac{x_i - \bar{x}}{s}$$

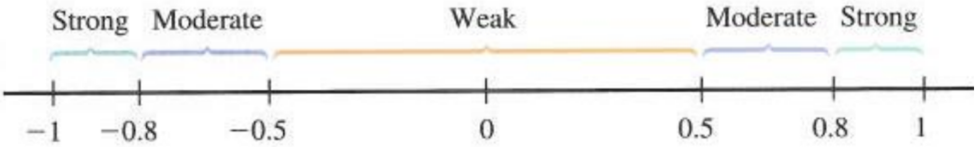
3 Bivariate data

When data depends on two dimensions, it can be represented using a scatterplot. Points within the two axes might have a linear relationship, named correlation and calculated with the Pearson coefficient:

$$r = \frac{1}{n-1} \sum \left(\frac{x - \bar{x}}{s_x}\right)\left(\frac{y - \bar{y}}{s_y}\right)$$

Properties of r:

- Does not depend on unit of measurement or axis selection;
- Always between +1 and -1;
- Not appropriate for non-linear relationships.



A linear relation has equation of $y = a = bx$ where b is the slope and a is the intercept. A measure for the goodness of fit of a line to bivariate data is the least squares criterion, using the vertical distance:

$$\sum \left[y - (a + bx)^2\right]$$

The smallest distance is found setting the partial derivatives to 0 and solving the equation.

Steps in linear regression:

1. Determining the explanatory and response variable;
2. Looking at the scatterplot to find any potential linear relationship;
3. Checking for unusual patterns;
4. Observing predictions and their accuracy.

The predicted or fitted values result from substituting each sample x value into the equation for the least squares line.

Residuals are the values $y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n$, and correspond to the vertical difference. Plotting those results can be explicative for potential problems or under-prediction.

Another indicator consists in the total sum of squares $\sum(y - \bar{y})^2$ and measures the distance from the horizontal line.

Coefficient of determination:

$$r^2 = 1 - \frac{\sum y^2 - \frac{(\sum y)^2}{n}}{\sum y^2 - a\sum y - b\sum xy} = \frac{\text{total sum of squares}}{\text{reidual sum of squares}}$$

It ranges between 0 and 1, higher values represent a better prediction. r^2 is the percent of variation in y that can be explained by x.

The standard deviation about the least squares line is denoted s_e and interpreted as the typical amount by which an observation deviates from the least squares line.

$$s_e = \sqrt{\frac{\text{SSResid}}{n-2}}$$

If the scatterplot does not look linear, it is possible to apply transformations to the variables (such as logarithm) to obtain a better fit.

Since observations occurring over time are not often independent, using a time series plot with time on the x-axis can be uninformative.

4 Simple Regression

The motivation of application of regression is seeing whether there is an association on trend of one variable with another. For instance, checking if the death rate due to melanoma changes according to the amount of sunshine the skin is exposed to.

The considered variables in this case are for instance state, mortality, latitude and longitude, population, bordering the ocean (dummy variable). Plotting mortality along with latitude highlights a relationship where the mortality decreases while latitude increases, but this association does not apply perfectly to all states.

Looking at the outliers is useful to identify potential errors or other factors which influence values. Some fields have not the possibility to repeat an experiment, but linear regression is a good method to have a simple assessment of magnitude of the relationship and variability around the regression line.

The linear relationship between x and y has the form of $y = \beta_0 + \beta_1 x$, where y is the dependent variable and β_0 and β_1 are respectively intercept and slope.

In other words, β_0 is the value of y when $x = 0$ and β_1 is the **change** in y for an **unit increase** in x. In this example, β_1 is the change of mortality as latitude increases of one point.

The regression line is estimated with the least squares criterion: given data (x_i, y_i) with $i = 1, 2, \ldots, n$, it chooses the values of b_0 and b_1 to minimize

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \qquad \hat{y} - b_0 = b_1 x_1$$

The solution is obtained setting the partial derivatives to 0 and solving to obtain the sample means of x and y:

$$\sum_{i=1}^{n}(y_i - (b_0 = b_1 x_i))^2 = (y_1 - (b_0 = b_1 x_1))^2 + \cdots + (y_n - (b_0 = b_1 x_n))^2$$

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{sample mean of } x \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{sample mean of } y \qquad b_0 = \bar{y} - b_1\bar{x}$$

The regression line always passes through $\bar{x}$, $\bar{y}$. Coefficients can be positive or negative.

4.1 Assumptions

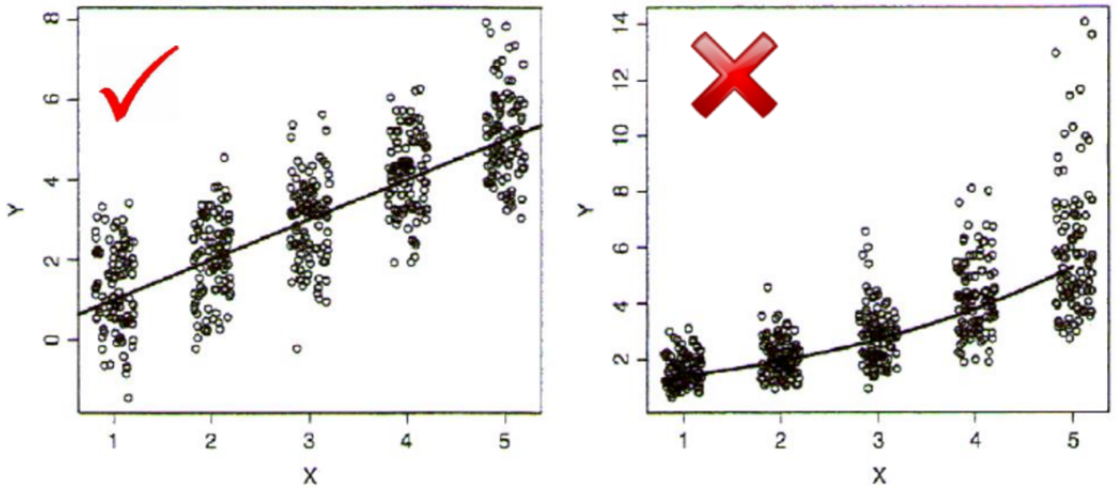
$$y = \beta_0 + \beta_1 x + \epsilon$$

Regression, although powerful, is not able to explain everything: it works assuming the random deviation ϵ has a normal distribution $\epsilon \sim N(0, \sigma^2)$ (has expected value $E(\epsilon) = 0$). Furthermore, there are other assumptions to be made:

- Independence of the y;
- Linearity of the mean of y in terms of x;
- Homogeneity of variance of y for each x;
- Normal distribution of y for each x.

Independence depends on how data were collected, and can be quantitatively checked for autocorrelation. It is usually stated at the moment of collection.

All the other constraints can be checked plotting raw data along with its variance.



The expectation is linear ($\bar{x}$ is constant):

$$E[a + by] = E(a) + E(by) = a + bE(y)$$

Furthermore, the following properties are applicable while making proofs:

- $V(aY) = a^2V(y)$;

- $V(a) = 0$;
- $E[y_1 + y_2] = E[y_1] + E[y_2]$ with y_1, y_2 random;
- $V[y_1 + y_2] = V[y_1] + V[y_2] + 2Cov[y_1, y_2] = V[y_1] + V[y_2]$ if y_1, y_2 are independent;
- $V[y] = E[y^2] - [E(y)]^2$.

Given an independent set of observations $()x_i, y_i)$, $i = 1, \dots, n$ it is possible to assume that each follows the regression model

$$y_i = \beta_0 + \beta_1 x_i = \epsilon_i \qquad \epsilon_i \sim N(0, \sigma^2)$$

Then an estimate of σ^2 is $s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$, where $n - 2$ represents the degrees of freedom (2 estimated parameters, the intercept and the slope).

4.2 Maximum likelihood

Parameters are chosen making the probability of the observed data as large as possible, maximizing the likelihood. The obtained values are the same as the least squares estimates, since maximizing L is equivalent to minimizing the sum of squares.

$$\hat{\sigma}^2_{mle} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Instead of maximum likelihood, it is used the proposed estimate s^2 with $\frac{1}{n-2}$, since it is unbiased if the assumptions are true.

The standard error should have the same scale of the data, therefore the square root is applied.

The true value of the slope can be assessed through a t-test, assuming that the variable equals to 0 (no correlation between x and y).