Introduction to regression:

Consider a bi (or multi)-variate data set

\[(Y_1, X_1), (Y_2, X_2) \ldots (Y_n, X_n)\]

obtained measuring two (or more) quantitative variables on n units selected at random from the population of interest.

i.e. an iid sample from the joint distribution of \((Y, X)\)

--notice this is a set of pairs (multiplets), one for each unit--

We can study the data from \(Y\) and \(X\) (X’s) separately (univariate analysis)

but we might be interested in the dependence of

\[Y = \text{response on } X \text{ (X’s) = explanatory(ies) or predictor(s)}\]

As we will see, \(Y\) will be usually taken to contain a normal term (error). So \(Y\) is usually assumed to be a continuous quantitative variable. From now on (for a while), we will consider a single \(X\).

Is there a dependence? What form does it take? Can we summarize it numerically and/or graphically? Can we adopt and estimate a functional form for the dependence?

In full generality, regression concerns the study of conditional distributions.

Does the conditional distribution of \(Y \mid X\) vary as a function of \(X\), and how?

In other words, if we divided the original population in subpopulations based on the value of \(X\), would \(Y\) behave differently within each subpopulation, and how so?

We want to extract from the sample data useful information on this.
The most frequently studied traits of the conditional distribution are (at the population level)

- conditional mean $E(Y|X)$; mean function --as a function of the value assumed by $X$ it expresses how the mean of $Y$ changes across $X$-subpopulations
- conditional variance $\text{var}(Y|X)$; variance function --as a function of the value assumed by $X$ it expresses how the variance of $Y$ changes across $X$-subpopulations

but $Y$ cold depend on $X$ only (... or also) through other traits --moments of order higher than 2, median or other quantiles, etc.

If there are several replicates for each observed $X$-value in the sample, or dividing the range of $X$ in classes $a_j - a_{j+1}$, we can calculate “conditional” numerical and graphical summaries.

Under the assumption that as $X$ varies within each class $Y|X$ does not vary (or varies very little with respect to inter-class variation):

$$\text{Av}(Y|a_0 < X \leq a_1), \text{Av}(Y| a_1 < X \leq a_2), \ldots$$

constitutes a “step” estimate of the mean function

and

$$s^2(Y|a_0 < X \leq a_1), s^2(Y| a_1 < X \leq a_2), \ldots$$

constitutes a “step” estimate of the variance function

(...fairly constant variance)
Also, collection of “conditional” box-plots

and a collection of “conditional” histograms --dividing also the range of Y in classes

(in all the above graphical representations one refers to the observed X-values with several replicates, or to the mid-points of the X-classes).

Another typical numerical summary:

\[
c(Y,X) = \frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \text{Av}(Y))(X_i - \text{Av}(X)) \quad --\text{sample covariance}
\]

\[
r(Y,X) = \frac{c(Y,X)}{s(Y) s(X)} = \frac{\sum_{i=1}^{n} (Y_i - \text{Av}(Y))(X_i - \text{Av}(X))}{\sqrt{\sum_{i=1}^{n} (Y_i - \text{Av}(Y))^2 \sum_{i=1}^{n} (X_i - \text{Av}(X))^2}}
\]

sample correlation coefficient, which estimates the population correlation coefficient \(\rho_{Y,X} = \frac{\sigma_{Y,X}}{\sigma_Y \sigma_X}\)

i.e. the covariance rescaled by the square roots of the variances to be a number in \([-1,+1]\).

This expresses strength and sign of the linear component of the relationship between Y and X.

It is important to remark that non-linear components of the relationship will not be captured by the correlation coefficient.
Another typical graphical summary: Scatter plot of $Y$ (vertical axis) vs $X$ (horizontal axis)

Two pictorial examples of samples with “0” correlation coefficient, while $Y$ clearly depends on $X$

- Data along a parabola with no linear component
- Data in two clusters with centers having no vertical component -- different radius
(Simple; one explanatory) Regression models:

Adopting and estimating a functional form for the dependence of Y on X.

Important remark:

We do not try to treat Y and X equally. In particular, the values of X are often taken as known or fixed (i.e. not random) on the units. This is rigorously the case

- in experimental settings, where a collection of values of X (e.g. the amount of a fertilizer per plot) are chosen for the units (plots), and the response Y (e.g. crop yield per plot) subsequently measured on them
- in survey settings in which sampling, and therefore measurement of Y (e.g. yearly disposable income per household) on the units (households), occurs within a chosen collection of X-subpopulations (e.g. number of individuals in the household)

but in very many survey setting sampling is not “stratified” on a given collection of X-subpopulations... rather, Y and X are both random, i.e. not controled for at the outset and simultaneously observed on the units. Here, taking the values of X as given is only a convenient ex-post expedient.

Consider the very general additive form:

\[ Y = f(X) + e \]

Y is equal to a function of X, plus an error term. The mean and variance functions will be

\[ E(Y|X) = f(X) + E(e|X), \quad \text{var}(Y|X) = \text{var}(e|X) \]

In terms of sample data, we will have

\[ Y_i = f(X_i) + e_i, \quad i=1...n \]

where \( e_i, \quad i=1...n \) are the errors associated with each sample unit.

Standard assumptions:

- The errors are identically distributed across sample units. Since units differ in terms of X, this means that the distribution of e does not depend on X, i.e. that the conditional distribution of e|X is the same for all X-subpopulations. This implies E(e|X) = E(e) = \( \xi \) and \( \text{var}(e|X) = \text{var}(e) = \sigma^2 \) both constant. Therefore, as X varies, f(X) models the mean fct. up to a constant, and the variance fct. is constant.

\[ E(Y|X) = f(X) + \xi \]
\[ \text{var}(Y|X) = \text{var}(e|X) = \sigma^2 \]
• **The common error mean is** $E(e) = \xi = 0$ (the errors have no “systematic” component to them). This implies that $f(X)$ models the mean fct. completely

$$E(Y|X) = f(X)$$

• $f(X)$ can be expressed in a parsimonious, parameterizable, and easily understandable way. The obvious starting point is a linear form

$$f(X) = \beta_0 + \beta_1X$$

![linear form in X](image)

- $\beta_0$ = intercept. Intersection with vertical axis. Pos is above orgin, neg if below, 0 if through origin.
- $\beta_1$ = slope. Inclination wrt the horizontal axis. Pos if less than 90dg, neg if more, 0 if parallel to horizontal axis.

• **The errors are independent across sample units** (recall this implies uncorrelation).

In summary: $Y = \beta_0 + \beta_1X + e$, with $e$ indep. of $X$, $E(e)=0$, var($e$)=$\sigma^2$

In terms of sample data: $Y_i = f(X_i) + e_i$, $i=1...n$, with $e_i$ identical ~ $e$ and independent

Another standard assumption further specifying the nature of the errors:

• **The common error distribution is normal** --with mean 0 and variance $\sigma^2$ from the above: $e \sim N(0, \sigma^2)$

Under normality: $Y = \beta_0 + \beta_1X + e$, with $e$ indep. of $X$, $e \sim N(0, \sigma^2)$

In terms of sample data: $Y_i = f(X_i) + e_i$, $i=1...n$, with $e_i$ iid from $e \sim N(0, \sigma^2)$