INSTITUTE OF ACTUARIAL & QUANTITATIVE STUDIES

Class: MSc

Subject : Probability and Statistics -2

Chapter: Unit 4 Chapter 1 - Part 1

Chapter Name: Correlation analysis and regression

Index

1 Introduction

- Actuaries, statisticians and many other professionals are increasingly engaged in analysing and interpreting large data sets, in order to determine whether there is any relationship between variables, and to assess the strength of that relationship. The methods in this and the following three chapters are perhaps more widely applied than any other statistical methods.

- Exploratory data analysis (EDA) is the process of analysing data to gain further insight into the nature of the data, its patterns and relationships between the variables, before any formal statistical techniques are applied.

1 Introduction

- Exploratory data analysis can be used to:
  - detect any errors (outliers or anomalies) in the data
  - check the assumptions made by any models or statistical tests
  - identify the most important/influential variables
  - develop parsimonious models – that is models that explain the data with the minimum number of variables necessary

1 Introduction

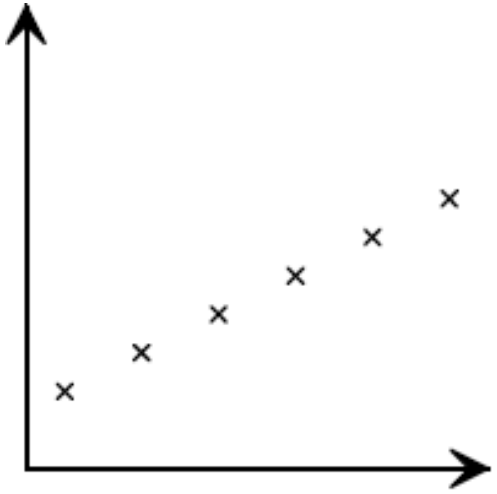
- For numerical data, this process will include the calculation of summary statistics and the use of data visualisations. Transformation of the original data may be necessary as part of this process.

- For a single variable, EDA will involve calculating summary statistics (such as mean, median, quartiles, standard deviation, IQR and skewness) and drawing suitable diagrams (such as histograms, boxplots, quantile-quantile (Q-Q) plots and a line chart for time series/ordered data).

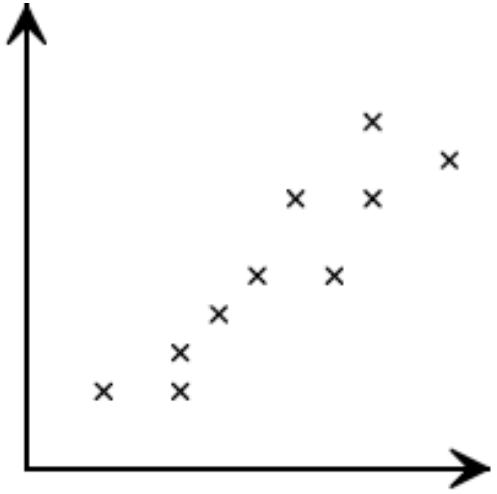
- For bivariate or multivariate data, EDA will involve calculating the summary statistics for each variable and calculating correlation coefficients between each pair of variables. Data visualisation will typically involve scatterplots between each pair of variables.

1 Introduction

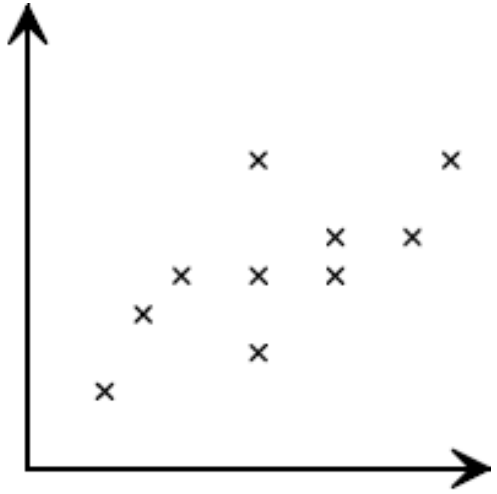
- Linear correlation between a pair of variables looks at the strength of the linear relationship between them. The diagrams below show the various degrees of positive correlation:



perfect positive correlation



strong positive correlation



weak positive correlation

2 Bivariate correlation analysis

- In a bivariate correlation analysis the problem of interest is an assessment of the strength of the relationship between the two variables Y and X .

- In any analysis, it is assumed that measurements (or counts) have been made, and are available, on the variables, giving us bivariate data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

2 Bivariate correlation analysis

Data Visualisation

- The starting point is always to visualise the data. For bivariate data, the simplest way to do this is to draw a scatterplot and get a feel for the relationship (if any) between the variables as revealed/suggested by the data.

- We are particularly interested in whether there is a linear relationship between Y , the response (or dependent) variable, and X , the explanatory (or independent, or regressor) variable. That is the expected value of Y , for any given value x of X , is a linear function of that value x, ie:

$$\boldsymbol{E[Y|x] = \alpha + \beta x}$$

- If a linear relationship (even a weak one) is indicated by the data, the methods of Linear Regression can be used to fit a linear model, with a view to exploiting the relationship between the variables to help estimate the expected response for a given value of the explanatory variable.

3 Sample correlation coefficients

- The degree of association between the x and y values is summarised by the value of an appropriate correlation coefficient each of which take values from –1 to +1.

- The coefficient of linear correlation provides a measure of how well a linear regression model explains the relationship between two variables. The values of r can be interpreted as follows:

Value	Interpretation
$r=1$	The two variables move together in the same direction in a perfect linear relationship.
$0<r<1$	The two variables tend to move together in the same direction but there is not a direct relationship.
$r=0$	The two variables can move in either direction and show no linear relationship.
$-1<r<0$	The two variables tend to move together in opposite directions but there is not a direct relationship.
$r=-1$	The two variables move together in opposite directions in a perfect linear relationship.

3 Sample correlation coefficients

- There are three broadly used correlation coefficients: Pearson, Spearman's rank and Kendall's rank.
- It is always important in data analysis to note that simply finding a mathematical relationship between variables tells one nothing in itself about the causality of that relationship or its continuing persistence through time. Qualitative as well as quantitative analysis is essential before making predictions or taking action.
- Jumping to a 'cause and effect' conclusion — that a change in one variable causes a change in the other — is a common misinterpretation of correlation coefficients. For example, the correlation may be spurious, or there may be another variable not part of the analysis that is causal.

4 Pearson's correlation coefficient

- Pearson's correlation coefficient r (also called Pearson's product-moment correlation coefficient) measures the strength of linear relationship between two variables and is given by:

$$\boldsymbol{r = \frac{S_{xy}}{\sqrt{S_{xx}\ \mathbf{x}\ S_{yy}}}}$$

- where:
  - $S_{xx} = \sum(x_i - \bar{x})^2 = \sum x_i^2 - (\sum x_i)^2/n$
  - $S_{yy} = \sum(y_i - \bar{y})^2 = \sum y_i^2 - (\sum y_i)^2/n$
  - $S_{xy} = \sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$

- Note that S_{xx} and S_{yy} , the sum of squares of x and y respectively, are the sample variances of x and y except we don't divide by (n -1) . Similarly S_{xy} is the sample covariance except we don't divide by n.

5 Spearman's rank correlation coefficient

- Spearman's rank correlation coefficient r_s measures the strength of monotonic (but not necessarily linear) relationship between two variables.
- So we are measuring how much they move together but the changes are not necessarily at a constant rate.
- Formally, it is the Pearson correlation coefficient applied to the ranks, $r(X_i)$ and $r(Y_i)$, rather than the raw values, (X_i, Y_i), of the bivariate data.
- So it just uses their relative sizes in relation to each other. We usually order them from smallest to largest.
- If all the X_i 's are unique, and separately all of the Y_i 's are unique, ie there are no 'ties', then this calculation simplifies to:

$$\boldsymbol{r_s = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}}$$

- where $d_i = r(X_i) - r(Y_i)$

6 Kendall rank correlation coefficient

- Kendall's rank correlation coefficient τ measures the strength of dependence of rank correlation between two variables.

- Like the Spearman rank correlation coefficient, the Kendall rank correlation coefficient considers only the relative values of the bivariate data, and not their actual values. It is far more intensive from a calculation viewpoint, however, since it considers the relative values of all possible pairs of bivariate data, not simply the rank of X_i and Y_i for a given i.

- Any pair of observations (X_i , Y_i); (X_j , Y_j) where $i \neq j$, is said to be concordant if the ranks for both elements agree, ie $X_i > X_j$ and $Y_i > Y_j$, or $X_i < X_j$ and $Y_i < Y_j$; otherwise they are said to be discordant.
- Let n_c be the number of concordant pairs, and let n_d be the number of discordant pairs. Assuming that there are no ties, the Kendall coefficient τ is defined as:

$$\tau = \frac{n_c - n_d}{n(n-1)/2}$$

- The numerator is the difference in the number of concordant and discordant pairs. The denominator is the total number of combinations of pairing each (X_i , Y_i) with each (X_j , Y_j). This could also be defined as $n_c + n_d$.

- So τ can be interpreted as the difference between the probability of these objects being in the same order and the probability of these objects being in a different order.

7 Inference

- To go further than a mere description/summary of the data, a model is required for the distribution of the underlying variables (X,Y).

Inference under Pearson's correlation

- The appropriate model is this: the distribution of (X,Y) is bivariate normal, with parameters $\mu_X, \mu_Y, \sigma_X, \sigma_Y$ and ρ.

- To assess the significance of any calculated r, the sampling distribution of this statistic is needed. The distribution of r is negatively skewed and has high spread/variability.

7 Inference

Result 1

- Under $\boldsymbol{H_0: \rho = 0}$, $\frac{\boldsymbol{r\sqrt{n-2}}}{\boldsymbol{\sqrt{1-r^2}}}$ has a t distribution with $\nu = n - 2$ degrees of freedom.
- From this result a test of $H_0: \rho = 0$ (the hypothesis of 'no linear relationship' between the variables) can be performed by working out the value of r which is 'significant' at a given level of testing, or by finding the probability value of the observed r.

Result 2 (Fisher's transformation of r)

- This is a more general result – it is not restricted to the case $\rho = 0$.
- If $\boldsymbol{W = \frac{1}{2}\ln\frac{1+r}{1-r}}$, then W has (approximately) a normal distribution with mean $\boldsymbol{\frac{1}{2}\ln\frac{1+\rho}{1-\rho}}$ and standard deviation $\boldsymbol{\frac{1}{\sqrt{n-3}}}$.
- This is usually referred to as the Fisher Z transformation (because the resulting z-values are approximately normal). Accordingly, the letter Z is usually used.
- From the result on W, tests of $H_0: \rho = \rho_0$ can be performed. Confidence intervals for μ_w and hence for ρ can also be found.

7 Inference

Notes:

(a) The bivariate normal assumption.

- The presence of ‘outliers’ – data points far away from the main body of the data – may indicate that the distributional assumption underlying the above methods is highly questionable.

(b) Influence

- Just as a single observation can have a marked effect on the value of a sample mean and standard deviation, so a single observation separated from the bulk of the data can have a marked effect on the value of a sample correlation coefficient.

7 Inference

Inference under Spearman's rank correlation

- Since we are using ranks rather than the actual data, no assumption is needed about the distribution of X , Y or (X,Y) , ie it is a non-parametric test.
- Under a null hypothesis of no association/no monotonic relationship between X and Y the sampling distribution of r_s can (for small values of n) be determined precisely using permutations. This does not have the form of a common statistical distribution.
- For larger values of n (>20) we can use Results 1 and 2 from before. The limiting normal distribution will have a mean 0 and a variance of 1/(n-1).
- Recall that Spearman's rank correlation coefficient is derived by applying Pearson's correlation coefficient to the ranks rather than the original data.

7 Inference

Inference under Kendall's rank correlation

- Again, since we are using ranks, we have a non-parametric test.
- Under the null hypothesis of independence of X and Y , the sampling distribution of τ can be determined precisely using permutations for small values of n.
- We can carry out a hypothesis test in the same way as described above but calculating n_c - n_d for each arrangement. However, again, for large n this will be time consuming.
- For larger values of n (>10), use of the Central Limit Theorem means that an approximate normal distribution can be used, with mean 0 and variance 2(2n+5)/9n(n-1).

8 Multivariate correlation analysis

- So far, we have only considered bivariate data. In most practical applications, there are many variables to consider. We now consider the case ($\underline{X}$,Y) , where Y remains the variable of interest, but $\underline{X}$ is now a vector of possible explanatory variables.

Data visualisation

- Again, the starting point is always to visualise the data. For multivariate cases it is no bother for a computer package to plot a scattergraph matrix, ie scattergraphs between each pair of variables to make the relationships between them clear.

Sample correlation coefficient matrix.

- Similarly it is no bother for a computer package to calculate correlation coefficients between each pair of variables and display them in a matrix.

Inference

- We can carry out tests on the correlation for each pair of variables using the method of Principal Component Analysis (PCA).

9 Principal component analysis

- Until now we have considered the variables in separate pairs, but in practice the amount of analysis required in this approach grows exponentially with each additional variable.

- Principal component analysis (PCA), also called factor analysis, provides a method for reducing the dimensionality of the data set, $\underline{X}$ – in other words, it seeks to identify the key components necessary to model and understand the data.

- For many multivariate datasets there is correlation between each of the variables. This means there is some 'overlap' between the information that each of the variables provide. The technical phrase is that there is redundancy in the data. PCA gives us a process to remove this overlap.

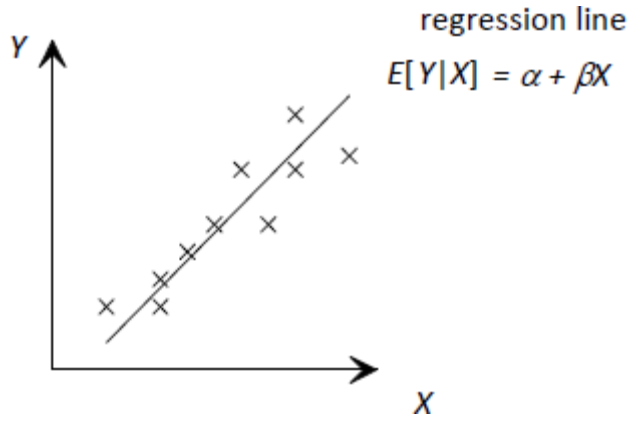
- The idea is that we create new uncorrelated variables, and we should find that only some of these new variables are needed to explain most of the variability observed in the data. The key thing is that each 'new' variable is a linear combination of the 'old' variables, so if we eliminate any of the new variables we are still retaining the most important bits of information.

9 Principal component analysis

- We then rewrite the data in terms of these new variables, which are called principle components.
- These components are chosen to be uncorrelated linear combinations of the variables of the data which maximise the variance.

10 Regression

- If there is a suitably strong enough correlation between the two variables (and there is cause and effect) we can justifiably calculate a 'regression line' which gives the mathematical form of this relationship:



10 Regression

- Regression analysis is used to assess the nature of the relationship between Y , the response (or dependent) variable, and X , the explanatory (or independent, or regressor) variable(s).
- The values of the response variable (our principal variable of interest) depend on, or are, in part, explained by, the values of the other variable(s), which is referred to as the explanatory variable(s).
- Ideally, the values used for the explanatory variable(s) are controlled by the experimenter — (in the analysis they are in fact assumed to be error-free constants, as opposed to random variables with distributions).
- Regression analysis consists of choosing and fitting an appropriate model — usually with a view to estimating the mean response (ie the mean value of the response variable) for specified values of the explanatory variable(s). A prediction of the value of an individual response may also be needed.
- Let's consider linear relationships for which we assume that the expected value of Y , for any given value x of X , is a linear function of that value x . For the bivariate case this simplifies to:

$$E[Y|x\} = \alpha + \beta x$$

- As always, before selecting and fitting a model, the data must be examined (eg in scatterplots) to see which types of model (and model assumptions) may or may not be reasonable.

10 Regression

- As always, before selecting and fitting a model, the data must be examined (eg in scatterplots) to see which types of model (and model assumptions) may or may not be reasonable.

- If a non-linear relationship (or no relationship) between the variables is indicated by the data, then the methods of analysis discussed here are not applicable for the data as they stand. However a well-chosen transformation of y (or x , or even both) may bring the data into a form for which these methods are applicable.