Chapter 3
2D Analysis: Correlation and Visualization of Two Features

3.1 General

Analysis of two features on the same entity set can be of interest assuming that the features are related in such a way that certain changes in one of them tend to co-occur with changes in the other. Then the relation – if observed indeed – can be used in various ways, of which two types of application are typically discernible: those oriented at

(i) prediction of values of one variable from those of the other;
(ii) addition of the relation to the knowledge of the domain by interpreting and explaining it in terms of the existing knowledge.

Goal (ii) is a subject in the discipline of knowledge bases as part of the so-called inferential approach, in which all relations are assumed to have been expressed as logical predicates and treated within a formal logic system – this approach will not be described here. We concentrate on another approach, referred to as the inductive one and related to the analysis of what type of information the data can provide with respect to the goals (i) and (ii). Typically, the feature whose values are to be predicted is referred to as the target variable and the other as the input variable. Examples of goal (i) are: prediction of an intrusion attack of a certain type (Intrusion data) or prediction of exam mark (Student data) or prediction of the number of Primary schools in a town whose population is known (Market town data). One may ask: why bother – all numbers are already in the file! Indeed, they are. But in the prediction problem, the data at hand are just a sample from a large population so that it is used as a training ground for devising a decision rule for prediction of the target feature at other, yet unobserved, entities. Typically, the input feature is readily available while the target feature is not. As to the goal (ii), the data usually are just idle empirical facts not necessarily noticeable unless they are generalized into a decision rule.

The mathematical structure and the visual portrayal of the problem differ depending on the type of feature scales involved, which leads us to considering all possible cases (see also Lohninger 1999):

B. Mirkin, *Core Concepts in Data Analysis: Summarization, Correlation and Visualization*, Undergraduate Topics in Computer Science,
DOI 10.1007/978-0-85729-287-2_3, © Springer-Verlag London Limited 2011

(1) both features are quantitative,
(2) one feature is quantitative, the other categorical, and
(3) both features are categorical.

3.2 Two Quantitative Features Case

P3.2.1 Scatter-Plot, Linear Regression and Correlation Coefficients

In the case when both features are quantitative, the three following concepts are popular: scatter plot, correlation and regression. We consider them in turn by using two features from the Market towns dataset, Population Resident and Number of Primary Schools. The data are taken from Table 1.4 (see below an extract for four towns out of 45):

	Pop (x)	PSchools (y)	(x,y)-point
Tavistock	10,222	5	(10,222,5)
Bodmin	12,553	5	(12,553,5)
Saltash	14,139	4	(14,139,4)
Brixham	15,865	7	(15,865,7)

Scatter plot is a presentation of entities as 2D points in the plane of two pre-specified features. On the left-hand side of Fig. 3.1, a scatter-plot of Market town features Pop (Axis *x*) and PSchools (Axis *y*) is presented.

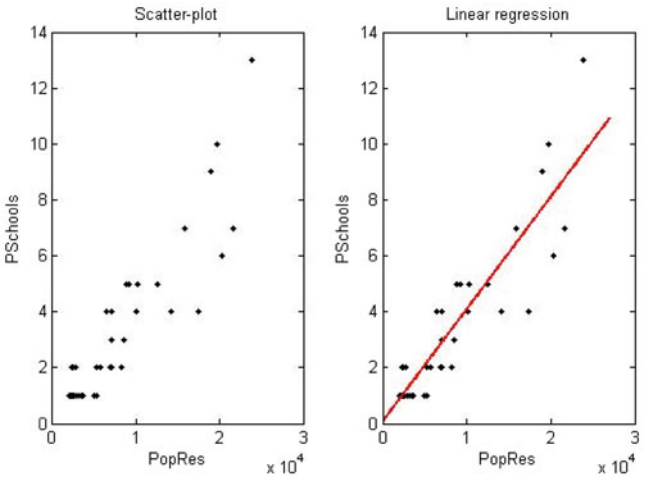

Fig. 3.1 Scatter plot of PopRes versus PSchools in Market town data. The *right hand graph* includes a regression line of PSchools over PopRes

One can think that these two features are related by a linear equation $y = ax+b$ where a and b are some constant coefficients, referred to as the slope and intercept, respectively, because the number of schools should be related to the number of children which is related to the number of residents. This equation is referred to as the linear regression of y over x. Obviously, most relations are not necessarily that simple because they also depend on other factors such as school sizes, population's age, etc. It would be a miracle if one equation fitted well all 45 towns. The possible inconsistencies in the equation can be modeled as additive errors, or residuals. The slope a and intercept b are taken in such a way that the inconsistencies of the equation on the 45 towns are minimized.

When a linear regression equation is fitted, its validity should be checked. A valid equation can be used for both (i) prediction and (ii) description.

The Galton-Pearson theory of linear regression involves a useful and very popular parameter, the correlation coefficient that shows the extent of linearity in the relation between the two features. Its square, referred to as the determination coefficient, can be used for a quick check of the validity of the regression: it shows the proportion of the variance of y that is taken into account by the regression. The correlation coefficient between the two features, Pop and PSchools, is 0.909. The correlation coefficient, in general, ranges between -1 and 1, and a value close to 1 or -1 indicates a high extent of the linear dependence between the features. In physics or chemistry, a high value of the correlation coefficient is rather usual; in social sciences, rather not – that is, the current features are highly related indeed.

Most other features in Market town data – such as the numbers of Post offices or Doctors – are also highly related to Pop feature, but not the number of Farmers markets. This latter feature appears to be binary here: a town either has a farmers market or not. The low value of the correlation coefficient, just below 0.15, shows that the size of the town does not much matter in this part of the world: a farmers market is as likely in a small town as it is in a larger town.

A low or even zero value of the correlation coefficient does not necessarily mean "no relation at all", but rather just "no *linear* relation". A zero correlation coefficient may hide a different type of functional relation, as shown on Fig. 3.2, which presents three different cases of the zero correlation. Only one of these, that on the left, case is genuine – there is no relation between x and y according to the picture indeed. Each of the other two cases relates to a rather high association between x and y. Specifically, the figure in the middle refers to a quadratic dependence and the figure on the right, to a split between two subsamples of highly linear but inverse relations.

Then the regression equation, estimated according to formulas (3.4–3.6) in Section 3.2.3.2, is this:

$$\text{PSchool} = 0.401^{*}\text{Pop} + 0.072 \tag{3.1}$$

where Population resident (Pop) is expressed in thousands to make the slope the thousand times greater than it would be if population is expressed in the absolute numbers. The slope expresses how much target changes when the input changes

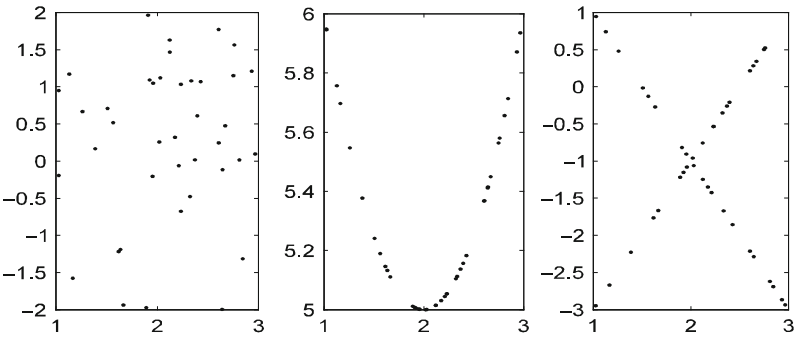

Fig. 3.2 Three scatter-plots corresponding to zero or almost zero correlation coefficient ρ; the case *on the left*: no relation between x and y; the case in the middle: a non-random quadratic relation $y = (x - 2)^2 + 5$; the case *on the right*: two symmetric linear relations, $y = 2x - 5$ and $y = -2x + 3$, each holding at a half of the entities

by 1. Because the target's values are integers, the value of slope can be rephrased as follows: the growth of population in a town by 2.5 thousand would lead, on average, to building one more primary school.

P3.2.2 Validity of the Regression

A regression function built over a data set should be validated. Three types of validity checks can be considered:

(a) The proportion of the variance of target variable taken into account by the regression, the determination coefficient: the greater the determination the better the fit.
(b) The confidence intervals of regression parameters – their ranges can give an idea of how stable the regression is.
(c) The direct testing of the accuracy of prediction both on data used for building the regression and data not used for that.

Worked example 3.1. Determination coefficient

Consider feature PSchools as target versus Pop as input, in Market Data (Fig. 3.1). The correlation coefficient between them is 0.909. The determination coefficient, in the case of linear regression, is its square, that is, $0.909^2 = 0.826$, which shows that the linear dependence on Pop decreases the variance of PSchools by 82.6%, a rather high value.

If the determination coefficient is not that high, still the hypothesis of linear relation may hold – depending on the distribution of residuals, that is, differences

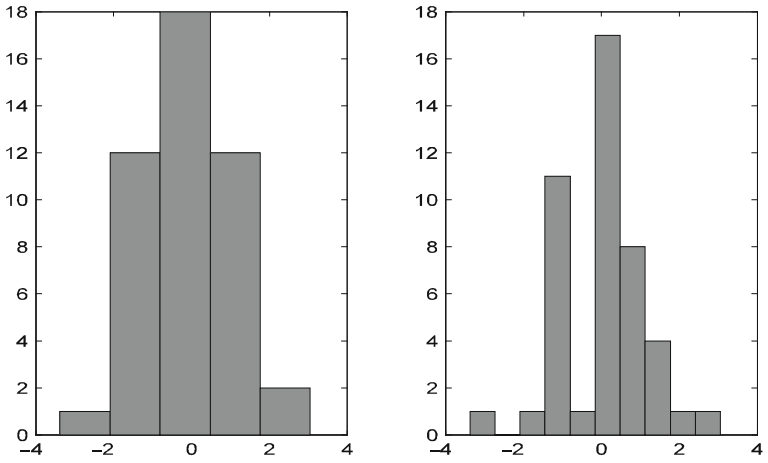

Fig. 3.3 Histograms of the residuals, the differences between values of PSchool as observed and those computed from Pop by using Equation (3.1), with 5 bins (*on the left*) and 10 bins (*on the right*). The dents in the finer histogram can be attributed to the fact that the sample of 45 instances is too small to have 10 bins

between the observed values of PSchool and those computed from Pop according to Equation (3.1). This distribution should be Gaussian or approximately Gaussian, so that the principle of maximum likelihood and formulas derived from it are appropriate. The distribution for the case under consideration is presented on Fig. 3.3. It is similar to a Gaussian distribution indeed, at the 5 bin histogram. The histogram with 10 bins is less so because it is somewhat dented – probably the sample is too small for this level of granularity: on average, only 4–5 entities fall in each of the bins.

A more straightforward validity test can be performed without any statistic theory at all – by purely computational means using the so-called bootstrapping which is a procedure for obtaining a multitude of random estimates of the parameters of interest by using random samples from the dataset as illustrated in Worked example 3.2.

Worked example 3.2. Bootstrap validity testing

Consider the linear regression of PSchools over PopRes in Equation (3.1) in the previous section. How stable are its slope and intercept regarding change of the sample? This can be tested by using bootstrap. One bootstrap trial involves three stages:

1. Randomly choose, with replacement, as many entities as there are in the sample – 45 in this case. Here is the sequence of indices of the entities randomly drawn

with replacement while writing this text: r = {26,17,36,11,29,39,32,25,27,26,29, 4,4,33,10,1,5,45,17,16,13,5,42,43,28,26,35,2,37,44,6,39,33,21,15,11,33,1,44,30, 26,25,5,37,24}. Some indices made it into the sample more than once, most notably 26 – four times, whereas many others did not make it into the sample at all – altogether, 16 entities such as 3,7,8 are absent from the sample. The proportion of the absent indices is 16/45= 0.356, which is rather close to the theoretic estimate 1/e=0.3679 derived in Project 2.3.

2. Take "resampled" versions of *Pop* and *PSchools* as their values on the elements drawn on step 1.
3. Find values of the slope and intercept for the resampled Pop and PSchools and store them.

The MatLab computation steps are similar to those in Project 3.1. After 400 trials the stored slopes and intercepts form distributions presented as 20 bin histograms on Fig. 3.4a, b respectively. After 4,000 trials, the respective histograms are c and d. One can easily see the smoothing effect of the increased number of trials on the histogram shapes – at 4,000 trials they do look Gaussian.

The bootstrapping trials give a diversity needed for estimating the average values of the slope and intercept. Moreover, one can draw confidence boundaries for the values.

How can one obtain, say, 95% confidence boundaries? According to the non-pivotal method, lower and upper 2.5% quantiles are cut out from the distribution in a symmetric way: 95% of the observations fall between the quantiles. For the case of 400 trials, 2.5% equals 10, so that the lower quantile corresponds to 11th and the upper quantile to 390th elements in the sorted set of values. For the case of 4,000 trials, 2.5% equals 100: these quantiles correspond to 101st and 3,900th elements of the sorted sets. They are shown in Table 3.1 at both of the cases, 400 and 4,000 trials. One can see that these provide consistent and rather tight boundaries for the slope: it is between 0.303 and 0.488 in 95% of all trials, according to 4,000-trial data, and

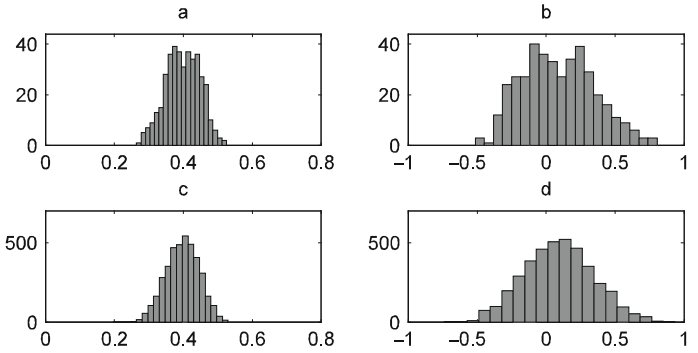

Fig. 3.4 Histograms of the distributions of the slope, *on the left* (**a**) and (**c**), and intercept, *on the right* (**b**) and (**d**), found at 400 (*on top*) and 4,000 (*below*) bootstrapping trials on PopResid, expressed in thousands, and PSchool features in Market town data

Table 3.1 Parameters of the linear regression of PopResid over PSchool found on the original set, as well as on 400 and 4,000 trials. The latter involves the average values as well as the lower and upper 2.5% quantiles

Regression		400 trials			4,000 trials		
Parameters	Set	Mean	2.5%	97.5%	Mean	2.5%	97.5%
Slope	0.401	0.399	0.296	0.486	0.398	0.303	0.488
Intercept	0.072	0.089	−0.343	0.623	0.092	−0.400	0.594

more or less the same at 400-trial data. The values of intercept are distributed with a greater dispersion and provide for a worsened accuracy. Symmetric 95% confidence intervals for the intercept are [−0.343,0.623] at 400 trials and [−0.400,0.594] at 4,000 trials.

Q.3.1. How a pivotal bootstrapping rule can be applied here? This would provide more stable evaluations than empirical distributions. The standard deviations of the slope and intercept are 0.0493 and 0.2606, respectively, at 400 boostrapping trials; they are somewhat smaller, 0.0477 and 0.2529, at the 4,000 trials. Can one derive from this a symmetric 95% confidence interval for the slope or intercept? Tip: in a Gaussian distribution, 95% of all values fall within interval mean ± 1.96*std. This is the so-called pivotal bootstrapping method.

Q.3.2. Can you give an estimate of the level of variance of the differences between PSchool observed and computed values?

A final validity test of the regression equation is probably the toughest one – by the prediction error (see Worked example 3.3).

Worked example 3.3. Prediction error of the regression equation

Compare the observed values of PSchool with those computed through Pop according to Equation (3.1). Table 3.2 presents a few examples taken from both ends of the sorted Pop feature.

Table 3.2 Observed numbers of Primary schools versus those predicted from the Population resident data on some Market towns

PS obs.	PS comp.	Pop	PS obse.	PS comp.	Pop
1	0.89	2,040	2	2.35	5,676
2	0.97	2,230	2	2.90	7,044
2	1.06	2,452	4	4.12	10,092
2	1.19	2,786	7	6.44	15,865
1	1.54	3,660	4	7.05	17,390

On average, the predictions are close, but, in some cases, are less so. One can easily estimate the relative error, which is [(1 − 0.89)/1]*100 = 11% at the first case, [(2 − 0.97)/2]*100 = 51.5% at the second case, etc. The average relative error of Equation (3.1) is equal to 30.7%. Can it be made smaller? On the first glance, no, it cannot, because Equation (3.1) minimizes the error. But, the error minimized by Equation (3.1) is the average quadratic error, not the relative error under consideration. The two errors do differ, and Equation (3.1) is not necessarily optimal with regard to the relative error.

The classical optimization theory has virtually nothing to propose for the minimization of the relative error – this criterion is neither linear, nor quadratic, nor convex. Yet the evolutionary optimization approach can be applied to the task. This approach uses a population of solutions randomly evolving, iteration after iteration, in the search for better solutions as explained in Project 3.2. Applying the algorithm from that project to minimize the criterion of relative error, one can find a different solution, in fact, a set of solutions each leading to the average relative error of 26.4%, a reduction of 4.3 points, one seventh of the relative error of Equation (3.1). The new solution is PSchool= 0.28*Pop + 0.33 expressing a smaller rate of increase in school numbers at the growth of population.

F3.2.3 Linear Regression: Formulation

F3.2.3.1 Fitting Linear Regression

Let us derive parameters of linear regression. Given target feature y and predictor x at N entities (x_1, y_1), $(x_2, y_2) \ldots$, (x_N, y_N), we are interested at finding a linear equation relating them so that

$$y = ax + b \tag{3.2}$$

The exact fit can occur only if all pairs (x_i, y_i) belong to the same straight line on *(x,y)*-plane, which is rather unlikely on real-world data. Therefore, Equation (3.2) will have an error at each pair (x_i, y_i) so that the equation should be rewritten as

$$y_i = ax_i + b + e_i \quad (i = 1, 2, \ldots, N) \tag{3.2'}$$

where e_i are referred to as errors or residuals. The problem is of determining the two parameters, a and b, in such a way that the residuals are least-squares minimized, that is, the average square error

$$L(a, b) = \Sigma_i e_i^2/N = \Sigma_i (y_i - ax_i - b)^2/N, \tag{3.3}$$

reaches its minimum over all possible a and b, given x_i and y_i $(i = 1, 2, \ldots, N)$. This minimization problem is easy to solve with the elementary calculus tools.

Indeed $L(a, b)$ is a "bottom down" parabolic function of a and b, so that its minimum corresponds to the point at which both partial derivatives of $L(a, b)$ are zero (the first-order optimality condition):

$$\partial L/\partial a = 0 \quad \text{and} \quad \partial L/\partial b = 0.$$

Leaving the task of actually finding the derivatives to the reader as an exercise, let us focus on the unique solution to the first-order optimality equations defined by the following formulas (3.4), for a, and (3.6), for b:

$$a = \rho\, \sigma(y)/\sigma(x) \tag{3.4}$$

where

$$\rho = \left[\Sigma_i (x_i - m_x)(y_i - m_v)\right] / \left[N\sigma(x)\sigma(y)\right] \tag{3.5}$$

is the so-called correlation coefficient, m_x, m_y are means of x_i, y_i, respectively and $\sigma^2(x), \sigma^2(y)$ are standard deviations;

$$b = m_v - am_x \tag{3.6}$$

By putting these optimal a and b into (3.3), one can express the minimum criterion value as

$$L_m(a, b) = \sigma^2(y)(1 - \rho^2) \tag{3.7}$$

The Equation (3.2) is referred to as the linear regression of y over x, index ρ in (3.4) and (3.5) as the correlation coefficient, its square ρ^2 in (3.7) as the determination coefficient, and the minimum criterion value L_m in (3.7) is referred to as the unexplained variance.

F3.2.3.2 Correlation Coefficient and Its Properties

The meaning of the coefficients of correlation and determination, in the data recovery framework of data analysis, is provided by Eqs. (3.3), (3.4), (3.5), (3.6) and (3.7). Here are some formulations.

Property 1 Determination coefficient ρ^2 shows the relative decrease of the variance of y after its linear relation to x has been taken into account by the regression (follows from (3.7)).

Property 2 Correlation coefficient ρ ranges between -1 and 1, because ρ^2 is between 0 and 1, as follows from the fact that value L_m in (3.7) cannot be negative because the items in its expression (3.3) are all squares. The closer ρ to either 1 or

-1, the smaller are the residuals in the regression equation. For example, $\rho = 0.9$ implies that y's unexplained variance L_m is $1 - \rho^2 = 19\%$ of the original value.

Property 3 The slope a is proportional to ρ according to (3.4); a is positive or negative depending on the sign of ρ. If $\rho = 0$, the slope is 0: in this case, y and x are referred to as not correlated.

Property 4 The correlation coefficient ρ does not change under shifting and rescaling of x and/or y, which can be seen from Equation (3.5). Its formula (3.5) becomes especially simple if the so-called z-scoring has been applied to standardize both x and y.

To perform z-scoring over a feature, its mean m is subtracted from all the values and the results are divided by the standard deviation σ:

$$x_i' = (x_i - m_x)/\sigma(x) \quad \text{and} \quad y_i' = (y_i - m_v)/\sigma(y), \quad i = 1, 2, \ldots, N$$

Using the z-score standardization, formula (3.5) can be rewritten as

$$\rho = \Sigma_i x_i' y_i' / = <x', y'>/N \tag{3.5'}$$

where $<x', y'>$ denotes the inner product of vectors $x' = (x_i')$ and $y' = (y_i')$.

The next property refers to one of the fundamental discoveries by K. Pearson, an interpretation of the correlation coefficient in terms of the bivariate Gaussian distribution. A generic formula for the density function of this distribution, in the case in which the features have been pre-processed by using z-score standardization described above, is

$$f(u, \Sigma) = C \exp\left\{-u^T \Sigma^{-1} u/2\right\} \tag{3.8}$$

where $u = (x, y)$ is a two-dimensional vector of the two variables x and y under consideration and Σ is the so-called correlation matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

In formula (3.8), ρ is a parameter with a very clear geometric meaning. Consider a set of points $u = (x, y)$ on (x, y)–plane making function $f(u, \Sigma)$ in (3.8) equal to a pre-specified constant. Such a set makes the values of $u^T \Sigma u$ constant too. That means that a constant density set of points $u = (x, y)$ must satisfy equation $x^2 - 2\rho xy + y^2 = \text{const}$. This equation is known to define a well-known quadratic curve, the ellipsis. At $\rho = 0$ the equation becomes an equation of a circle, $x^2 + y^2 = const$, and the greater the difference between ρ and 0, the more skewed is the ellipsis, so that at $\rho = \pm 1$ the ellipsis becomes a bisector line $y = \pm x + b$ because the left part of the equation makes a full square, in this case, $x^2 \pm 2xy + y^2 = const$, that is,

$(y \pm x)^2 = const$. The size of the ellipsis is proportional to the constant: the greater the constant the greater the size.

Property 5 The correlation coefficient (3.5) is a sample based estimate of the parameter ρ in the Gaussian density function (3.8) under the conventional assumption that the sample points (y_i, x_i) are drawn from a Gaussian population randomly and independently.

This striking fact is behind a long standing controversy. Some say that the usage of the correlation coefficient is justified only when the sample is taken from a Gaussian distribution, because the coefficient has a clear-cut meaning only in this model. This logic seems somewhat overly restrictive. True, the usage of the coefficient for estimating the density function is justified only when the function is Gaussian. However, when trying to linearly represent one variable through the other, the coefficient has a very different meaning in the approximation context, which has nothing to do with the Gaussian distribution, as expressed above with Eqs. (3.4), (3.5), (3.6) and (3.7).

F3.2.3.3 Linearization of Non-linear Regression

Non-linear dependencies also can be fit by using the same criterion of minimizing the square error. Consider a popular case of exponential regression, that is, representing correlation between target y and predictor x as $y = ae^{bx}$ where a and b are unknown constants and e the base of natural logarithm. Given some a and b, the average square error is calculated as

$$E = \left(\left[y_1 - a\exp(bx_1)\right]^2 + \ldots + \left[y_N - a\exp(bx_N)\right]^2\right)/N = \Sigma_i\left[y_i - a\exp(bx_i)\right]^2/N \tag{3.9}$$

There is no method that would straightforwardly lead to a globally optimal solution of the problem of minimization of E in (3.9) because it is too complex function of the unknown values. This is why conventionally the exponential regression is fit by what should be referred to as its linearization: transforming the original problem to that of linear regression. Indeed, let us take the logarithm of both parts of the equation that we want to fit, $y = ae^{bx}$. The resulting equation is $\ln(y) = \ln(a) + bx$. This equation has the format of linear equation, $z = \alpha x + \beta$, where $z = \ln(y)$, $\alpha = b$ and $\beta = \ln(a)$. This leads to the following idea. Let us take the target be $z = \ln(y)$ with its values $z_i = \ln(y_i)$. By fitting the linear regression equation with data x_i and z_i, one finds optimal α and β, so that the original exponential parameters are found as $a = \exp(\beta)$ and $b = \alpha$. These values do not necessarily minimize (3.9), but the hope is that they are close to the optimum anyway. Unfortunately, this may be very wrong sometimes as the material in Project 3.2 clearly demonstrates.

Q.3.3. Find the derivatives of L over a and b and solve the first-order optimality conditions.

Q.3.4. Derive the optimal value of L in (3.7) for the optimal a and b.

Q.3.5. Prove or find a proof in the literature that any linear equation $y = ax+b$ corresponds to a straight line on Cartesian xy plane for which a is the slope and b intercept.

Q.3.6. Find the inverse matrix Σ^{-1} for $\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. **A.** $\Sigma^{-1} = \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \Big/ (1-\rho^2)$.

C3.2.4 Linear Regression: Computation

Regression is a technique for representing the correlation between x and y as a linear function (that is, a straight line on the plot), $y = slope^*x + intercept$ where *slope* and *intercept* are constants, the former expressing the change in y when x is added by 1 and the latter the level of y at x=0. The best possible values of slope and intercept (that is, those minimizing the average square difference between real y's and those found as $slope^*x + intercept$) are expressed in MatLab, according to formulas (3.4), (3.5) and (3.6), as follows:

```
≫ rho = corrcoef(x,y);
%2×2 matrix whose off-diagonal entry is correlation coefficient
≫ slope = rho(1,2)*std(y)/std(x);
≫ intercept = mean(y) − slope*mean(x);
```

Here rho(1, 2) is the Pearson correlation coefficient between x and y (3.5) that can be determined with MatLab operation "corrcoef" which leads to an estimate of the matrix Σ above.

Project 3.1. 2D analysis, linear regression and bootstrapping

Let us take the Students data table as a 100×8 array a in MatLab, pick any two features of interest and plot entities as points on the Cartesian plane formed by the features. For instance, take Age as x and Computational Intelligence mark as y:

```
≫ x = a(:,4); % Age is 4-th column of array "a"
≫ y = a(:,8); % CI score is in 8-th column of "a"
```

Then student 1 (first row) will be presented by point with coordinates x = 28 and y = 90 corresponding to the student's age and CI mark, respectively. To plot them all, use command:

```
≫ plot(x,y,'k.')
% k refers to black colour, "." dot graphics; 'mp' stands for magenta pentagram;
% see others by using "help plot"
```

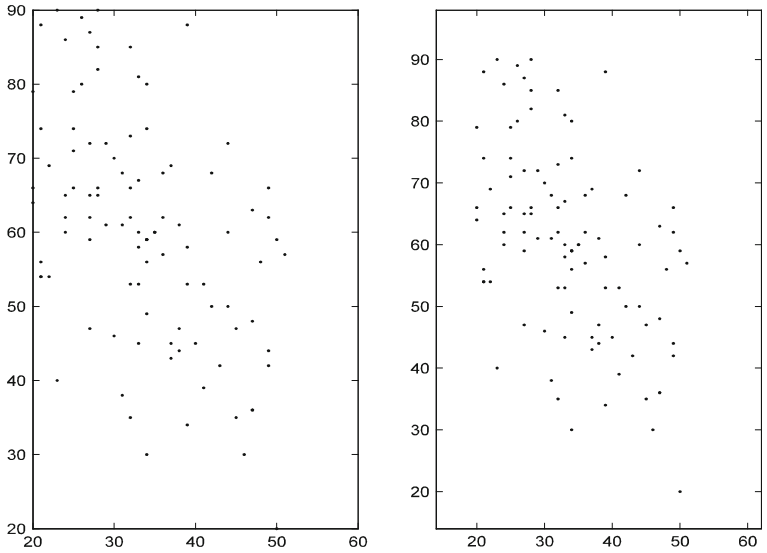

Fig. 3.5 Scatter plot of features "Age" and "CI score"; the display *on the right* is a rescaled version of that *on the left*

Unfortunately, this gives a very tight presentation: some points are on the borders of the drawing. To make the borders stretched out, one needs to change the axis, for example, as follows:

≫ d = axis; axis(1.2*d−10);

This transformation is presented on the right part of Fig. 3.5. To make both plots presented on the same figure, use "subplot" command of MatLab:

≫ subplot(1,2,1)
≫ plot(x,y,'k.');
≫ subplot(1,2,2)
≫ plot(x,y,'k.');
≫ d = axis; axis(1.2*d−10);

Whichever presentation is taken, no regularity can be seen on Fig. 3.5 at all. Let's try then whether anything better can be seen for different occupations. To do this, one needs to handle entity sets for each occupation separately:

≫ o1=find(a(:,1)==1); % set of indices for IT
≫ o2=find(a(:,2)==1); % set of indices for BA
≫ o3=find(a(:,3)==1); % set of indices for AN
≫ x1=x(o1);y1=y(o1); % the features x and y at IT students
≫ x2=x(o2);y2=y(o2); % the features at BA students
≫ x3=x(o3);y3=y(o3); % the features at AN students

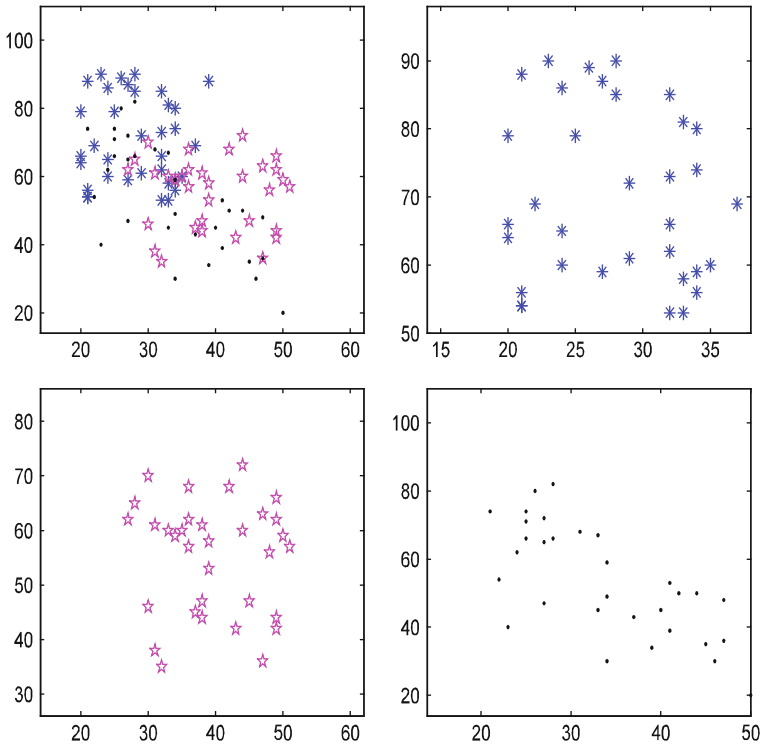

Fig. 3.6 Joint and individual displays of the scatter-plots at the occupation categories (IT *star*, BA *pentagrams*, AN *dots*)

Now we are in a position to put, first, all the three together, and then each of these three separately (again with the command "subplot", but this time with four windows organized in a two-by-two format, see Fig. 3.6).

≫ subplot(2,2,1); plot(x1,y1, '*b',x2,y2,'pm',x3,y3,'.k');% all three
≫ d=axis; axis(1.2*d−10);
≫ subplot(2,2,2); plot(x1,y1, '*b'); % IT plotted with blue stars
≫ d=axis; axis(1.2*d−10);
≫ subplot(2,2,3); plot(x2,y2,'pm'); % BA plotted with magenta pentagrams
≫ d=axis; axis(1.2*d−10);
≫ subplot(2,2,4); plot(x3,y3,'.k'); % AN plotted with black dots
≫ d=axis; axis(1.2*d−10);

Of the three occupation groups, some potential relation can be seen only in the AN group: it is likely that "the greater the age the lower the mark" regularity holds in this group (black dots in the Fig. 3.4's bottom right). To check this, let us utilize the linear regression.

Linear regression equation, $y = slope^*x + intercept$ is estimated by using MatLab, according to formulas (3.4), (3.5) and (3.6), as follows:

≫cc= corrcoef(x3,y3);rho=cc(1,2);% producing rho=−0.7082
≫ slope = rho*std(y3)/std(x3); % this produces slope =−1.33;
≫ intercept = mean(y3) − slope*mean(x3); % this produces intercept = 98.2;

Since we are interested in group AN only, we apply these commands at AN-related values x3 and y3 to produce the linear regression as $y3 = 98.2 - 1.33^*x3$. The slope value suggests that every year added to the age, in general decreases the mark by 1.33, so that aging by 3 years would lead to the loss of 4 mark points. Obviously, care should be taken to draw realistic conclusions.

Altogether, the regression equation explains $rho\hat{}2 = 0.50 = 50\%$ of the total variance of y3 – not too much, as is usual in social and human sciences.

Let us take a look at the reliability of the regression equation with bootstrapping, the popular computational experiment technique for validating data analysis results that was introduced in Project 2.3 (see also Carpenter and Bithell 2000, Davison and Hinkley 2005).

Bootstrapping is based on a pre-specified number of random trials, for instance, 5,000. Each trial consists of the following steps:

(i) randomly selecting an entity N times, with replacement, so that the same entity can be selected several times whereas some other entities may be never selected in a trial. (As shown above in Project 2.3, on average only 63% entities get selected into the sample.) A sample consists of N entities because this is the number of entities in the set under consideration. In our case, N=31. One can use the following MatLab command:

≫ N=31;ra=ceil(N*rand(N,1));
% rand(N,1) produces a column of N random real numbers, between 0 and 1 each.
% Multiplying this by N stretches them to (0,N) interval; ceil rounds the numbers up to integers.

(ii) the sample ra is assigned with their data values according to the original data table:

≫xt=xx(ra);yt=yy(ra);
% here xx and yy represent the predictor and target, respectively;
% they are x3 and y3, respectively, which can be taken into account with assignments
% xx=x3; and yy=y3.

so that coinciding entities get identical feature values.

(iii) a data analysis method under consideration, currently "linear regression", that basically computes the rho, the slope and the intercept, applies to this data sample to produce the trial result.

To do a number (5,000, in this case) of trials, one should run (i)–(iii) in a loop:

≫ for k=1:5000; ra=ceil(N*rand(N,1));
xt=xx(ra);yt=yy(ra);

```
    cc=corrcoef(xt,yt);
    rh(k)=cc(1,2);
    sl(k)=rh(k)*std(yt)/std(xt); inte(k)=mean(yt)−sl(k)*mean(xt);
end
% the results are 5000-strong columns rh (correlations), sl (slopes)
% and inte (intercepts)
```

Now we can check the mean and standard deviation of the obtained distributions. Commands

```
≫mean(sl); std(sl)
```

produce values −1.33 and 0.24. That means that the original value of slope=−1.33 is confirmed with the bootstrapping, but now we have obtained its standard deviation, 0.24, as well. Similarly mean/std values for the intercept and rho are computed. They are, respectively, 98.2 / 9.0 and −0.704 / 0.095.

We can plot the 5,000 values found as 30-bin histograms (see Fig. 3.7):

```
≫ subplot(1,2,1); hist(sl,30)
≫ subplot(1,2,2); hist(in,30)
```

Command subplot(1,2,1) creates one row consisting of two windows for plots and puts the follow-up plot into the first window (that on the left). Command subplot(1,2,2) changes the action into the second window which is on the right.

To derive the 95% confidence boundaries for the slope, intercept and correlation coefficient, one may use both pivotal and non-pivotal methods.

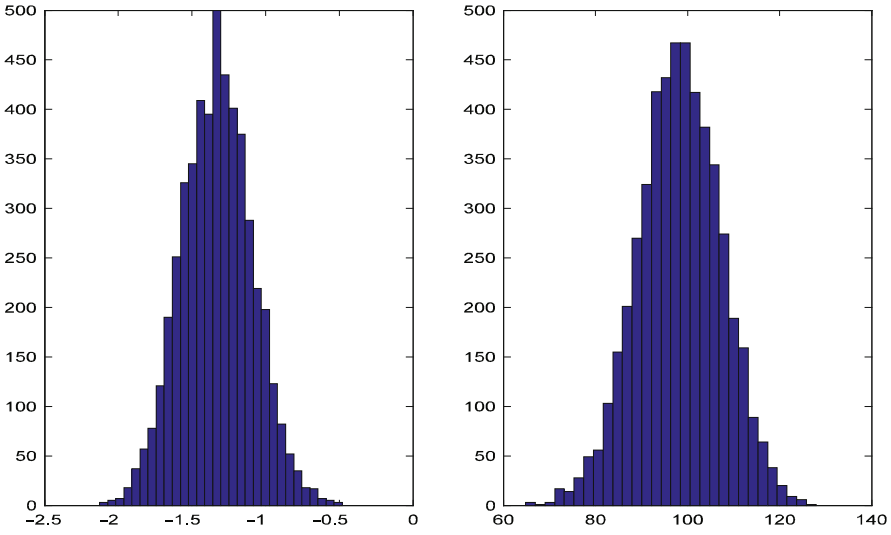

Fig. 3.7 30-bin histograms of the slope (*left*) and intercept (*right*) after 5,000 bootstrapping trials

The pivotal method uses the hypothesis that the bootstrap sample is indeed a random sample from a Gauusian distribution. Parameters of this distribution for slope are determined with the following commands:

≫ msl=mean(sl); ssl=std(sl);

Since 95% of the Gaussian distribution fall within interval of plus-minus 1.96*std, the 95% confidence boundaries are derived, for the slope, as follows:

≫ lbsl=msl − 1.96*ssl; rbsl=msl + 1.96*ssl

The non-pivotal estimates require no such a hypothesis and are based on the bootstrap distribution as is. One just sorts all the values and takes 2.5% quantiles on both extremes of the range:

≫ ssl=sort(sl); lbn=ssl(126);rbn=ssl(4875);

Indeed, we need to cut out 5% items from the sample, to make a 95% confidence interval. Since 5% of 5,000 is 250, conventionally divided in two halves, this requires cutting off first 125 observations as well as the last 125 observations of the presorted list of the bootstrap values, which brings us to ssl(126) and ssl(4875) as the non-pivotal boundaries for the slope value.

All these estimates are presented in Table 3.3. The pivotal and non-pivotal estimates do not fall too far apart. Either can be taken as parameters of the boundary regressions.

This all can be visualized by, first, defining the three regression lines, the regular one and two corresponding to the lower and upper estimate boundaries, respectively, with

≫ y3reg=slope*x3+intercept;
≫ y3regleft=lbsl*x3+lbintercept;
≫ y3regright=rbsl*x3+rbintercept;

and then plotting the four sets onto the same figure Fig. 3.8:

≫ plot(x3,y3,‘*k’,x3,y3reg,‘k’,x3,y3regleft,‘r’,x3,y3regright,‘r’)
% x3,y3,‘*k’ presents student data as black stars; x3,y3reg,‘k’ presents the
% real regression line in black
% x3,y3regleft,‘g’ and x3,y3regright,‘g’ for boundary regressions in green

The lines on Fig. 3.8 show the boundaries of the regression line for 95% of trials.

Table 3.3 Parameters of the bootstrap distributions and pivotal and non-pivotal boundaries

			Pivotal boundaries		Non-pivotal boundaries	
	Mean	St. dev.	Left	Right	Left	Right
Slope	−1.337	0.241	−1.809	−0.865	−1.800	−0.850
Intercept	98.51	9.048	80.776	116.244	80.411	116.041
Corr. coef.	−0.707	0.094	−0.891	−0.523	−0.861	−0.493

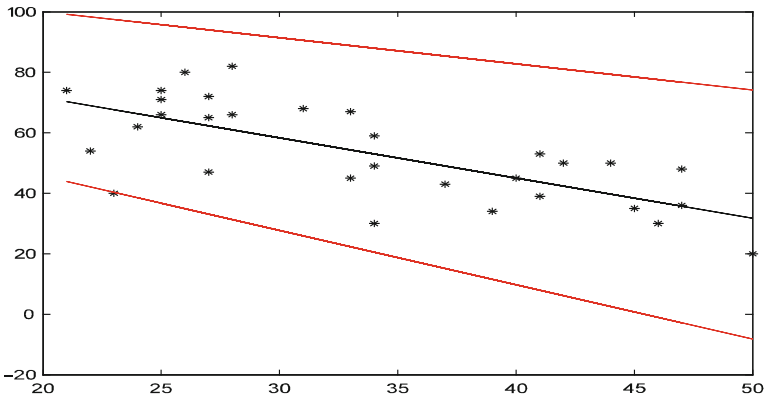

Fig. 3.8 Regression of CI score over Age (*black line*) within occupation category AN with boundaries covering 95% of potential biases due to sample fluctuations

Project 3.2. Non-linear and linearized regression: a nature-inspired algorithm

In many domains the correlation between features is not necessarily linear. For example, in economics, processes related to the inflation over time are modeled by using the exponential function. A similar way of thinking applies to the processes of growth in biology. Variables describing climatic conditions obviously have a cyclic character. The power law in social systems is nonlinear too.

Consider, for example, a power law function $y = ax^b$ where x is predictor and y predicted variables whereas a and b are unknown constant coefficients. Given the values of x_i and y_i on a number of observed entities $i = 1, \ldots, N$, the power law regression problem can be formulated as the problem of minimizing the summary squared or absolute error over all possible pairs of coefficients a and b. There is no method that would straightforwardly lead to a globally optimal solution of the problem because minimizing a sum of many exponents is a complex problem.

This is why conventionally the power law regression is fit by transforming it into a linear regression problem. Indeed, the equation of the power law regression, taken with no errors, is equivalent to the equation of linear regression with $\log(x)$ being predictor and $\log(y)$ target: $\log(y) = b\log(x) + \log(a)$. This gives rise to the very popular strategy of linearization of the problem. First, transform x_i and y_i to $v_i = \log(x_i)$ and $z_i = \log(y_i)$ and fit the linear regression equation for given v_i and z_i; then convert the found coefficients into those of the original exponential function. This strategy seems especially suitable since the logarithm of a variable typically is much smoother so that the linear fit is better under the logarithm transformation.

There is one caveat, however: the fact that found coefficients are optimal in the linear regression problem does not necessarily imply that the converted exponents are necessarily optimal in the original problem. This we are going to explore in this project.

Nature-inspired optimization is a computational intelligence approach to minimize a non-linear function. Rather than look and polish a single solution to the optimization problem under consideration, this approach utilizes a population of solutions iteratively evolving from generation to generation, according to rules imitating a real-world evolutionary process. The rules typically include: (a) random changes from generation to generation such as "mutations" and "crossovers" in earlier, genetic, algorithms, and (b) policies for selecting and maintaining the best found solutions, the "elite". After a pre-specified number of iterations, the best solution among those observed is reported as the outcome.

To start the evolutionary optimization process, one should first define a restricted area of admissible solutions so that no member of the population may leave the area. This warrants that the population will not explode by moving solutions to the infinity. Under the hypothesis of a power law relation $y = ab^x$, for any two entities i and j, the following equations should hold: $z_i = b^*v_i + c$ and $z_j = b^*v_j + c$ where $c = \log(a)$, $z_j = \log(y_i)$ and $v_i = \log(x_i)$. From these, b and c can be expressed as follows: $b = (z_i - z_j)/(v_i - v_j)$, $c = (v_i{}^*z_j - v_j{}^*z_i)/(v_i - v_j)$, which may lead to different values of b and c at different i and j. Denote bm and bM the minimum and the maximum of $(z_i - z_j)/(v_i - v_j)$, and cm and cM the minimum and maximum of $(v_i{}^*z_j - v_j{}^*z_i)/(v_i - v_j)$ over those i and j for which $v_i = v_j \neq 0$. One would expect that the admissible b and c should be within these boundaries, which means that the area of admissible solutions should be defined by the inequalities $(bm, cm) \leq (b, c) \leq (bM, cM)$. Since the optimal values of *(b,c)* should be around the averages of the ratios above, that is, lie deep inside the area between their maxima and minima, it helps to speed up the computation if one takes only those pairs *(i,j)* at which the values of v_i, v_j and z_i, z_j are not too close to 0 so that their logarithms are not that far away from 0, and, similarly, the differences between them should be neither that small nor that high. This approach is implemented in MatLab code ddr.m in Appendix A4.

For the step of producing the next generation, let us denote the population's $p \times 2$ array by f, at the current iteration, and by f', at the next iteration. The transition from f to f' is done in three steps. First, take the row of mean values within the columns of f and repeat it p times in a $p \times 2$ array mf. Then make a Gaussian random move:

$$\text{fn} = \text{f} + \text{randn(p,2).*mf/20} \tag{3.10}$$

Here randn(p,2) is a p×2 array of (pseudo) random numbers generated according to Gaussian distribution N(0,1) with 0 expectation and 1 variance. The symbol .* denotes the operation of multiplication of corresponding elements in matrices, so that (aij).*(bij) is a matrix whose *(i,j)*-th elements are products aij*bij. This random matrix is scaled down by *mf/20* so that the move accounts for about 5% (one twentieth) of the average f values.

Since the move is to be restricted within the admissibility area, any a-element (first column of fn) which is greater than aM, is to be changed for aM, and any a-element smaller than am is to be changed for am. Similar trimming applies to b-elements. Denote result by fr.

At the next step, take a $p \times 2$ array el whose rows are the same stored elite solution and arrive at the next generation f' by using the following "elite mix":

$$f' = 0.7fr + 0.3el$$

The elite mix moves all population members in the direction of the best solution found so far by 30%, which has been found work well in the examples of our interest.

This procedure is implemented in MatLab code nlr.m that relies on ddr.m at step 1 and a subroutine, delta, for evaluating the fitness (see A4 in Appendix).

Consider now this experiment. Generate predictor x as a 50-long vector of random positive entries between 0 and 10, x=10*rand(1,50), and define $y = 2^*x^{1.07}$ with the normal additive noise $2^*N(0, 1)$ where 0 is the mean and 1 the variance, which is suppressed when overly negative, according to the Matlab code line

≫for ii=1:50;yy=2*x(ii)^1.07 +2*randn;y(ii)=max(yy,1.01);end;

When using the conventional linearized regression model by linearly mapping log(x) to log(*y*), to extract b and a (as the exponent of the found c) from this, the program llr.m implementing this approach produces $a = 3.0843$ and $b = 0.8011$ leading to the averaged squared error $y - ax^b$ equal to 4.41, so that the standard error is 2.10, about 20% of the mean y value, 10.1168. It is not only that the error is high, but also a wrong law is identified. The generated function y stretches x out ($b>1$), whereas the found function stretches x in ($b<1$).

Minimization of the averaged squared error $y - ax^b$ of the original model directly by using the code nlr.m, that implements the nature-inspired algorithm, the values are $a = 2.0293$ and $b = 1.0760$ leading to the average squared error of 0.0003 and the standard error of 0.0180. In contrast to the values found at the linearized scheme, the parameters a and b here are very close to those generated.

This obviously considerably outperforms the conventional procedure. Similar results can be found at different values of the noise variance.

Case-study 3.1. Growth of Investment

Let us apply a similar approach to the following example involving variables x and y defined over a period of 20 time moments as presented in Table 3.4.

Table 3.4 Data of investment y at time moments x from 0.10 to 2.00

x	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	1.10	1.20	1.30	1.40	1.50
1.60	1.70	1.80	1.90	2.00											
y	1.30	1.82	2.03	4.29	3.30	3.90	3.84	4.24	4.23	6.50	6.93	7.23	7.91	9.27	9.45
11.18	12.48	12.51	15.40	15.91											

Variable x can be thought of as related to the time periods whereas y may represent the value of a fund. In fact, the components of x are numbers from 1 to 20 divided by 10, and y is obtained from them in MatLab according to formula y=2*exp(1.04*x)+0.6*randn where randn is the normal (Gaussian) random variable with the mathematical expectation 0 and variance 1.

Let us, first, try a conventional approach of finding the average growth of the fund during all the period.

The average growth of the investment according to these data is conventionally expressed as the root 19, or power 1/19, of the ratio y_{20}/y_{01}, that is, 1.14. This estimates the average growth as 14% per period – which is by far greater than 4% in the data generating model.

Let us now try to make sense of the relation between x and y by applying the conventional linearization strategy to this data.

The strategy of linearization of the exponential equation outlined in Section 3.2.3.3 leads to values 1.1969 and 0.4986 for b and c, respectively, to produce $a = e^c = 1.6465$ and $b = 1.1969$ according to formulas there. As one can see, these differ from the original $a = 2$ and $b = 1.04$ by the order of 15–20%. The value of the squared error here is E = 13.90. See Fig. 3.9 representing the data.

Let us now apply the nature inspired approach to the original non-linear least-squares problem.

The program nlrm.m implementing the evolutionary approach described in Project 3.2 found $a = 1.9908$ and $b = 1.0573$. These are within 1–2% of the error from the original values $a = 2$ and $b = 1.04$. The summary squared error here is E = 7.45, which is by far smaller than that found with the linearization strategy.

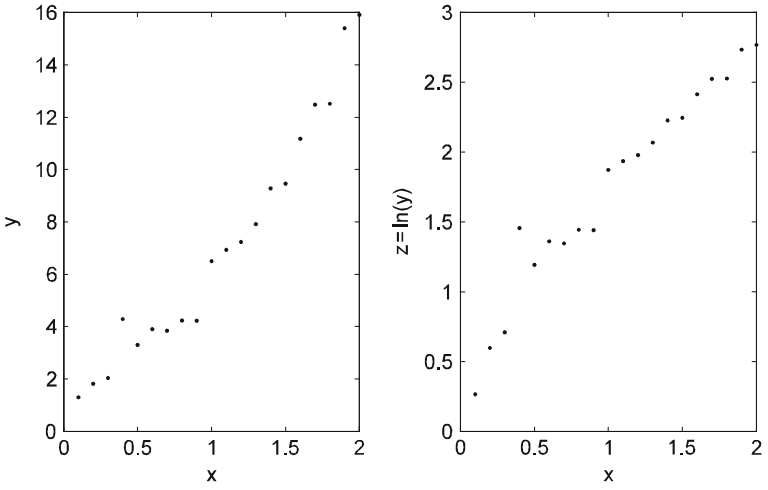

Fig. 3.9 Plot of the original pair (x,y) in which y is a noisy exponential function of x (*on the left*) and plot of the pair (x,z) in which $z = \ln(y)$. The plot *on the right* looks somewhat straighter indeed, though the correlation coefficients are rather similar, 0.970 for the plot *on the left* and 0.973 for the plot *on the right*

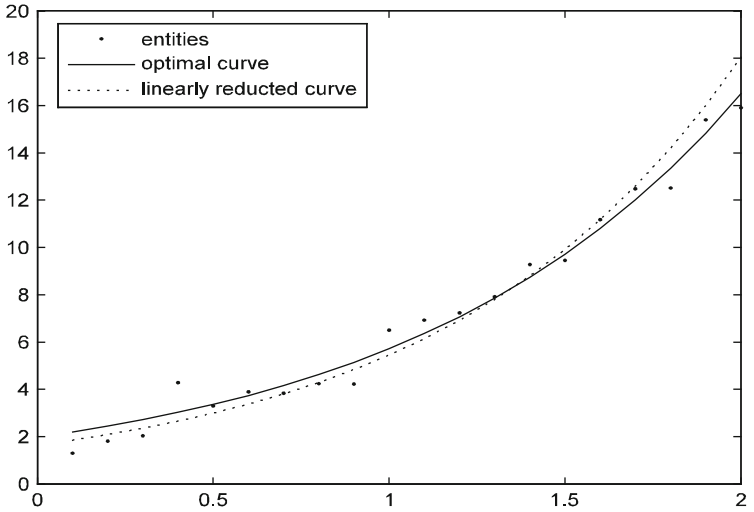

Fig. 3.10 Two fitting exponents are shown, with *stars and dots*, for the data in case study 3.1

The two found solutions can be represented on the scatter-plot graph, see Fig. 3.10. One can see that the linearized version has a much steeper exponent, which becomes visible at later periods.

Q.3.7. Consider a binary feature defined on seven entities so that it is category A on the first three of them, and category B on the next four. Let us draw two dummy 1/0 variables, xA and xB, corresponding to each so that xA = 1 on the first three entities and xA = 0 on the rest, whereas xB = 0 on the first three entities and xB = 1 on the rest. What can be said of the correlation coefficient between xA and xB? **A.** The correlation coefficient between xA and xB is −1 because xA+xB=1 for all entities so that xA = −xB+1.

Q.3.8. Extend the nature-inspired approach to the problem of fitting a linear regression with a nonconventional criterion such as the average relative error defined by formula $1/N \sum_{i=1}^{N} |e_i/y_i|$.

Case-study 3.2. Correlation Between Iris Sepal Length and Width

Take x and y from the Iris set in Table 1.3 as the Sepal's length and width, respectively.

A scatter plot of x and y is presented on the left part of Fig. 3.11. This is a loose cloud of points which looks similar to that on the left part of Fig. 3.2, of no correlation. Indeed the correlation coefficient value here is not only very small,

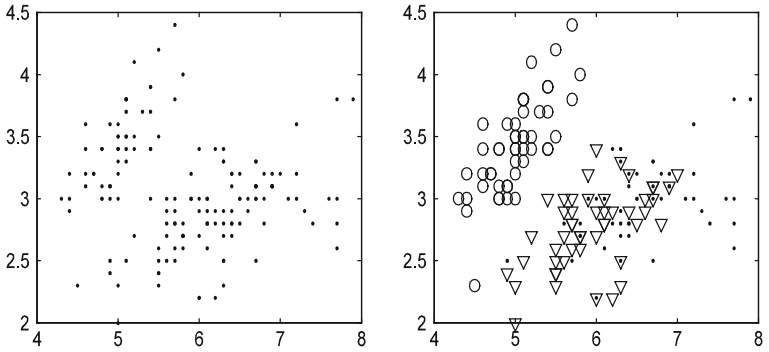

Fig. 3.11 Scatter plot of Sepal length and Sepal width from Iris data set (Table 1.3), as a whole *on the left* and taxon-wise *on the right*. Taxon 1 is presented by *circles*, taxon 2 by *triangles*, and taxon 3 by *dots*

−0.12, but also negative, which is somewhat odd, because intuitively the features should be positively correlated as reflecting the size of the same flower.

To see a particular reason for the low, and negative, correlation, one should take into account that the sample is not homogeneous: the Iris set consists of 50 specimens of each of three different taxa. When the taxa are separated (see Fig. 3.11 on the right), the positive correlation is restored. The correlation coefficients are 0.74, 0.53 and 0.46 for taxon one, two and three, respectively. Here is a nice example of the negative effect of the non-homogeneity of the sample on the data analysis results.

3.3 Mixed Scale Case: Nominal Feature Versus a Quantitative One

P3.3.1 Box-Plot, Tabular Regression and Correlation Ratio

Consider x a categorical feature on the same entities as a quantitative feature y, such as Occupation and Age at Students data set. The within-category distributions of y can be used to investigate the correlation between x and y. The distributions can be visualized by using just ranges as follows: present categories with equal-size bins on x axis, draw two lines parallel to x axis to present the minimum and maximum values of y (in the entire data set), and then present the within category ranges of y as shown on Fig. 3.12.

The correlation between x and y is higher when the within-category spreads are tighter because the tighter the spread within an x-category, the more precise is prediction of y at it. Figure 3.13 illustrates an ideal case of a perfect correlation – all within-category y-values are the same leading to an exact prediction of Age when Occupation is known.

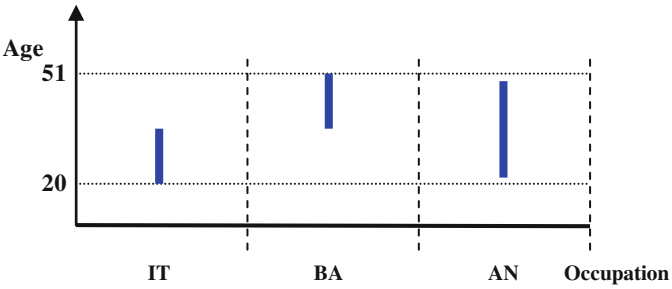

Fig. 3.12 Graphic presentation of within category ranges of Age at Student data

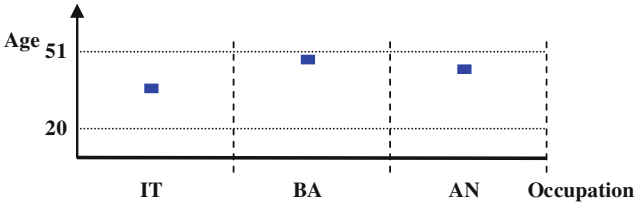

Fig. 3.13 In a situation of ideal correlation, with zero within-category variances, knowledge of the Occupation category would provide an exact prediction of the Age within it

Figure 3.14 presents another extreme, when knowledge of an Occupation category does not lead to a better prediction of Age than when the Occupation is unknown.

A simple statistical model extending that for the mean will be referred to as tabular regression. The tabular regression of quantitative y over categorical x is a table comprising three columns corresponding to:

(1) Category of x
(2) Within category mean of y
(3) Within category standard deviation of y

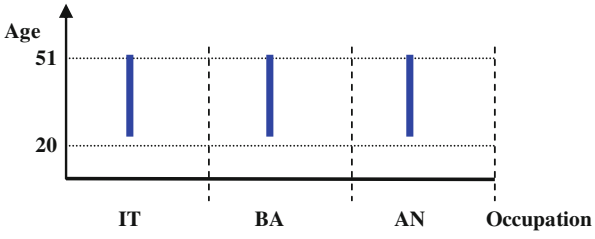

Fig. 3.14 Wide within-category distributions: the case of full variance within categories in which the knowledge of Occupation would give no information of Age

The number of rows in the tabular regression thus corresponds to the number of x-categories; there should be a marginal row as well, with the mean and standard deviation of y on the entire entity set.

Worked example 3.4. Tabular regression of Age (quantitative target) over Occupation (categorical predictor) in Students data

Let us draw a tabular regression of Age over Occupation in Table 3.5. The table suggests that if we know the Occupation category, say IT, then we can safely predict the Age as being 28.2 within the margin of plus/minus 5.6 years. With no knowledge of the Occupation category, we could only say that the Age is on average 33.7 plus/minus 8.5, a somewhat less precise estimate.

The table can be visualized in a manner similar to Figs. 3.12, 3.13 and 3.14, this time presenting the within category averages by horizontal lines and the standard deviations by vertical strips (see Fig. 3.15).

One more way of visualization of categorical/quantitative correlation is the so-called box-plot. The within-category spread is expressed here with a quantile (percentile) box rather than with the standard deviation. First, a quantile level should be defined such as, for instance, 40%, which means that we are going to show the within-category range over only 60% of its contents by removing 20% off of both its top and bottom extremes. These are presented with box' heights such as on Fig. 3.16; the full within-category ranges are shown with whiskers.

Table 3.5 Tabular regression of Age over Occupation in Students data

Occupation	Age Mean	Age StD
IT	28.2	5.6
BA	39.3	7.3
AN	33.7	8.7
Total	33.7	8.5

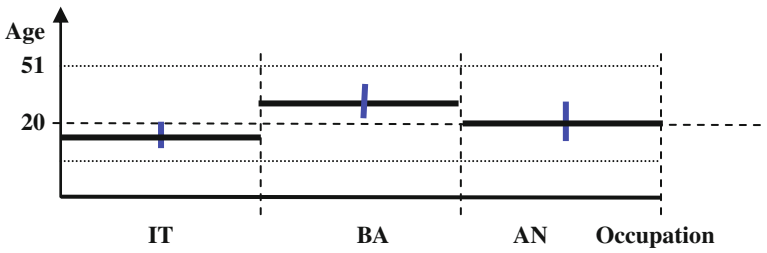

Fig. 3.15 Tabular regression visualized with the within-category averages and standard deviations represented by the position of *solid horizontal lines* and *vertical line sizes*, respectively. The *dashed line's* position represents the overall average (*grand mean*)

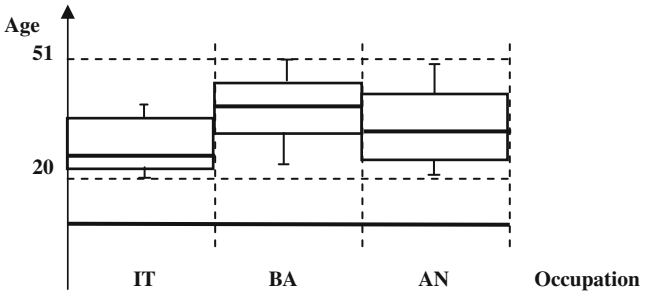

Fig. 3.16 Box-plot of the relationship between Occupation and Age with 20% quantiles; *the box heights* reflect the Age within-category 60% ranges, whiskers show the total ranges. Within-box *horizontal lines* show the within category averages

Worked example 3.5. Box-plot of Age at Occupation categories at Students data

With the quantile level specified at 40%, at the category IT, Age ranges between 20 and 39, but if we sort it and remove 7 entities of maximal Age and 7 entities of minimal Age (there are 35 students in IT so that 7 makes 20% exactly), then the Age range on the remaining 60% is from 22 to 33. Similarly, Age 60% range is from 32 to 47 on BA, and from 25 to 44 on AN (see box heights on Fig. 3.16). The whiskers reflect 100% within category ranges, which are intervals [20, 39], [27, 51] and [21, 50], respectively.

The box-plot proved useful in studies of quantitative features too: one of the features is partitioned into a number of bins that are treated then as categories.

Consider now one more tabular regression, this time of the OOProgramming mark over Occupation (Table 3.6)

A natural question emerges: In which of the tables the correlation is greater, 3.5 or 3.6?

This can be addressed with an integral characteristic of the tabular regression, the correlation ratio. This coefficient scores the extent at which the within group variance is smaller on average than the variance of the feature on the set before the split – a determination coefficient for the tabular regression.

Table 3.6 Tabular regression OOProg/Occupation

Occupation	OOP Mean	OOP StD
IT	76.1	12.9
BA	56.7	12.3
AN	50.7	12.4
Total	61.6	16.5

Worked example 3.6. Correlation ratio

Let us address the question above: Is the correlation in Table 3.5 is greater than in Table 3.6?

Correlation ratios for the tables computed by using formulas (3.14) and (3.12) are:

Occupation/Age	28.1%
Occupation/OOProg	42.3%

The drop in variance expressed by the correlation ratio is greater at the second table, that is, the correlation between Occupation and OOProgramming is greater than that between Occupation and Age.

Q.3.9. In Table 3.6, there is a positive relation between the Occupation and the OOP mark, with the largest mark, 76.1, going to IT and the smallest mark, 50.7, to AN. There is no such a relation in Table 3.5 in which AN's Age is in the middle between that at the other two groups. Is it that feature of Table 3.6 that leads to a higher correlation ratio? **A.** No; the order of means is irrelevant at the tabular regression. The correlation ratio is higher at Table 3.6 than at Table 3.5 because of the tighter boundaries on the quantitative feature within the groups in Table 3.6.

F3.3.2 Tabular Regression: Formulation

Given a quantitative feature y, with no further information, its average, $\bar{y} = \sum_{i \in I} y_i/|I|$, would represent a proper summarization of the data. If, however, a set of categories of another variable, x, is additionally present, a more detailed summarization can be provided: the within category averages. Let S_k denote the set of entities falling in k category of x, then the within-category averages are $\bar{y}_k = \sum_{i \in S_k} y_i/|S_k|$.

This can be considered the least-squares solution to the model of tabular regression which extends the data recovery model (2.5) for the mean on page 40 as follows. Find a set of c_k values such that the summary square error $L = \Sigma_{i \in I} e_i^2$ is minimized, where $e_i = y_i - c_k$ according to equations

$$y_i = c_k + e_i \text{ for all } i \in S_k \tag{3.11}$$

The equations underlie the tabular regression and are referred to sometimes as the piece-wise regression. It is not difficult to prove that the optimal c_k in (3.11) is the within category average $\bar{y}_k$, which implies that the minimum value of L is equal to $L_m = \sum_{k=1}^{K} \sum_{i \in S_k} (y_i - \bar{y}_k)^2$. By dividing and multiplying the interior sum by the number

of elements in S_k, $|S_k|$, we can see that in fact $L_m = N\sigma^2{}_w$ where $\sigma^2{}_w$ is the average within category variance defined as

$$\sigma_w^2 = \Sigma_k p_k \sigma_k^2 \tag{3.12}$$

where $p_k = |S_k|/N$ is the proportion of category k and $\sigma^2{}_k$ the variance of y within S_k.

To further analyze this, consider equation

$$(y_i - \bar{y}_k)^2 = y_i{}^2 + \bar{y}_k^2 - 2y_i\bar{y}_k$$

and sum it up over all $i \in S_k$. This would lead to the summary right-hand item being similar to that in the middle, thus producing $\sum_{i \in S_k} (y_i - \bar{y}_k)^2 = \sum_{i \in S_k} y_i{}^2 - |S_k|\bar{y}_k{}^2$. Summing these equations over k and moving the right-hand item to the other side of the equation, would lead to the following decomposition:

$$\sum_{i \in I} y_i{}^2 = \sum_{k=1}^{K} |S_k|\bar{y}_k^2 + \sum_{k=1}^{K} \sum_{i \in S_k} (y_i - \bar{y}_k)^2 \tag{3.13}$$

Note that the right-hand item in (3.13) is the summary least-squares criterion of model in (3.11) L_m. This allows us to interpret the Equation (3.13) as a decomposition of the scatter of variable y, the item on the left, in two parts on the right: the explained part, in the middle, and the unexplained part L_m.

The explained part sums contributions of individual categories k, $|S_k|\bar{y}_k^2$. The value of the contribution is proportional to both the category frequency and the squared value – the greater the better.

Another expression of decomposition (3.13) can be obtained under the assumption that variable y is centered, so that its mean is 0, by relating it to N:

$$\sigma^2 = \sum_{k=1}^{K} p_k \bar{y}_k^2 + \sum_{k=1}^{K} p_k \sigma_k^2 \tag{3.14}$$

where σ^2 is the variance of y, the item on the right the minimum value L_m/N from (3.12), and the item in the middle, the weighted summary squared distance between the grand mean $\bar{y}$=0 and within-category means $\bar{y}_k$.

Equation (3.14) is very popular in statistics as the decomposition of the variance into the within-group variance, the item on the right, and the between-group variance, the item in the middle, as the base of a popular method for comparison of within-category means which is referred to as ANOVA (ANalysis Of VAriance). In the context of the tabular regression model (3.11) viewed as a data recovery model, the original decomposition (3.13) of the quantitative feature scatter into part

explained by the nominal feature and part remaining unexplained is more appropriate, as will be seen later in Sections 4.4 and 6.3. Viewed in this light, decomposition (3.14) shows that the category k contribution to the total variance of y is proportional to its frequency and the squared difference between within-category mean $\bar{y}_k$and grand mean $\bar{y} = 0$.

The correlation ratio shows the relative drop in the variance of y when it is predicted according to model (3.11) or, in other words, the relative proportion of the explained part of the variance. Correlation ratio is usually denoted by η^2 and can be defined by the following formula:

$$\eta^2 = 1 - {\sigma^2}_w/\sigma^2 \tag{3.15}$$

The definition implies the following properties:

- The range of η^2 is between 0 and 1.
- Correlation ratio $\eta^2 = 1$ when all within-category variances ${\sigma^2}_k$ are zero (that is, when y is constant within each group S_k).
- Correlation ratio $\eta^2 = 0$ when all ${\sigma^2}_k$ are of the order of ${\sigma^2}_k$.

Q.3.10. Consider two quantitative features x and y. Divide the range of x in five equal-sized bins to produce a categorical variable xc. Is there any relation between the correlation coefficient between x and y and the correlation ratio coefficient between xc and y? **A.** None, the former can be greater than the latter in some cases, and smaller in some others.

3.3.3 Nominal Target

In the case when it is the quantitative variable that is predictor while the categorical variable is the target, one can use all the wealth of methods developed for pattern recognition or machine learning. The problem may be stated variously depending on the learning task. A machine learning task typically assumes a training dataset for deriving a rule that can be applied to entities from a testing dataset, under the assumption that structures of the training and testing datasets are similar – see a discussion in Chapter 4. All features under consideration are assumed known on both of the sets, except that the categories are not known on the testing dataset.

A most popular problem to address would be like this: given a value of the quantitative predictor on an entity, tell the category of the target feature on the entity. We present two approaches to this.

3.3.3.1 Nearest Neighbor Classifier

One of the most popular is the so-called Nearest-Neighbor classifier. It is applicable at any data admitting distances or (dis)similarities between entities. The NN classifier works as this: find, in the training dataset, an entity which is the nearest to that

under consideration and extrapolate its category to the entity in question. One can take a look at the results of application of the NN classification rule to two feature pairs, one from Intrusion data set, and the other from Student dataset, in the follow up examples. The results are very different – the former, in Table 3.7, is very successful whereas the other, in Table 3.9, not. An explanation to this is the difference in the strength of correlation between the two variables – very strong in one case and rather weak in the other (see Tables 3.8 and 3.10).

The NN classifier can be easily extended to the so-called k-NN classifier; the latter usually supplies the category supported by a majority of the k nearest neighbors of the entity in question. This classifier may also lead to the so-called "reject option" – giving no answer when there is no clear-cut majority.

Table 3.7 Applying NN classifier SH⇒Attack to a random subsample of the Intrusion dataset

Random sample	9	29	37	51	63	70	72	80	86	89
True target category	apa	nor	nor	nor	nor	nor	nor	sai	sai	sai
Predictor's value PV	24	10	1	14	2	3	1	482	482	483
Nearest Neighbor's PV	23	11	1	13	2	3	1	482	482	482
NN predicted category	apa	nor	nor	nor	nor	nor	nor	sai	sai	sai

Table 3.8 Tabular regression of SHCo over Attacks in Intrusion data: comparatively small within-category standard deviations

Attack	Number	Mean	Standard deviation
Apache	23	33.61	12.13
Saint	11	484.64	8.42
Smurf	10	508.40	5.13
Normal	56	5.13	5.59
Total	100	114.75	198.09
Correlation ratio	0.988		

Table 3.9 Applying NN classifier CI⇒Occupation to a random subsample of the Student dataset; wrong category assignments are highlighted in bold

Random sample	4	11	24	42	44	61	87	89	94	100
True target category	IT	IT	IT	BA	BA	BA	AN	AN	AN	AN
Predictor's value PV	72	65	54	65	44	62	72	48	34	45
Nearest Neighbor's PV	72	65	54*	65	44	62**	72	47	35*	45*
NN predicted category	**BA**	**AN**	IT	**AN**	BA	BA	**BA**	**BA**	AN	AN

* – of two other entities having different categories that with the matching one has been selected;
** – of several entities, the most frequent category has been selected.

Table 3.10 Tabular regression of CI mark over Occupation in Student data: comparatively high within-category standard deviations

Occupation	Number	Mean	Standard deviation
IT	35	70.57	12.73
BA	34	54.79	10.60
AN	31	53.35	16.29
Total	100	59.87	15.37
Correlation ratio	0.250		

Worked example 3.7. Nearest neighbor classifier

Consider two features from the dataset Intrusion: the type of attack Att, the target, and the number of connections to the same host as the current one in the past two seconds, SH. To make the method work fast, first, sort entities in the ascending order of SH. Take a random 10-element subset (upper row in Table 3.7) along with their Att categories (the second row) and SHCo values (the third row). Now take the entities whose SH values are nearest neighbors of those in the third row: these SH values are in the fourth row, and look at their Att categories (the bottom row). A striking success: all ten are predicted correctly!

Q.3.11. Build a tabular regression of the SHCo over Attack categories and find the correlation ratio.**A.** See Table 3.8.

Q.3.12. Apply NN classifier to predict Occupation from CI Mark over Student dataset. **A.** See Table 3.9.

Q.3.13. Consider a data table for 8 students and 2 features, as follows:

Student	Mark	Occupation
1	50	IT
2	80	IT
3	80	IT
4	60	AN
5	60	AN
6	40	AN
7	40	AN
8	50	AN

(i) Build a regression table for prediction Mark by Occupation.
(ii) Predict the mark for a new student whose occupation is IT.
(iii) Find the correlation ratio for the table.

A. (i) Regression table of Mark over Occupation contains Occupation category frequencies as well as Mark within-category averages and variances is this:

	Mark		
	Frequency	Average	Variance
IT	3	70	14.1
AN	5	50	8.9

(ii) For an IT student the likely mark will be 70±14.1.
(iii) The correlation ratio is determined by the weighted within-category variance, which is (3*14.1 + 5*8.9)/8 = (42.3 + 44.5)/8 = 10.85, and the total variance, which is calculated on all the data set with the mean = 57.5, and equal to 14.79. Then correlation ratio is $\eta^2 = 1 - 10.85/14.79 = 0.266$. This means that the tabular regression explains only 26.6% of the variance of Mark.

Q.3.14. Build tabular regression of the CI mark over Occupation in Student data and find the correlation ratio. **A.** See Table 3.10.

3.3.3.2 Interval Predicate Classifier

Another, more human friendly, classifier can be built in terms of quantitative feature x intervals. To predict a target feature category k, such a classifier would rely on an interval predicate $x(a(k), b(k))$ which is true if and only if the value of x is between $a(k)$ and $b(k)$. Then an interval predicate rule would be a production $x(a(x), b(k)) \Rightarrow k$. Consider, for example, "Saint" Attack in Intrusion data: there are 11 cases of this type and all, except one, have SHCo values 482 or 483. Thus, the interval predicate rule SHCo(482, 483)⇒ Saint would make only 9% of errors.

How one can infer which of the categories are more likely to be well covered by an interval predicate rule? One of the proposals is to rely on category contributions to the variance of x in (3.13), $p_k(\bar{x} - \bar{x}_k)^2$, in the denotations of this section, where p_k is proportion of entities in category k, $\bar{x}$is grand mean and $\bar{x}_k$ is within category k mean. The mechanism making sense of this proposal is illustrated on Fig. 3.17 (top): the further away the within-category mean is, the more plausible that the entire category is further away. Yet, in many cases, many entities in a category fall apart from their averages thus leading to errors in the interval based prediction (Fig. 3.17, bottom).

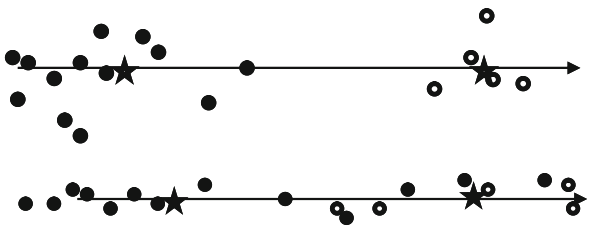

Fig. 3.17 A group of white *circles falls* apart from the rest *on the top*, and much intermixes with the rest *on the bottom*

Worked example 3.8. Category contributions for interval predicate productions

Consider the same features Att and SHCo from Intrusion dataset as those considered in Worked example 3.7 and determine the Att category contributions according to formula (3.13), $p_k(\bar{x} - \bar{x}_k)^2$ (see Table 3.11).

With respect to the data in Table 3.11, one can try to build interval predicate based productions for the largest contributing Saint and Smurf categories. We already observed that SH(482, 483)⇒ Saint makes 9% error, which is a false negative. This is caused by a SH value of 510 corresponding to Saint at 90-th row of the Intrusion data table – this does not satisfy the production's subject. Now one can see that rule SH(490, 512) ⇒ Smurf would fail only once too, on the same observation – but this time this would be a false positive, satisfying the subject but being not Smurf. The next contributing category, lagging far behind, is "Normal" corresponding to the range of x values from 1 to 28 which overlaps the range (16, 42) of x values corresponding to Apache category. Yet the rule SH(1, 15)⇒ Normal is true for 53 of 56 cases, the three false negative errors making about 5% only. The rule SHCo(16, 42)⇒ Apache has the same three cases as false positives.

Q.3.15. Build a category contribution table like Table 3.11 for CI mark and Occupation features in Students dataset. **A.** See Table 3.12.

A rather successful usage of interval based productions in Worked example 3.8 is due to the tight correlation between SH and Attack. In a less comfortable situation, such as that of pair CI mark – Occupation at Student dataset, the interval based descriptions make no sense at all. Consider the most contributing category IT – its CI mark range is from 53 to 90. If one takes the entire range to make it into rule CI(53, 90)⇒ IT, this would make no false negatives at all. Yet there are 22 entities

Table 3.11 Category contributions according to formula (3.13)

Attack	Proportion	Mean	Contribution
Apache	0.23	33.61	1514.3
Saint	0.11	484.64	15049.8
Smurf	0.10	508.40	15496.0
Normal	0.56	5.13	6729.9
Total	1.00	114.75	38790.0

Table 3.12 Occupation category contributions to CI Mark

Occupation	Proportion	Mean	Squared diff.	Contribution
IT	0.35	70.574	4, 980.327	1, 743.114
BA	0.34	54.794	3, 002.395	1, 020.814
AN	0.31	38.774	1, 503.438	466.066
Total	1.00	55.350		

of BA category and 15 of AN category whose CI mark falls within (53, 90) interval too, totaling to 37 false positive errors! One can try to somewhat reduce the interval predicate range, to lessen the false positive errors, with the price of admitting some false negatives. Consider, for example, CI(62, 90)⇒ IT rule to admit 12 false negative errors as well as 10 BA and 11 AN false positives, a drop to 33 errors altogether – quite a high error rate! Yet the interval based rules follow human way of thinking, which may lead to overall acceptance of such a rule, possibly amended by another feature interval added.

3.4 Two Nominal Features Case

P3.4.1 Analysis of Contingency Tables: Presentation

P3.4.1.1 Deriving Conceptual Relations from Statistics

To analyze interrelations between two nominal features, they are cross-classified in the so-called contingency table. A contingency table has its rows corresponding to categories of one feature and columns to categories of the other feature, with the entries reflecting the counts of entities falling in the overlap of the corresponding row and column categories.

Worked example 3.9. Contingency table on Market towns data

To cross-classify features Banks and Farmer's Market on Market towns data, we first need to categorize the quantitative feature Banks. Consider, for example, the four-category partition of the range of Banks feature at Market towns set presented in Table 3.13.

These categories are cross-classified with FM "yes" and "no" categories in Table 3.14. Besides the cross-classification counts, the table also contains summary within category counts, the totals, on the margins of the table, the last row and last column – this is why they are referred to as marginal frequencies. The total count balances the sheet in the bottom-right corner.

The same contingency data converted to relative frequencies by relating them to the total number of entities are presented in Table 3.15.

Table 3.13 Definition of Ba categories on the Market town dataset

Category	Definition	Notation
1	Ba ≥ 10	10+
2	10 > Ba ≥ 4	4+
3	4 > Ba ≥ 2	2+
4	Ba = 0 or 1	1−

Table 3.14 Cross classification of the Ba categories with FM categories

	Bank/Building Society categories				
FarmMarket	10+	4+	2+	1–	Total
Yes	2	5	1	1	9
No	4	7	13	12	36
Total	6	12	14	13	45

Table 3.15 BA/FM cross-classification relative frequencies, per cent

FM | Ba	10+	4+	2+	1–	Total
Yes	4.44	11.11	2.22	2.22	20
No	8.89	15.56	28.89	26.67	80
Total	13.33	26.67	31.11	28.89	100

Table 3.16 Protocol/Attack contingency table for Intrusion data

Category	Apache	Saint	Smurf	Norm	Total
Tcp	23	11	0	30	64
Udp	0	0	0	26	26
Icmp	0	0	10	0	10
Total	23	11	10	56	100

Q.3.16. Build a contingency table for features "Protocol-type" and "Attack type" in Intrusion data. **A.** See Table 3.16.

A contingency table can be used for assessment of correlation between two sets of categories. The highest level of correlation is that of a conceptual association. A conceptual association may exist if a row, k, has all its entries, not marginal of course, except just one, say l, equal to 0, which would mean that all of the extent of category k belongs to the column category l. The data, thus, indicate that category k implies category l.

Worked example 3.10. Equivalence and implication from a contingency table

Such are rows "Udp" and "Icmp" in Table 3.16. There is a perfect match in this table: a row category k= "Icmp" and a column category l = "Smurf", that contains the only non-zero count. No other combination (k, l') or (k', l) is possible according to the table. In such a situation, one may claim that, subject to the sampling error, category l may occur if and only if k does, that is, k and l are equivalent.

A somewhat weaker, but still very much valuable is the case of "Udp" row in Table 3.16. It appears, Udp protocol implies "Norm" column category – a no-attack situation, though there is no equivalence here because the "Norm" column contains another positive count, in row "Tcp".

Case study 3.3. Trimming Contingency Data: A Bad Option

Unfortunately, there are no zeros in Table 3.14: thus, no conceptual relation between the number of Banks and the presence of a Farmer's market. But some of the entries are really close to 0, which may make us tempted to trim the data a bit. Imagine, for example, that in row "Yes" of Table 3.14, two last entries are 0, not 1s. This would imply that a Farmers Market may occur only in a town with 4 or more Banks. A logical implication, that is, a production rule, "If BA is 4 or more, then a Farmer's market must be present", could be derived then from thus modified table. One may try taking this path and cleaning the data of smaller entries, by removing corresponding entities from the table of course, to not obscure our "vision" of the pattern of correlation. Thus trimmed Table 3.17 is obtained from Table 3.14 by removing just 13 entities from "less popular" entries. This latter table expresses, with no exception, a very simple conceptual statement "A town has a Farmer's market if and only if the number of Banks in it is 4 or greater". However nice the rule may sound, let us not forget the cost of the trimming which is the 13 towns, almost 30% of the sample, that have been removed as those not fitting the stated perspective. Such a data doctoring borders with forgery – one of the reasons for a famous quip attributed to B. Disraeli, a celebrated British politician of XIX century: "There are three gradations of lies: lies, damned lies and statistics." The issue of sample adjustment so far has received no reasonable solution, even with respect to outliers – values falling way beyond the feature range one would expect normally. Anyway, the conclusion of the trimming exercise is that one should try finding ways of expressing conceptual relations without much doctoring the sample.

P3.4.1.2 Capturing Relationships with Quetelet Indexes

Quetelet index provides for a strategy for visualization of correlation patterns in contingency tables without removal of "not-fitting" entities. In 1832, A. Quetelet, a

Table 3.17 A trimmed BA/FM cross classification "cleaned" of 13 towns, to sharpen the view

	Number of Banks/Build. Societies				
FMarket	10+	4+	2+	1–	Total
Yes	2	5	0	0	7
No	0	0	13	12	25
Total	2	5	13	12	32

founding father of statistics, proposed to measure the extent of association between row and column categories in a contingency table by comparing the local count with an average one.

Let us consider correlation between the presence of a Farmer's Market and the category "10 or more Banks" according to Table 3.15. We can see that their joint probability/frequency is the entry in the corresponding row and column: P(Ba = 10+ & FM = Yes) = 2/45 = 4.44% (joint probability/frequency rate). Of the 20% entities that fall in the row "Yes", this makes the proportion of "Ba = 10+" under condition "FM = Yes" equal to P(Ba = 10+ /FM = Yes) = P(Ba = 10+ & FM =Yes)/ P(FM = Yes) = 0.0444/0.20 = 0.222 = 22.2%. Such a ratio expresses the conditional probability/rate.

Is this high or low? Hard to tell without comparing this with the unconditional rate, that is, with the frequency of category "Ba = 10+" in the whole dataset, which is P(Ba = 10+) = 13.33%. Let us compute the (relative) difference between the two, which is referred to as Quetelet index q:

$$q(\text{Ba} = 10+/\text{FM} = \text{Yes}) = [\text{P}(\text{Ba} = 10+ \text{FM} = \text{Yes}) - \text{P}(\text{Ba} = 10+)]/\text{P}(\text{Ba} = 10+) = [0.2222 - 0.1333]/0.1333 = 0.6667 = 66.7\%$$

That means that condition "FM = Yes" raises the frequency of the Bank category by 66.7%. This logic concurs with our everyday intuition. Consider, for example, the risk of getting a serious illness, say tuberculosis, which may be, say, about 0.1%, one in a thousand, in a given region. Take a condition such as "Bad housing" and count the rate of tuberculosis under this condition, amounting to, say 0.5% – which is very small by itself, yet a five-fold increase over the average tuberculosis rate. This is exactly what Quetelet index measures: $q(l/k) = (0.5 - 0.1)/0.1 = 400\%$ to show that the change of the average rate is 4 times.

Worked example 3.11. Quetelet index in a contingency table

Let us apply the general Quetelet index formula (3.15) to entries in Table 3.14. This leads to Quetelet index values presented in Table 3.18. By highlighting positive values in the table, we obtain the same pattern as on the "purified" data as in Case-study 3.3, but this time in a somewhat more realistic manner, keeping the sample intact. Specifically, one can see that "Yes" FM category provides for a strong increase in the probabilities, whereas "No" category leads to much weaker changes.

Table 3.18 BA/FM Cross classification Quetelet coefficients, % (positive entries highlighted)

FMarket	10+	4+	2+	1–
Yes	**66.67**	**108.33**	−64.29	−61.54
No	−16.67	−27.08	**16.07**	**15.38**

Table 3.19 Quetelet indices for the Protocol/Attack contingency Table 3.16, per cent

Category	Apache	Saint	Surf	Norm
Tcp	**56.25**	**56.25**	−100.00	−16.29
Udp	−100.00	−100.00	−100.00	**78.57**
Icmp	−100.00	−100.00	**900.00**	−100.00

Q.3.17. Compute Quetelet coefficients for Table 3.16. **A.** See Table 3.19 in which positive entries are highlighted in bold.

Case-study 3.4. Has There Been a Bias in S'nS' Policy?

Take on the case of Stop-and-Search policy in England and Wales 2005 represented according to race (B - black, A - asian and W - white), by numbers in Table 2.4 in Section 2.3 – these are overwhelmingly in category W. The criticism of this policy came out of comparison of this distribution with the distribution of the entire population. Such a distribution, according to the latest pre-2005 census 2001, can be easily found on web. By subtracting from that the numbers of Stop-and-Search occurrences, under the assumption that nobody has been subjected to this more than once, Table 3.20 has been drawn. Its last column gives the numbers that were used for the claim of a racial bias: indeed category B members have been subjects of the policy six times more frequently than category W members. A similar picture emerges when Quetelet coefficients are used (see Table 3.21). Category B is subject to Stop-and-Search policy 400% more frequently than on average, whereas category W is 15% less.

Yet some would consider drawing a table like Table 3.20, and of course the derived Table 3.21, as something nonsensical, because it is based on an implicit

Table 3.20 Distribution of Stop-and-Search policy cross-classified with race

	S'n'S	Not S'n'S	Total	S'n'S-to-Total
Black	131,723	1,377,493	1,509,216	0.0873
Asian	70,252	2,948,179	3,018,431	0.0233
White	676,178	46,838,091	47,514,269	0.0142
Total	878,153	51,163,763	52,041,916	0.0169

Table 3.21 Relative Quetelet coefficients for cross-classification in Table 3.20, per cent

	S'n'S	Not S'n'S
Black	417.2	−7.2
Asian	37.9	−0.6
White	−15.7	0.3

assumption that the Stop-and-Search policy applies to the population randomly. They would argue that police apply the policy only when they deem it necessary, so that the comparison should involve not all of the total population but only those criminal. Indeed, the distribution of subjects to Stop-and-Search policy by race has been almost identical to that of the imprisoned population of the same year. Therefore, the claim of a racial bias should be declared incorrect.

P3.4.1.3 Chi-Square Contingency Coefficient As a Summary Correlation Index

A somewhat more refined visualization of the contingency table comes from the Quetelet indexes weighted by the probabilities of corresponding entries, as explained in Section 3.4.2. They sum to a most popular concept in the analysis of contingency tables, the celebrated chi-square contingency coefficient. This coefficient was introduced by K. Pearson (1900) to express the deviation of the observed bivariate distribution, represented by the relative frequencies in a contingency table, from the situation of statistical independence between the features.

Worked example 3.12. Visualization of contingency table using weighted Quetelet coefficients

Let us multiply Quetelet coefficients in Table 3.18 by the frequencies of the corresponding entries in Table 3.14. Quetelet coefficients in Table 3.18 are taken relative to unity, not per cent. This leads us to Table 3.22 whose entries sum to the value of Pearson's chi-square coefficient for Table 3.14, 6.86. Note that entries in Table 3.20 can be both positive and negative; those with absolute value greater than $6.86/4 = 1.72$ are highlighted in bold – they show the entries of an extraordinary deviation from the average. Of them, column 4+ supplies the highest positive impact and the highest negative impact.

A pair of categories, one from one nominal feature and the other from another nominal feature, are said to be statistically independent if the probability of their co-occurrence is equal to the product of probabilities of these categories. Take, for example, category "Yes" of FM and "4+" of Banks in Table 3.15: the probability of their co-occurrence is 0.111. On the other hand, the probability of FM = "Yes" is 0.2 and that of Banks = 4+ is 0.267, according to the table. If these two categories were independent they would have co-occurred at the level of $0.2 \times 0.267 = 0.053$,

Table 3.22 BA/FM chi-squared (NQ = 6.86) and its decomposition according to (3.19)

FMarket	10+	4+	2+	1–	Total
Yes	1.33	**5.41**	−.064	−.062	5.48
No	−.067	**−1.90**	**2.09**	**1.85**	1.37
Total	0.67	3.51	1.45	1.23	**6.86**

about twice as less than in reality, which means that the pair highly deviates from the statistical independence. Two features are said to be statistically independent if all pairs of their mutual categories are statistically independent. K. Pearson was concerned with the situation at which two features are independent in the population at large but this may not necessarily be reflected in the sample under consideration because of the randomness of sampling. Thus he proposed to take the squared differences between observed frequencies and those that would occur under the independence assumption and relate them to the "theoretical" probabilities that should be true in the population. The summary index is referred to as the Pearson chi-square coefficient, see (3.18) later. The distribution of the summary chi-square index, under conventional assumptions of independence in sampling, converges to the so-called chi-square distribution, which allows for statistical testing of the hypothesis of independence between the features. This suggests that the coefficient should be used only for testing the hypothesis, but not as a measure of correlation. The claim would be – and often has been – that the index can only distinguish between two cases, statistical independence or not, and thus cannot be used for comparison of the extent of the dependence. Yet practitioners are always tempted to ignore this commandment and do compare the extent of dependence at different pairs of categorical features. Indeed, as formula (3.19) shows there is nothing wrong in using chi-square contingency coefficient as an index of correlation – it is indeed the summary Quetelet index, thus showing the average degree of relationship between two features.

Worked example 3.13. A conventional decomposition of chi-square coefficient

Let us consider a conventional way of visualization of contingency tables, by putting Pearson indexes, the square roots *x(k,l)* of the chi-square coefficient items in (3.21) as the table's elements. These are in Table 3.23. The table does show a similar pattern of positive and negative associations. However, it is not the entries of the table that sum to the chi-square coefficient but rather the squares of the entries. The fact that the summary values on the margins in Tables 3.22 and 3.23 are the same is not by chance: it exemplifies a mathematical property (see Equation (3.19)).

Table 3.23 Square roots of the items in Pearson chi-squared ($X^2 = 6.86$); the items themselves are in parentheses

FMarket	10+	4+	2+	1–	Total
Yes	**0.73**(0.53)	**1.68(2.82)**	−1.08 (1.16)	−0.99 (0.98)	(5.49)
No	−0.36 (0.13)	−0.84 (0.70)	**0.54 (0.29)**	**0.50 (0.25)**	(1.37)
Total	(0.67)	(3.52)	(1.45)	(1.23)	(6.86)

Q.3.18. In Table 3.22, all marginal values, the sums of rows and columns, are positive, in spite of the fact that many within-table entries are negative. Is this just due to specifics of the distribution in Table 3.14 or a general property? **A**: A general property: the within-row or within-column sums of the elements, $N_{lk}\ q(l/k)$, must be positive, see (3.19).

Q.3.19. Find a similar decomposition of chi-squared for OOPmarks/Occupation in Student data. Hint: First, categorize quantitative feature OOPmarks somehow: you may use equal bins, or conventional boundary points such as 35, 65 and 75, or any other considerations.

Q.3.20. Can any logical production rules come from the columns of Table 3.16? **A.** Yes, both Apache and Saint attacks may occur at the tcp protocol only.

Q.3.21. Among the shoppers in Q.2.21, those who spent £60 each are males only and those who spent £100 each are females only, whereas among the rest 30 individuals half are men and half are women. Build a contingency table for the two features, gender and spending. Find and interpret the value of Quetelet coefficient for females who spent £100 each.
A. The contingency table (of co-occurrence counts):

Spending, £				
Gender	60	100	150	Total
Female	0	20	15	35
Male	50	0	15	65
Total	50	20	30	100

This table of absolute co-occurrence counts coincides with that of proportions expressed per cent because the number of shoppers is 100.

Quetelet coefficient for (Female/£100) entry is

$$Q = 100^*20/(20^*35) - 1 = 2.86 - 1 = 1.86$$

This means that being female in this category of spending is more likely than the average, by 186%.

F3.4.2 Analysis of Contingency Tables: Formulation

Consider two sets of disjoint categories on an entity set I: $l = 1, \ldots, L$ (for example, occupation of individuals constituting I) and $k = 1, \ldots, K$ (say, family or housing type). Each makes a partition of the entity set I; they are crossed to see if there is any correlation between them. Combine a pair of categories $(k, l) \in K \times L$ and count the number of entities that fall in both. The *(k,l)* co-occurrence count is

denoted by N_{kl}. Obviously, these counts sum to N because the categories are not overlapping and cover the entire dataset. A table housing these counts, N_{kl} , or their relative values, frequencies $p_{kl} = N_{kl}/N$, is referred to as a contingency table or just cross-classification. The totals, that is, within-row sums $N_{k+} = \Sigma_l N_{kl}$ and within-column sums $N_{+l} = \Sigma_k N_{kl}$ (as well as their relative frequency counterparts) are referred to as marginals (because they are located on margins of the contingency table).

The (empirical) probability that category l occurs under condition of k can be expressed as $P(l/k) = p_{kl}/p_{k+} = N_{kl}/N_{k+}$. The probability $P(l)$ of the category l with no condition is just $p_{+l} = N_{+l}/N$. Similar notation is used when l and k are swapped. The relative difference between the two probabilities is referred to as (relative) Quetelet index (Mirkin 2001):

$$q(l/k) = \frac{P(l/k) - P(l)}{P(l)} \tag{3.16}$$

where $P(l) = N_{+1}/N$, $P(k) = N_{k+}/N$, $P(l/k) = N_{kl}/N_{k+}$ That is, Quetelet index expresses correlation between categories k and l as the relative change in the probability of l when k is taken into account.

With little algebra, one can derive a simpler expression

$$q(l/k) = [N_{kl}/N_{k+} - N_{+l}/N]/(N_{+l}/N) = N_{kl}N/(N_{k+}N_{+l}) - 1 = \frac{p_{kl}}{p_{k+}p_{+l}} - 1 \tag{3.16'}$$

Highlighting high positive and negative values in a Quetelet index table, such as Tables 3.18 and 3.21, visualizes the pattern of correlation between the two sets of categories.

This visualization can be extended to a theoretically sound presentation. Let us define the summary Quetelet correlation index Q as the sum of pair-wise Quetelet indexes weighted by their frequencies/probabilities:

$$Q = \sum_{k=1}^{K}\sum_{l=1}^{L} p_{kl}q(l,k) = \sum_{k=1}^{K}\sum_{l=1}^{L} p_{kl}\left(\frac{p_{kl}}{p_{k+}p_{+l}} - 1\right) = \sum_{k=1}^{K}\sum_{l=1}^{L} \frac{{p_{kl}}^2}{p_{k+}p_{+l}} - 1 \tag{3.17}$$

The right-hand expression for Q in (3.17) is very popular in statistical analysis of contingency data. In fact, this is equal to chi-squared correlation coefficient proposed by K. Pearson (1900) in a very different context – as a measure of deviation of the contingency table entries from the statistical independence.

To explain this in more detail, let us first introduce the concept of statistical independence. The sets of k and l categories are said to be statistically independent if $p_{kl} = p_{k+}p_{+l}$ for all k and l. Obviously, such a condition is hard to fulfill in reality. K. Pearson suggested using relative squared errors to measure the deviations of observed frequencies from the statistical independence. Specifically, he introduced the following coefficient usually referred to as Pearson's chi-squared association

coefficient:

$$X^2 = N \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} = N \left(\sum_{k=1}^{K} \sum_{l=1}^{L} \frac{{p_{kl}}^2}{p_{k+}p_{+l}} - 1 \right) \tag{3.18}$$

The equation on the right can be proven with little algebra. Consider, for example, this part of the expression on the left in (3.18):

$$\sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} = \sum_{l=1}^{L} \frac{{p_{kl}}^2 - 2p_{kl}p_{k+}p_{+l} + (p_{k+}p_{+l})^2}{p_{k+}p_{+l}}$$
$$= \sum_{l=1}^{L} \frac{{p_{kl}}^2}{p_{k+}p_{+l}} - 2\sum_{l=1}^{L} p_{kl} + \sum_{l=1}^{L} p_{k+}p_{+l} = \sum_{l=1}^{L} \frac{{p_{kl}}^2}{p_{k+}p_{+l}} - p_{k+}$$

The expression on the right in the above is derived by using equations $\Sigma_l p_{kl} = p_{k+}$ and $\Sigma_l\ p_{+l} = 1$. Summing these equations over k will produce (3.18). On the other hand, the expression on the right in (3.18) is obviously equal to $\Sigma_l p_{kl} q(l/k)$ so that

$$\sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} = \sum_{l=1}^{L} p_{kl} q(l/k) \tag{3.19}$$

By comparing the right-hand parts of (3.17) and (3.18), it is easy to see that $X^2 = NQ$. The same follows from summing Equation (3.19) over k.

The popularity of X^2 index in statistics and related fields rests on the theorem proven by K. Pearson: if the contingency table is based on a sample of entities independently drawn from a population in which the statistical independence holds (so that all deviations are due to just randomness in the sampling), then the probabilistic distribution of X^2 converges to the chi-squared distribution (when N tends to infinity) introduced by Pearson earlier for similar analyses. The probabilistic chi-squared distribution is defined as the distribution of the sum of squares of random variables distributed according to the standard Gaussian distribution.

This theorem is not always of interest to a computational data analyst, because they analyze data that are not necessarily random or not necessarily independently sampled. However, Pearson's chi-squared coefficient is frequently used just for scoring correlation in contingency tables, and the equation $X^2 = NQ$ gives a credible support to it. According to this equation, X^2 also is not necessarily a measure of deviation from the statistical independence. It also has a different meaning of a measure of interrelation between categories: that of the averaged Quetelet coefficient.

To get more intuition on the underlying correlation concept, let us take a look at the extreme values that X^2 can take and situations at which the extreme values are reached (Mirkin 2001). It appears that at $K \leq L$, that is, the number of columns is not greater than that of rows, X^2 ranges between *0* and $K - 1$. It reaches *0* if there is a statistical independence at all *(k,l)* entries so that all $q_{kl} = 0$, and it reaches $K - 1$

if each column l contains only one non-zero entry $p_{k(l)l}$, which is thus equal to p_{+l}. The latter can be interpreted as the logical implication $k \to l(k)$.

Representation of chi-squared through Quetelet coefficients,

$$X^2 = \sum_{k=1}^{K} \sum_{l=1}^{L} N p_{kl} q(l/k) \tag{3.20}$$

amounts to decomposition of X^2 into the sum of $N_{kl}\ q(l/k)$ items and allows for visualization of the items within the contingency table format, such as that presented in Table 3.21.

In fact not only the total sum of these items coincide with that of the original chi-squared items $N(p_{kl} - p_{k+}p_{+l})^2/p_{k+}p_{+l}$, but also the within-column and within-row sums coincide too, as (3.19) clearly demonstrates for the latter case.

However all the original chi-squared items in (3.18) are positive and cannot show whether the contribution of an individual entry is positive or negative. To overcome this shortcoming, another visualization of X^2 is in use. That visualization involves the square roots of the chi-squared items

$$r(k,l) = \frac{p_{kl} - p_{k+}p_{+l}}{\sqrt{p_{k+}p_{+l}}} \tag{3.21}$$

that are convenient to refer to as Pearson indexes. Obviously, $X^2 = N\Sigma_{k,l}\ r(k,l)^2$. Pearson indexes indeed have the same signs as $q(l/k)$, and in fact are closely related: $q(l/k) = r(k,l)[p_{k+}\ p_{+l}]^{1/2}$. It is less clear what interpretation of its own $r(k,l)$ may have, although they are useful in Correspondence analysis of contingency tables (Section 5.4, see also normalized Laplacian in Section 8.2).

Q.3.22. Take two binary features presented as 1/0 variables and build their contingency table, sometimes referred to as a four-fold table (Table 3.24) when symbols *a, b, c, d* are used to denote the co-occurrence numbers.

Prove that Quetelet coefficient *q(Yes/Yes)* expressing the relative difference between $a/(a+c)$ and $(a+b)/N$ is equal to

$$q(Yes/Yes) = \frac{ad - bc}{(a+c)(a+b)},$$

Table 3.24 Four-fold contingency table between binary features

		Feature Y		
		Yes	No	Total
Feature X	Yes	*a*	*b*	*a+b*
	Not	*c*	*d*	*c+d*
Total		*a+c*	*b+d*	$N = a+b+c+d$

and the summary Quetelet coefficient Q, or Pearson's X^2/N, is equal to

$$Q = \frac{(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}.$$

Q.3.23. Prove that the correlation coefficient between two 1/0 binary features can be expressed in terms of the four-fold table as $\rho = \sqrt{Q}$, that is,

$$\rho = \frac{ad - bc}{\sqrt{(a + c)(b + d)(a + b)(c + d)}}.$$

Q.3.24. Given a $K \times L$ contingency table P and a pair of categories, $k \in K$ and $l \in L$, consider an absolute Quetelet index $a(l/k) = P(l/k) = P(l)$ – the change from the frequency of $l \in L$ on the whole entity set I to the frequency of l on entities falling in category $k \in K$. In terms of P, $P(l) = p_{+1}$ and $P(l/k) = p_{kl}/p_{+l}$. Prove that the summary Quetelet index $A = \Sigma_{k,l}\, p_{kl} a(l/k) = \Sigma_{k,l} {p_{kl}}^2/p_{k+} - \Sigma_l\, {p_{+l}}^2$ is equal to the following expression, an asymmetric analogue to Pearson chi-squared:

$$A = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}} \tag{3.22}$$

which also is the numerator of the so called Goodman-Kruskal "tau-b" index (Kendall and Stewart 1973).

A. Indeed, by taking the square of the denominator, expression in (3.22) becomes equal to $\Sigma_{k,l}\ ({p_{kl}}^2 - 2p_{kl}p_{k+}p_{+l} + {p_{k+}}^2{p_{+l}}^2)/p_{k+}$, which is $\Sigma_{k,l}{p_{kl}}^2/p_{k+} - 2\Sigma_{k,l}p_{kl}p_{+l} + \Sigma_{k,l}\, {p_{k+}p_{+l}}^2 = \Sigma_{k.l}{p_{kl}}^2/p_{k+} - 2\Sigma_{k,l}\, {p_{k+l}}^2 + \Sigma_l {p_{+l}}^2$ because $\Sigma_k p_{kl} = p_{+l}$ and $\Sigma_k p_{k+} = 1$ .This is obviolsly $\Sigma_{k,l}{p_{kl}}^2/p_{k+} - \Sigma_l\, {p_{+l}}^2 = \Sigma_{k,l}\, p_{kl} a(l/k) = A$, which proves the statement.

3.5 Summary

The Chapter outlines several important characteristics of summarization and correlation between two features, and displays some of the properties of those. They are:

– linear regression and correlation coefficient for two quantitative variables;
– tabular regression, correlation ratio, decomposition of the quantitative feature scatter, and nearest neighbor classifier for the mixed scale case; and
– contingency table, Quetelet index, statistical independence, and Pearson's chi-squared for two nominal variables.

They all are applicable in the case of multidimensional data as well.

Some of the characteristics are rather unconventional. For example, the concepts of tabular regression and correlation ratio are not terribly popular in data mining. The Quetelet indexes are recognized by neither community, the more so the idea that Pearson chi-squared is a summary correlation measure, not necessarily a criterion of statistical independence.

Some examples of non-linear regression and nature-inspired approaches for fitting that are outlined. Computational bootstrap based validation is considered.

References

Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**, 1141–1164 (2000)

Davison, A.C., Hinkley, D.V.: Bootstrap Methods and Their Application. Cambridge University Press, Cambridge (7th printing) (2005)

Kendall, M.G., Stewart, A.: Advanced Statistics: Inference and Relationship, 3d edn. Griffin, London, ISBN: 0852642156 (1973)

Lohninger, H.: Teach Me Data Analysis. Springer, Berlin-New York-Tokyo, ISBN 3-540-14743-8 (1999)

Mirkin, B.: Eleven ways to look at the chi-squared coefficient for contingency tables. Am. Stat. **55**(2), 111–120 (2001)

Pearson, K.: On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling. Phil. Mag. **50**, 157–175 (1900)