Chapter 14 Analysis of Continuous Variables

14.1 Introduction

This chapter presents the methods used to measure the association between two variables when one or both of them is continuous. Of course, one option to avoid dealing with continuous variables is to bin them before using them. We explore this issue in the first section of this chapter. When one of the two variables is nominal or ordinal (or a binned continuous variable), we can extend the definitions of the Gini and entropy variances to the case of continuous variables. We could also use what is known as the F-test. These methods are presented in Section 14.3. Finally, we discuss correlation analysis of continuous variables in Section 14.4.

14.2 When Is Binning Necessary?

Binning is used in two situations: (1) as a requirement during modeling because of either the model form or the accepted practices and (2) as a tool to facilitate data exploration.

Let us consider the first situation. Most data mining algorithms deal efficiently with continuous variables, in their raw form, if not explicitly require them, as in the case of neural networks and regression models. However, the procedures used in the implementation and presentation of the model results sometimes force the analyst to bin continuous variables into ranges or new categories.

A common example of such a case is during the construction of scorecards. In most scorecards developed using logistic regression, continuous variables are binned into ranges, and a set of new indicator variables representing these ranges is used as independent variables. This process facilitates the construction of scorecards because the scoring parameters are then given directly by the model coefficients.

Similarly, when the final score is required to be produced using a simple set of IF–THEN–ELSE rules, this type of model is naturally generated using decision trees. These models are very efficient and therefore are the most used scoring models for online transactions. Continuous variables are then binned either using the decision tree algorithm or during the original data preparation procedures.

In the second situation, binning is used only so that we can implement some of the measures used for categorical and ordinal variables. This will help the analyst to evaluate the predictive power of the variables before building the model.

When one of the two variables we study is nominal or ordinal (either originally or as a result of binning a continuous variable), we can use one of the following measures to determine the association between the variables (Breiman et al. 1998).

- F-test
- Gini variance
- Entropy variance

These methods are described in detail in the next section.

14.3 Measures of Association

In the following, we always assume that the X variable is nominal and the Y variable is continuous.

Before we present the formulas and implementation of three association measures, we must mention that all these measures ignore the possible ordering relationships of the categories of the X variable.

14.3.1 Notation

Table 14.1 shows that variable X has k categories, $x_1, \ldots, x_k$. It also lists the values of the variable Y in each of these categories. For example, in the category x_i, there are n_i

Table 14.1 Notation for association measures of continuous variables.

X	Y
x_1	$y_{11} \cdots y_{1n_1}$
$\vdots$	$\vdots$
x_i	$y_{i1} \cdots y_{in_i}$
$\vdots$	$\vdots$
x_k	$y_{k1} \cdots y_{kn_k}$
Total	N records

records, with the variable Y taking the values of $y_{i1}, \ldots, y_{in_i}$. The total number of records is N, such that $N = \sum_{i=1}^{k} n_i$.

The sum of the values of Y in the category i is given as

$$y_i = \sum_{j=1}^{n_i} y_{ij}. \tag{14.1}$$

The average value of the variable Y in the category i is then calculated as

$$\bar{y}_i = \frac{y_i}{n_i}. \tag{14.2}$$

Similarly, the average of the variable Y is given by

$$\bar{y} = \frac{1}{N} \sum_{i=1}^{k} y_i. \tag{14.3}$$

The sum of the squared deviations from the average value of Y is calculated as

$$SSTO = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2. \tag{14.4}$$

Similarly, the weighted squared deviations of the category average from the global average is

$$SSR = \sum_{i=1}^{k} n_i (\bar{y}_i - \bar{y})^2. \tag{14.5}$$

And finally the sum of squared deviations within the categories from their means is

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2. \tag{14.6}$$

Borrowing from the notation of linear regression, we can define the *mean square* (MS) values for SSR and SSE by dividing each sum by its degrees of freedom as

$$MSR = \frac{SSR}{(k-1)} \tag{14.7}$$

and

$$MSE = \frac{SSE}{(N-k)}. \tag{14.8}$$

14.3.2 The F-Test

The F-test is similar to the test devised for the analysis of variance in linear regression. However, in this case we are testing the association between the discrete variable X and the continuous variable Y. To perform the test, we define the value F^* as

$$F^* = \frac{MSR}{MSE}. \tag{14.9}$$

Large values of F^* suggest a high association between the two variables. F^* follows the F-distribution with degrees of freedom of $(k-1, N-k)$. Therefore, we calculate the p-value of F^* from the inverse F-distribution function such that

$$p = F(F^*, k-1, N-k). \tag{14.10}$$

Small p-values indicate that the variables are not associated and high p-values indicate that they are.

14.3.3 Gini and Entropy Variances

In Section 10.2.5, we defined the Gini variance when both variables are categorical. We now extend the formulation to account for the continuous variable Y. In this case, the Gini ratio is simply given as

$$G_r = 1 - \frac{SSE}{SSTO}. \tag{14.11}$$

In fact, Equation 14.11 also defines the *coefficient of determination,* R^2 in the terminology of linear regression, as well as the *entropy ratio* to measure the association between the variables.

Table 14.2 Parameters of the macro `ContGrF()`.

Header	`ContGrF(DSin, Xvar, Yvar, M_Gr, M_Fstar, M_Pvalue);`
Parameter	*Description*
`DSin`	Input dataset
`Xvar`	X variable name
`Yvar`	Y variable name
`M_Gr`	Output entropy/Gini ratio (Equation 14.11)
`M_Fstar`	Output F^* used in F-test (Equation 14.9)
`M_Pvalue`	Output p-value of F^* (Equation 14.10)

The following macro calculates both the value of F^* and the ratio G_r. The macro simply calculates the values defined in Equations 14.1 through 14.11.

Step 1

Begin by using PROC FREQ to find the unique categories of the X variable XVar. Assume that it does not have missing values. The categories are stored in the dataset Temp_Cats.

```
proc freq data=&DSin noprint;
 tables &XVar /out=Temp_Cats;
run;
```

Step 2

Convert the categories X_i and their frequencies n_i into macro variables. Also find the number of categories K and the total number of records N.

```
Data _null_;
 retain N 0;
 set Temp_Cats;
  N=N+count;
  call symput ("X_" || left(_N_), compress(&XVar));
  call symput ("n_" || left(_N_), left(count));

  call symput ("K", left(_N_));
  call symput ("N", left(N));
Run;
```

Step 3

Calculate the average of the variable Y, that is, $\bar{y}$, as well as the averages for each category, $\bar{y}_i$.

```
proc sql noprint;
 /* Ybar */
  select avg(&YVar) into :Ybar from &DSin;
  /* Ybar_i */
  %do i=1 %to &K;
    select avg(&YVar) into :Ybar_&i
           from &DSin where &XVar = "&&X_&i";
  %end;
```

Step 4

Calculate the remaining terms *SSTO*, *SSE*, and *SSR*.

```
  select var(&YVar) into: SSTO from &DSin;
%let SSTO=%sysevalf(&SSTO *(&N-1));
```

```
%let SSR=0;
%let SSE=0;
  %do i=1 %to &K;
     select var(&YVar) into: ssei
            from &DSin where &Xvar="&&X_&i";
       %let SSE=%sysevalf(&SSE + &ssei * (&&n_&i - 1));
      %let SSR=%sysevalf(&SSR+ &&n_&i *
               (&&Ybar_&i - &Ybar)*(&&Ybar_&i - &Ybar));
    %end;

  quit; /* end of Proc SQL */
```

Step 5

Substitute into the equations of MSR, MSE, F^*, G_r, and p-value.

```
%let MSR=%sysevalf(&SSR/(&K-1));
%let MSE=%sysevalf(&SSE/(&N-&K));
%let &M_Gr=%Sysevalf(1-(&SSE/&SSTO));
%let &M_Fstar=%sysevalf(&MSR/&MSE);
%let &M_PValue=
        %sysevalf(%sysfunc(probf(&Fstar,&K-1,&N-&K)));
```

Step 6

Clean the workspace and finish the macro.

```
/* clean workspace */
 proc datasets library=work nolist;
  delete temp_cats;
 run; quit;

%mend;
```

Let us demonstrate this macro with an example. The following code generates a dataset containing two variables: Loan, which varies between 0 and 2000, and Debt, which takes one of two values, Low or High.

```
data Credit;
 do CustID=1 to 10000;
  Loan=int(2000*ranuni(0));
  Err=(100-200*ranuni(0));
  /* if Loan + Err>1000 then Debt='High'; */
  /* if ranuni(0)>0.5 then Debt='High';   */
                      else Debt='Low';
  output;
 end;
 drop Err;
run;
```

The code contains two *commented* statements. You need to remove the comments from one of them for the code to work properly. The first one will generate data with strong association between the Loan value and the Debt status (high or low). The second statement will assign the status randomly independent of the value of the loan, thus creating low association between the two variables. We then invoke the macro using macro variables, as follows.

```
%let dsin=Credit;
%let XVar=Debt;
%let YVar=Loan;
%let Gr=;
%let Fstar=;
%let pvalue=;

%ContGrF(&DSin, &Xvar, &YVar, Gr, Fstar, Pvalue);

%put Gr=&Gr;
%put Fstar=&Fstar;
%put pvalue=&Pvalue;
```

If you investigate the two measures, G_r and F^*, you will discover that the p-value of the F-test saturates quickly; that is, it reaches either 1 or 0 and is less discriminant than the Gini/entropy ratio G_r.

It is worth noting that these measures are used as purity measures in *decision tree* models with continuous dependent variables (i.e., *regression trees*). Some analysts like to use decision tree software to identify the most significant variables (i.e., to *filter* them). The preceding macro is a simple alternative.

14.4 Correlation Coefficients

This section presents the well-known correlation concept as a way of determining the association between continuous variables. However, we have to remember that in this case, *both* variables have to be continuous or at least have to have been assigned scores on a numerical scale, as in the case of ordinal variables.

The correlation coefficient is defined in statistics as the *Pearson correlation coefficient*. It is defined as

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right\}^{1/2}}, \tag{14.12}$$

where x_i, y_i, $i = 1, \ldots, n$ are the n observations of the variables x and y, with mean values of $\bar{x}$ and $\bar{y}$, respectively.

The value of r is always between -1.0 and 1.0. Variables that have a correlation coefficient near 0 are called uncorrelated; those with r closer to 1 or -1 are said to be correlated. Figure 14.1 shows cases where r is positive, negative, and zero.

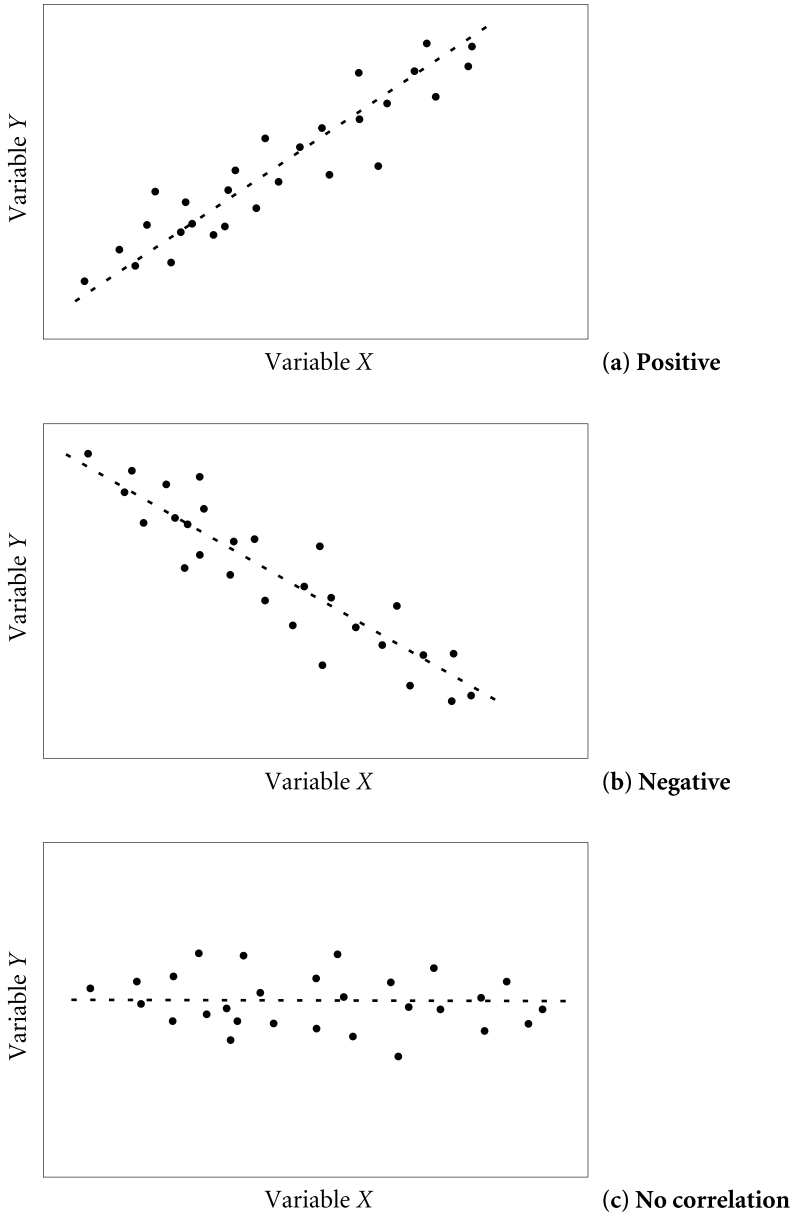


Figure 14.1 Correlation between continuous variables.

The figure shows the best linear fit in each case. It is well known that the slope of this line is the correlation coefficient defined in Equation 14.12.

Equation 14.12 does not prevent us from using a categorical variable if the two categories are represented by numbers, for example, 0 and 1.

Figure 14.2(a) shows this case. However, Figure 14.2 (b) shows that the scale of the variable x, which is not binary, *does* play a significant role in the value of r. If the variable x is normalized, between 0 and 1 in this case, we may obtain a larger value for r. Therefore, if we use any binary variables in association with the correlation coefficient, normalization is a necessary step.

Note that the normalization of two continuous variables does not change the value of r between them. It only plays a role when one of them is binary. Figure 14.3 shows the cases when r is positive, negative, and zero (after normalizing the continuous variable).

The problem with the Pearson correlation coefficient is that it is sensitive to outliers. Let us demonstrate this feature with a simple example.

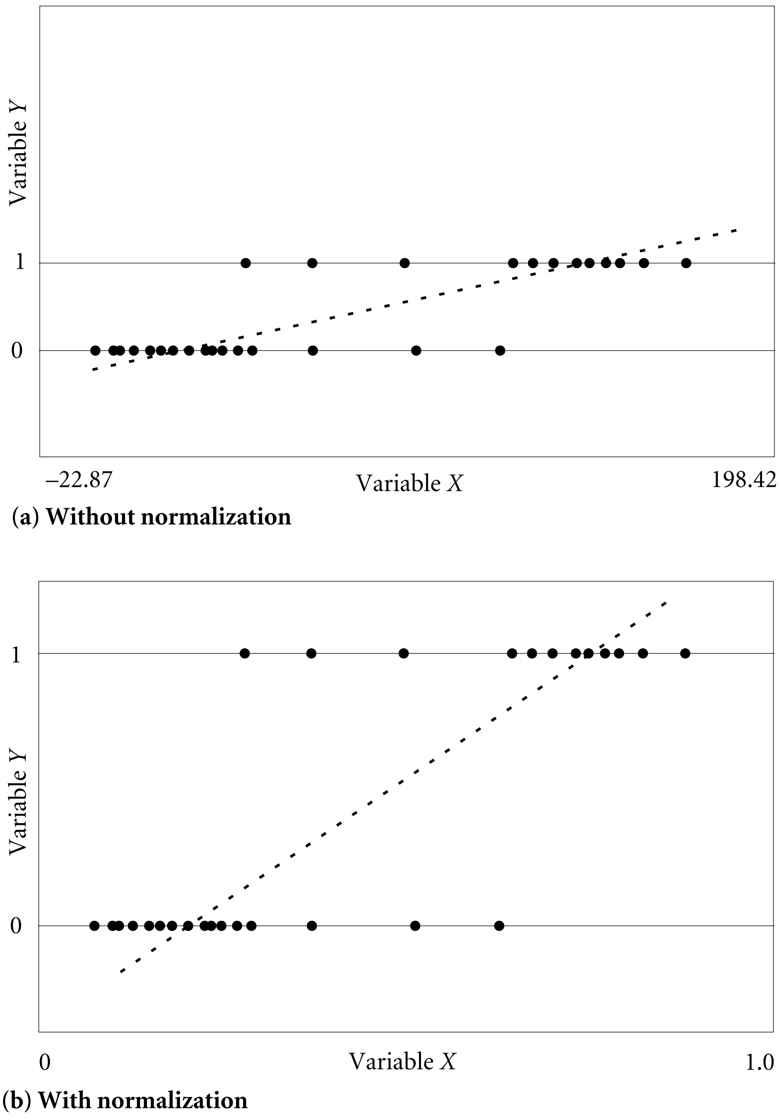
Example 14.1 The income and home value (in $1000s) of 18 individuals are shown in Table 14.3 The value of the correlation coefficient between income and home value is 0.92. Suppose now that home values of two individuals have changed, as shown in bold in Table 14.4; the new data is also shown in bold. The value of the Pearson correlation coefficient has dropped to 0.77, i.e., a difference of 16%. ◆

Table 14.3 Incomes and home values.

Income	52	64	25	37	36	100	99	31	25
Home value	285	364	136	203	269	526	613	192	205
Income	48	**40**	22	83	22	20	37	81	**100**
Home value	194	**364**	165	514	120	129	324	448	**419**

Table 14.4 Values of Table 14.3 with two changes.

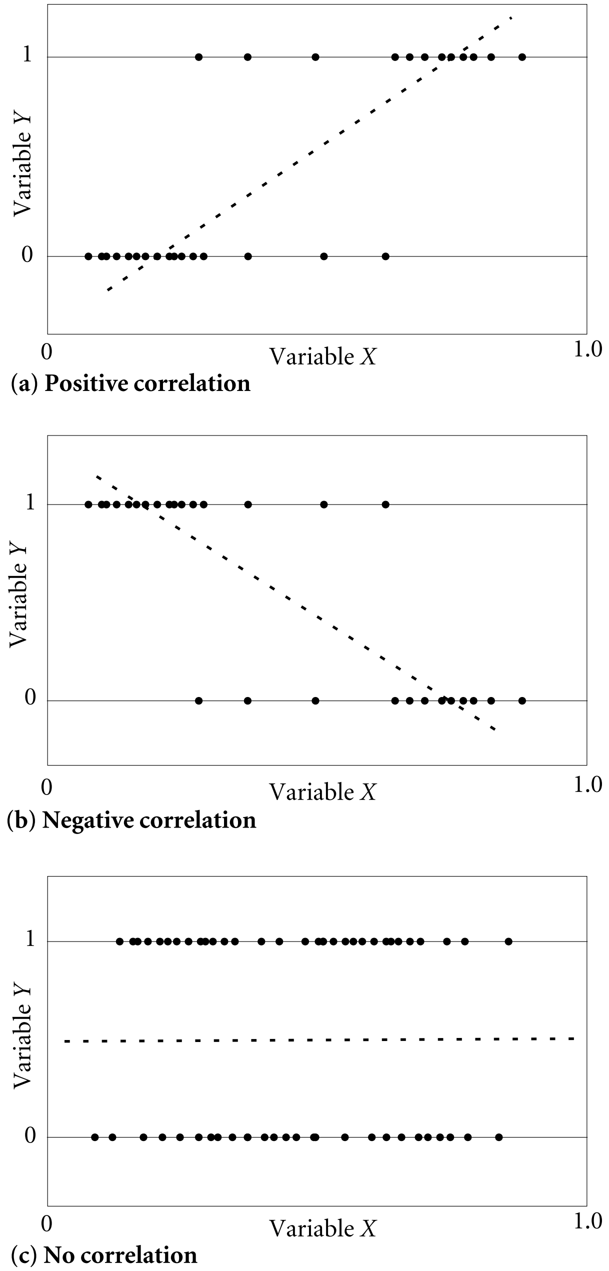
Income	52	64	25	37	36	100	99	31	25
Home value	285	364	136	203	269	526	613	192	205
Income	48	40	22	83	22	20	37	81	100
Home value	194	**759**	165	514	120	129	324	448	**667**



(a) Without normalization

(b) With normalization

Figure 14.2 Correlation between a continuous variable and a binary variable.



(a) Positive correlation

(b) Negative correlation

(c) No correlation

Figure 14.3 Correlation between a continuous variable and a binary variable.

To remedy this weakness of the Pearson correlation coefficient, another coefficient of correlation is suggested based on the concept of the *rank* of the observations with in the variable. This is known as the Spearman rank correlation coefficient. It is defined as

$$\theta = \frac{\sum_{i=1}^{n}(R_i - \bar{R})(S_i - \bar{S})}{\left\{\sum_{i=1}^{n}(R_i - \bar{R})^2 \sum_{i=1}^{n}(S_i - \bar{S})^2\right\}^{1/2}}, \qquad (14.13)$$

where $R_i, S_i, i = 1, \ldots, n$ are the ranks of the variables x and y, and $\bar{R}$ and $\bar{S}$ are the mean values of these ranks, respectively.

Example 14.2 Continuing from Example 14.1, the values of Spearman's coefficient for the two cases are 0.91 and 0.86, respectively. Therefore, the two outliers caused the Spearman's coefficient to drop 5% only. This shows that Spearman's coefficient is less sensitive to such data outliers.
◆

PROC CORR of SAS/STAT calculates these two correlation coefficients, along with other coefficients and tests of their significance. The following macro wraps the procedure and stores the resulting coefficients in a dataset with the appropriate labels (see Table 14.5).

Table 14.5 Parameters of macro VarCorr().

Header	VarCorr(DSin, VarX, VarY, CorrDS);
Parameter	*Description*
DSin	Input dataset
VarX	*X* variable name
VarY	*Y* variable name
CorrDS	Output dataset with correlation coefficients

```
%macro VarCorr(DSin, VarX, VarY, CorrDS);
/* Calculation of the correlation coefficients between
   VarX and VarY in the dataset DSin.
   The results are stored in the CorrDS dataset
   with the names of the coefficients */

/* Step 1: put the variable names in uppercase */
%let x=%upcase(&VarX);
%let y=%upcase(&VarY);
```

```
/* Step 2: invoke proc corr */
proc corr data=&DSin pearson spearman hoeffding kendall
 outp=temp_P outs=temp_S outh=temp_H outk=temp_K noprint;
 var &x &y;
run;
/* Step 3: Get the coefficients from the temporary datasets */
proc sql noprint;
 select &x into : xyP from temp_P where upcase(_NAME_) eq "&y";
 select &x into : xyS from temp_S where upcase(_NAME_) eq "&y";
 select &x into : xyH from temp_H where upcase(_NAME_) eq "&y";
 select &x into : xyK from temp_K where upcase(_NAME_) eq "&y";

 create table &CorrDS (Type char(10), Value num);
 insert into &CorrDS values('Pearson'  , &xyP)
                     values('Spearman' , &xyS)
                     values('Hoeffding', &xyH)
                     values('Kendall'  , &xyK);
quit;
/* Clean the workspace and finish the macro.*/
proc datasets library=work nolist;
  delete temp_P temp_s temp_H temp_K;
quit;

%mend;
```

You should have noticed by now that two more correlation measures were implemented in the macro `VarCorr()`, namely, the *Hoeffding measure of dependence D* and the *Kendall* τ_b. Like the Spearman coefficient, both of these coefficients are nonparametric measures of correlation. In fact, they are *more* nonparametric than the Spearman coefficient. Refer to the SAS/STAT help for more details on these two coefficients.

This Page Intentionally Left Blank