第5章 回归分析

回归分析是使用最为广泛的统计学分支，在质量管理、市场营销、宏观经济管理等领域都有非常广泛的应用。本章介绍一元线性回归、多元线性回归、多项式回归，这三种回归方法应用非常广泛。通过本章的学习，读者可以掌握基本的回归分析原理及应用方法。

5.1 回归分析概述

回归分析(Regression Analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法，应用很广泛。回归分析按照涉及变量的多少，分为一元回归分析和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。如果在回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，则这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且自变量之间存在线性相关，则称为多元线性回归分析。

回归分析主要解决两个问题：一是确定几个变量之间是否存在相关关系，如果存在，则找出它们之间适当的数学表达式；二是根据一个或几个变量的值，预测或控制另一个或几个变量的值，且要估计这种控制或预测可以达到何种精确度。

在经济管理和其他领域中，人们经常需要研究两个或多个变量(现象)之间的相互(因果)关系，并使用数学模型加以描述和解释。例如商品销售量与价格之间的关系。

5.1.1 变量间的两类关系

1. 确定性关系

确定性关系是指当一些变量的值确定以后，另一些变量的值也随之完全确定的关系，这些变量间的关系完全是已知的，变量之间的关系可以用函数关系表示。

例如，圆的面积 S 与半径 r 之间的关系 $S=\pi r^2$；电路中电阻值 R、电压 U 与电流 I 之间的关系 $U=IR$，等等。

图 5-1 表示价格不变时，某商品的销售收入与销售量之间的关系，属于确定性关系。

2. 非确定性关系

非确定性关系是指变量之间有一定的依赖关系，变量之间虽然相互影响和制约，但由于受到无法预计和控制的因素的影响，使变量间的关系呈现不确定性，当一些变量的值确定以后，另一些变量值虽然随之变化，却不能完全确定，这时，变量间的关系就不可以精确地用函

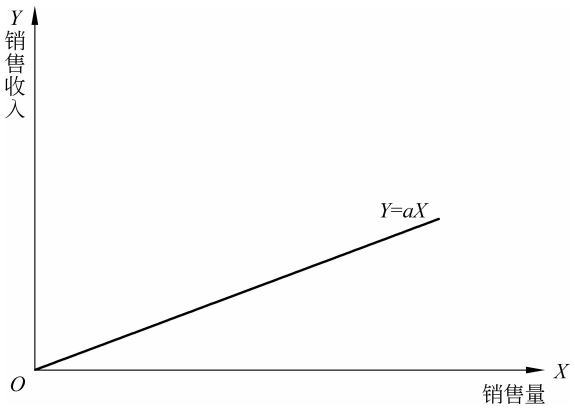

图 5-1　某商品销售收入与销售量的关系

数表示，即不能由一个或若干变量的值精确地确定另一个变量的值。

例如，子女的身高与父母的身高之间有一定的关系，但这种关系不是确定的，即不能根据父亲与母亲的身高精确地得出子女的身高；再如，某块农田粮食的产量与施肥量之间的关系；某件商品的销售量与广告费之间的关系，等等。

5.1.2　回归分析的步骤

回归分析的主要步骤如下。

（1）确定变量

明确预测的具体目标，也就是确定因变量。例如，预测的具体目标是下一年度的销售量，那么销售量 Y 就是因变量。通过市场调查和查阅资料，寻找与预测目标的相关影响因素，即自变量，并从中选出主要的影响因素。

（2）建立预测模型

依据自变量和因变量的历史统计资料进行计算，在此基础上建立回归分析方程，即回归分析预测模型。

（3）进行相关分析

回归分析是对具有因果关系的影响因素（自变量）和预测对象（因变量）所进行的数理统计分析处理。只有当自变量与因变量确实存在某种关系时，建立的回归方程才有意义。因此，作为自变量的因素与作为因变量的预测对象是否有关、相关程度如何以及判断这种相关程度的把握有多大，就成为进行回归分析必须要解决的问题。进行相关分析一般要求出相关关系，以相关系数的大小判断自变量和因变量的相关程度。

（4）计算预测误差

回归预测模型是否可用于实际预测取决于对回归预测模型的检验和对预测误差的计算。回归方程只有通过各种检验，且预测误差较小，才能将回归方程作为预测模型进行预测。

（5）确定预测值

利用回归预测模型计算预测值，并对预测值进行综合分析，确定最后的预测值。

注意：应用回归预测法时，应首先确定变量之间是否存在相关关系。如果变量之间不

存在相关关系，则对这些变量应用回归预测法就会得出错误的结果。

5.2 一元线性回归

一元线性回归分析是处理两个变量之间关系的最简单模型，研究对象是两个变量之间的线性相关关系。通过对这个模型的讨论，不仅可以掌握有关一元线性回归的知识，而且可以从中了解回归分析方法的基本思想、方法和应用。

5.2.1 原理分析

1. 一元线性回归模型

一元线性回归模型只包含一个解释变量（自变量）和一个被解释变量（因变量），是最简单的线性回归模型。一元线性回归模型为

$$Y = a + bX + \varepsilon \tag{5-1}$$

其中，X 为自变量，Y 为因变量；a 为截距，即常量；b 为回归系数，表示自变量对因变量的影响程度；ε 为随机误差项。

一元线性回归模型的特点如下。

① Y 是 X 的线性函数加上误差项。

② 线性部分反映了由于 X 的变化而引起的 Y 的变化。

③ 误差项 ε 是随机变量，反映了除 X 和 Y 之间的线性关系之外的随机因素对 Y 的影响，它是一个期望值为 0 的随机变量，即 $E(\varepsilon)=0$；也是一个服从正态分布的随机变量，且相互独立，即 $\varepsilon \sim N(0,\sigma^2)$。

④ 对于一个给定的 X 值，Y 的期望值为 $E(Y)=a+bX$，称为 Y 对 X 的回归。

2. 回归方程

记 $\hat{a}$，$\hat{b}$ 分别为参数 a 和 b 的点估计，并记 $\hat{Y}$ 为 Y 的条件期望 $E(Y|X)$ 的点估计，由式(5-1)得

$$\hat{Y} = \hat{a} + \hat{b}X \tag{5-2}$$

式(5-2)称为回归方程。其中，$\hat{a}$ 和 $\hat{b}$ 为回归方程的回归系数，$\hat{a}$ 是回归直线在 y 轴上的截距，$\hat{b}$ 是直线的斜率。$\hat{y}$ 表示 X 每变动一个单位时 Y 的平均变动值。对于每一个 x_i 值，由回归方程可以确定一个回归值 $\hat{y}_i=\hat{a}+\hat{b}x_i$。

5.2.2 回归方程求解及模型检验

1. 最小二乘法

可以使用最小二乘法（Least Square Estimation，LSE）求解一元线性回归方程。对每一个点 (x_i,y_i)，y_i 为其实际测量值，$\hat{y}_i$ 是通过式(5-2)得到的预测值。最小二乘法的原理就是找到一组 $\hat{a}$ 和 $\hat{b}$，使所有点的实际测量值 y_i 与预测值 $\hat{y}_i$ 的偏差的平方和最小。其中，称

$\Delta y=y_i-\hat{y}_i$ 为残差。残差平方和(Residual Sum of Squares,RSS)的定义为

$$Q(\hat{a},\hat{b})=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2 \tag{5-3}$$

由式(5-3)知,$Q(\hat{a},\hat{b})$是关于 $\hat{a}$ 和 $\hat{b}$ 的二次函数,所以 $Q(\hat{a},\hat{b})$存在最小值。由微积分知识,分别对 $\hat{a}$ 和 $\hat{b}$ 求一阶偏导并令其一阶偏导值为 0。即

$$\frac{\partial Q}{\partial \hat{a}}=-2\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)=0 \tag{5-4}$$

$$\frac{\partial Q}{\partial \hat{b}}=-2\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)x_i=0 \tag{5-5}$$

式(5-4)和式(5-5)称为正规方程组,据此可求解出 $\hat{a}$ 和 $\hat{b}$ 的值为

$$\hat{a}=\bar{y}-\hat{b}\bar{x} \tag{5-6}$$

$$\hat{b}=\frac{\bar{x}\cdot\bar{y}-\overline{xy}}{\bar{x}^2-\overline{x^2}} \tag{5-7}$$

其中 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$,$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i$,$\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i$,$\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2$。

将求得的 $\hat{a}$ 和 $\hat{b}$ 的值代入方程 $y=\hat{a}+\hat{b}x$ 中,得到的方程就是最佳拟合曲线。

2. 拟合优度检验

拟合优度指所求得的回归直线对观测值的拟合程度。若观测值与回归直线之间的距离近,则认为拟合优度较好,反之则较差,这里用决定系数(Coefficient of Determination)度量拟合优度。

首先给出离差、回归差、残差的概念。

离差定义为 $y_i-\bar{y}$,表示实际值与平均值之差。

回归差定义为 $\hat{y}_i-\bar{y}$,表示估计值与平均值之差。

残差定义为 $y_i-\hat{y}_i$,表示实际值与估计值之差。

其中,离差=回归差+残差。三者的关系如图 5-2 所示。

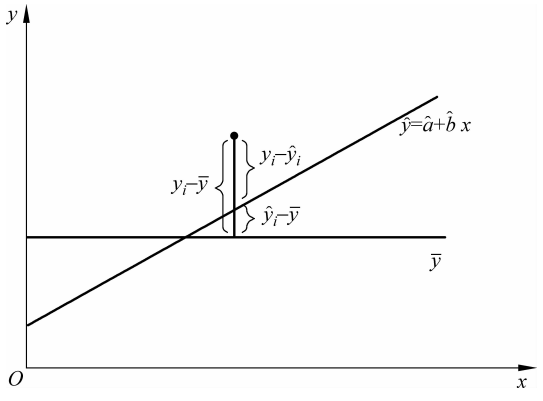

图 5-2　离差、回归差、残差三者的关系

用总平方和(Total Sum of Squares,TSS)表示因变量的 n 个观察值与其均值的误差的总和,TSS 是各个数据离差的平方和,即

$$\mathrm{TSS}=\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{5-8}$$

用回归平方和(Explained Sum of Squares,ESS)表示自变量 x 的变化对因变量 y 取值变化的影响,ESS 是各个数据回归差的平方和,即

$$\mathrm{ESS}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{5-9}$$

用残差平方和(Residual Sum of Squares,RSS)表示实际值与拟合值之间的差异程度,RSS 是各个数据残差的平方和。即

$$\mathrm{RSS}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5-10}$$

TSS、ESS、RSS 三者之间的关系为 TSS=ESS+RSS 即

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{5-11}$$

证明:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}[(y_i-\hat{y}_i)+(\hat{y}_i-\bar{y})]^2$$

$$=\sum_{i=1}^{n}[(y_i-\hat{y}_i)^2+2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})+(\hat{y}_i-\bar{y})^2]$$

$$=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2+\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$=\mathrm{RSS}+\mathrm{ESS}+2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$$

$$2\sum_{i=1}^{n}(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})=2\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)(\hat{a}+\hat{b}x_i-\bar{y})$$

$$=\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)(\hat{a}-\bar{y})+\hat{b}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)x_i=0$$

由式(5-4)和式(5-5)知,$\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)=0$,$\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)x_i=0$,$(\hat{a}-\bar{y})$ 和 $\hat{b}$ 为常数。

证明结束。

拟合优度(Goodness of Fit)是指回归直线对观测值的拟合程度。度量拟合优度的统计量是决定系数(也称确定系数)R^2,计算公式如式(5-12)所示。

$$R^2=\frac{\mathrm{ESS}}{\mathrm{TSS}}=\frac{\mathrm{TSS}-\mathrm{RSS}}{\mathrm{TSS}}=1-\frac{\mathrm{RSS}}{\mathrm{TSS}} \tag{5-12}$$

其中:$R^2\in[0,1]$,R^2 越接近于 1,说明回归曲线拟合度越好;R^2 越小,说明回归曲线拟合度越差。$R^2=0$ 时,表示自变量 x 与因变量 y 没有线性关系。当 $R^2=1$ 时,表示回归曲线与样本点重合。

3. 线性关系的显著性检验

采用 F 检验度量一个或多个自变量同因变量之间的线性关系是否显著。F 检验(F test)运用服从 F 分布的统计量或方差比作为统计检验,通过**显著性水平**(Significant Level)**检验**度量回归方程的线性关系是否显著。

F 检验的计算方式为

$$F = \frac{\mathrm{ESS}/k}{\mathrm{RSS}/(n-k-1)} \tag{5-13}$$

且服从 F 分布 $F=(k,n-k-1)$。

其中 k 为自由度(自变量的个数),n 为样本总量。一元线性回归方程只有一个自变量 x,所以 $k=1$。

F 值越大,说明自变量和因变量之间在总体上的线性关系越显著,反之线性关系越不显著。

4. 回归参数的显著性检验

采用 t 检验对回归参数进行显著性检验,t 检验检测变量 x 是否是被解释变量 y 的一个显著性的影响因素,t 检验是用于样本的两个平均值差异程度的检验方法,它使用 T 分布理论推断差异发生的概率,从而判断两个平均数的差异是否显著。

t 检测的计算方式为

$$t_i = \frac{\hat{b}_i}{s_{\hat{b}_i}} \tag{5-14}$$

其中,$s_{\hat{b}_i}$ 的计算公式如式(5-15)所示。

$$s_{\hat{b}_i} = \frac{\sqrt{\dfrac{\mathrm{RSS}}{n-k-1}}}{\sqrt{\sum_{i=1}^{n} x_i^2 - \dfrac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}} \tag{5-15}$$

其中,$\hat{b}_i$ 是自变量 $\hat{x}_i$ 的回归参数,$s_{\hat{b}_i}$ 是回归参数 $\hat{b}_i$ 的抽样分布的标准差,k 为自由度,n 为样本总量,RSS 为残差平方和。

对于一元线性回归模型,只有一个自变量 x_i,所以 $\hat{b}_i=\hat{b}$,自由度 $k=1$。

如果某个自变量 x_i 对因变量 y 没有产生影响或者影响很小,应当将自变量 x_i 的系数取值为 0,即 $\hat{b}_i=0$。

5.2.3　一元线性回归实例

例 5.1　某种商品与家庭平均消费量的关系。

以某家庭为调查单位,某种商品在某年各月的家庭平均月消费量 Y(kg)与其价格 X(元/kg)之间的调查数据如表 5-1 所示。

图 5-3 为该商品的家庭平均月消费量与价格之间呈现的关系。

表 5-1 商品价格与消费量的关系

价格 X	5.0	5.2	5.8	6.4	7.0	7.0	8.0	8.3	8.7	9.0	10.0	11
消费量 Y	4.0	5.0	3.6	3.8	3.0	3.5	2.9	3.1	2.9	2.2	2.5	2.6

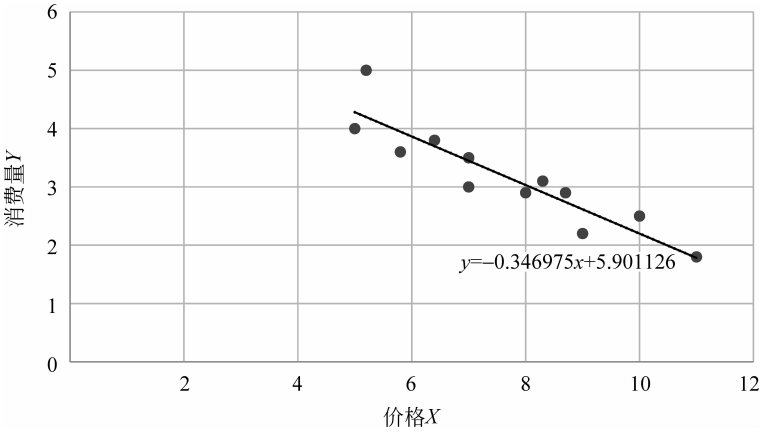

图 5-3 商品价格与消费量之间的线性关系

由图 5-3 可知，该商品在某家庭月平均消费量 Y 与价格 X 间基本呈线性关系，这些点与直线间的偏差是由其他一些无法控制的因素和观察误差引起的，根据 Y 与 X 之间的线性关系及表 5-1 中数据可以求得两者之间的回归方程。

(1) 求解一元线性回归方程

解：根据表 5-1 中的数据求得 $\bar{x}$、$\bar{y}$、$\overline{xy}$、$\overline{x^2}$。

$$\bar{x} = \frac{1}{12}(5.0+5.2+5.8+6.4+7.0+7.0+8.0+8.3+8.7+9.0+10.0+11)$$
$$= 7.616667$$

$$\bar{y} = \frac{1}{12}(4.0+5.0+3.6+3.8+3.0+3.5+2.9+3.1+2.9+2.2+2.5+2.6)$$
$$= 3.258333$$

$$\overline{xy} = \frac{1}{12}(5.0\times4.0+5.2\times5.0+5.8\times3.6+6.4\times3.8+7.0\times3.0+$$
$$7.0\times3.5+8.0\times2.9+8.3\times3.1+8.7\times2.9+9.0\times2.2+10.0\times$$
$$2.5+11\times2.6)$$
$$= 23.688333$$

$$\overline{x^2} = \frac{1}{12}(5.0^2+5.2^2+5.8^2+6.4^2+7.0^2+7.0^2+8.0^2+8.3^2+$$
$$8.7^2+9.0^2+10.0^2+11^2) = 61.268333$$

根据 $\bar{x}$、$\bar{y}$、$\overline{xy}$、$\overline{x^2}$求解 $\hat{b}$ 的值。

$$\hat{b} = \frac{7.616667\times3.258333-23.688333}{7.616667^2-61.268333} = -0.346975$$

根据 $\hat{b}$、$\bar{x}$、$\bar{y}$ 求解 $\hat{a}$。

$$\hat{a} = 3.258333 - (-0.346975) \times 7.616667 = 5.901126$$

故求得的线性回归方程为 $y=-0.346975x+5.901126$。

(2) 回归方程拟合优度检验

例 5-1 中，价格 x、平均月销售量预测 $\hat{y}$、实际销售量 y 的数据集如表 5-2 所示。

表 5-2　x、$\hat{y}$ 与 y 的数据集

x	$\hat{y}$	y
5.0	4.166251	4.0
5.2	4.096856	5.0
5.8	3.888671	3.6
6.4	3.680486	3.8
7.0	3.472301	3.0
7.0	3.472301	3.5
8.0	3.125326	2.9
8.3	3.0212335	3.1
8.7	2.8824435	2.9
9.0	2.778351	2.2
10.0	2.431376	2.5
11.0	2.084401	2.6

解：根据式(5-8)和式(5-9)求解 TSS、ESS。

$$\begin{aligned}\text{TSS} =& (4.0-3.258333)^2+(5.0-3.258333)^2+(3.6-3.258333)^2+\\ &(3.8-3.258333)^2+(3.0-3.258333)^2+(3.5-3.258333)^2+\\ &(2.9-3.258333)^2+(3.1-3.258333)^2+(2.9-3.258333)^2+\\ &(2.2-3.258333)^2+(2.5-3.258333)^2+(2.6-3.258333)^2\\ =&6.263581\end{aligned}$$

$$\begin{aligned}\text{ESS} =& (4.166251-3.258333)^2+(4.096856-3.258333)^2+\\ &(3.888671-3.258333)^2+(3.680486-3.258333)^2+\\ &(3.472301-3.258333)^2+(3.472301-3.258333)^2+\\ &(3.125326-3.258333)^2+(3.0212335-3.258333)^2+\\ &(2.8824435-3.258333)^2+(2.778351-3.258333)^2+\\ &(2.431376-3.258333)^2+(2.084401-3.258333)^2\\ =&4.702097\end{aligned}$$

根据式(5-12)求解 R^2。

$$R^2 = \frac{4.702097}{6.263581} = 0.750704$$

R^2 接近于 1，说明该回归方程拟合度较好。

(3) 回归方程线性关系的显著性检验

对于例 5-1，求解其 F 值。

解：首先求解 F 分布的值。

例 5-1 中，$k=1$，$n=12$，假设 $\alpha=0.05$，经查 F 值表有：

$$F_{0.05}(k,n-k-1)=F_{0.05}(1,10)=7.507$$

然后根据式(5-11)利用 ESS、TSS 和 RSS 三者之间的关系求解 RSS。

$$\mathrm{RSS}=6.263581-4.702097=1.561484$$

最后根据式(5-13)求解 F 值。

$$F=\frac{4.702097/1}{1.561484/(12-1-1)}=30.128$$

求得的 F 值为 $30.128>F_{0.05}(1,10)=7.507$，所以在显著性概率为 0.05 的条件下，回归方程显著成立。

(4) 回归参数的显著性检验

对于例 5-1，求解其 t 值。

解：首先根据 t 分布表求解 t 分布值。

例 5-1 中，$n=12$，在**置信度水平**(Confidence Level)为 0.05 的情况下，经查 t 分布表知 t 值为 1.782。

然后根据式(5-15)求解得 $s_{\hat{b}_i}$ 的值为 0.068396。

最后，根据式(5-14)求得 t 值为

$$t=\frac{-0.346975}{0.068396}=-5.07301$$

$|t|=5.07301>1.782$。所以变量 x 对于因变量 y 有显著影响。

其中，置信度水平是指总体参数值落在样本统计值某一区内的概率，用来表示区间估计的把握程度。假设置信度水平为 0.05，表示真值发生的概率为 95‰。

5.2.4 案例分析：使用 Weka 实现一元线性回归

例 5.2 信用卡积分与月收入之间的线性关系。

某家银行想统计信用卡积分与使用者月收入之间的关系，现有一个文件 bank.arff，该文件包含 7 个属性(月收入、每月工作天数、当前信用卡额度、历史统计的按时还款比例、曾经的最大透支额、银行贷款的数目、信用卡积分)，但是银行只想统计信用卡积分与月收入之间的关系，所以在构建模型的时候需要去除其余 6 个属性的影响，只留下“月收入”这一个属性。

该文件为自定义文件，文件 bank.arff 的内容如下。

```
@RELATION creditCardScore
%%%%
%SECTION1:PERSONAL INFO
%%%%
%
%月收入
%
@ATTRIBUTE personInfo.monthlySalary  NUMERIC
%%%%
%SECTION2: BUSINESS INFO
%%%%
%
```

```
%每月工作天数
%
@ATTRIBUTE businessInfo.workingDayPerMonth NUMERIC
%%%%
%SECTION 3: CREDIT CARD INFO(信用卡信息)
%%%%
%
%当前额度
%
@ATTRIBUTE creditCardInfo.currentLimit  NUMERIC
%
%月度正常还款比例
%
@ATTRIBUTE creditCardInfo.percentageOfNormalReturn NUMERIC
%
%曾经最大透支额
%
@ATTRIBUTE creditCardInfo.maximumOverpay NUMERIC
%%%%
%SECTION 4: FINANCIAL INFO(财政信息)
%%%%
%
%贷款数目
%
@ATTRIBUTE financialInfo.personalLoan NUMERIC
%%%%
%RESULT: CREDIT SCORE(积分)
%%%%
@ATTRIBUTE creditScore NUMERIC
@DATA
10000,22,20000,1,0,200000,55
15000,20,30000,0.5,14200,20000, 78
20000,18,40000,0.6,50000,200000,87
30000,22,60000,0.2,30000,150000,67
22000,15,30000,0.7,20000,140000,71
13200,21,18000,0.9,40000,500000,43
15500,20,30000,0.4,14200,20000, 59
25000,26,40000,0.5,50000,200000,88
28670,23,40000,0.7,30000,120000,68
22000,15,40000,0.7,20000,140000,72
10000,18,20000,0.6,30000,150000,47
14300,20,29800,0.5,14200,20000, 72
20000,18,40000,0.9,50000,200000,88
34335,22,50000,0.6,30000,150000,74
24555,15,20000,0.9,20000,120000,79
```