变量选择

1.皮尔逊相关系数

两个变量

衡量的是变量间的线性相关程度，取值范围$[-1,1]$

0为无相关性

接近-1,1为强相关性

$$\rho=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{[\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2]^{1/2}}$$

> x_i和y_i分别为变量x和y的取值
>
> $\bar{x}$和$\bar{y}$分别为变量x和y的均值

2.斯皮尔曼相关系数

两个顺序变量

取值升序排列时，取值的等级就是该取值的顺序

$$\rho=\frac{\sum_{i=1}^{N}(R_i-\bar{R})(S_i-\bar{S})}{[\sum_{i=1}^{N}(R_i-\bar{R})^2\sum_{i=1}^{N}(S_i-\bar{S})^2]^{1/2}}$$

> R_i和S_i分别为观测值i取值的等级
>
> $\bar{R}$和$\bar{S}$分别为变量x和y的平均等级

3.卡方统计量（皮尔逊卡方统计量）

两个名义（顺序）变量

x/y	y_1	$\cdots$	y_c	合计
x_1	n_{11}	$\cdots$	n_{1c}	n_{1*}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
x_r	n_{r1}	$\cdots$	n_{rc}	n_{r*}
合计	n_{*1}	$\cdots$	n_{*c}	N

> x变量分为r类
>
> y变量分为c类
>
> 各变量记录的频率数为n_{ij}
>
> 每一行的合计频率数为$n_{1*},\cdots,n_{*c}$

数据集中观测值的总量是N

$$N=\sum_{i=1}^{r}\sum_{j=1}^{c}n_{ij}$$

μ_{ij}用第i行和第j列中记录的总数计算第i行和第j列的预期单位数

$$\mu_{ij}=\frac{n_{i*}n_{*j}}{N}$$

皮尔逊卡方统计量定义如下：

$$\chi^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$$

4.概率比统计量

逻辑回归

改变变量x和变量y中的任意一个或全部两个类比的顺序不改变概率比的取值

x	y违约(1)	y违约(2)	合计
x_1	n_{11}	n_{12}	n_{1*}
x_2	n_{21}	n_{22}	n_{2*}
合计	n_{*1}	n_{*2}	n

> 当变量x取值为x_1，违约结果比率为n_{11}/n_{12}
>
> 当变量x取值为x_2,违约结果比率为n_{21}/n_{22}
>
> 概率比(θ)的定义为这以上两个数值的比率

$$\theta=\frac{n_{11}/n_{12}}{n_{21}/n_{22}}=\frac{n_{11}n_{22}}{n_{12}n_{21}}$$

5.F检验

衡量一个连续变量与名义变量之间的相关性和关联性

哪一个变量为因变量均可

假设连续变量用y表示，名义变量用x表示

> $y_i=\sum_{j=1}^{n_j}y_{ij}$,值的平均值为、$\bar{y_i}=\frac{y_i}{n_i}$
>
> 变量y的所有值得和为$y=\sum_{i=1}^{r}y_i$
>
> 变量y的总体平均值表示为$\bar{y}=\frac{y}{N}$

定义名义变量x每个类别平均值的离差平方的加权总和为：

$$SSTR=\sum_{i=1}^{r}n_i(\bar{y_i}-\bar{y}^2)$$

名义变量x对应的连续变量y所有取值的离差平方总和为：

$$SSE=\sum_{i=1}^{r}\sum_{j=1}^{n_i}(y_{ij}-\bar{y_i})^2$$

名义变量x对应的连续变量y所有取值的离差平方和为：

$$STD=\sum_{i=1}^{r}(\sum_{j=1}^{n_i}(y_{ij}-\bar{y_i})^2)$$

将$SSTR$和SSE的均值分别定义为：

$$MSSTR=\frac{SSTR}{r-1}$$

$$MMSE=\frac{SSE}{N-r}$$

故，F检验的统计量定义为：

$$F=\frac{MSSTR}{MSSE}$$

判断标准：F值越大，表明变量间关联性越大

6.基尼方差

一个连续变量和一个名义变量

两个名义变量

两个顺序变量

$$G_r=1-\frac{SSE}{STD}$$

7.信息值

两个名义变量（其中一个是二元变量），二元变量取值为0和1

$$IV=\sum_{i=1}^{r}(p_i-q_i)loh(\frac{p_i}{q_i})$$

> p_i和q_i分别为第i行中变量y第一类和第二类记录的百分比
>
> $p_i=\frac{n_{i1}}{n_{*i}}$
>
> $q_i=\frac{n_{i2}}{n_{*2}}$

标准：IV的预测力解释

IV的范围	预测力
小于0.02	无预测力
0.02~0.10	较弱
0.10~0.30	中等
大于0.30	较强