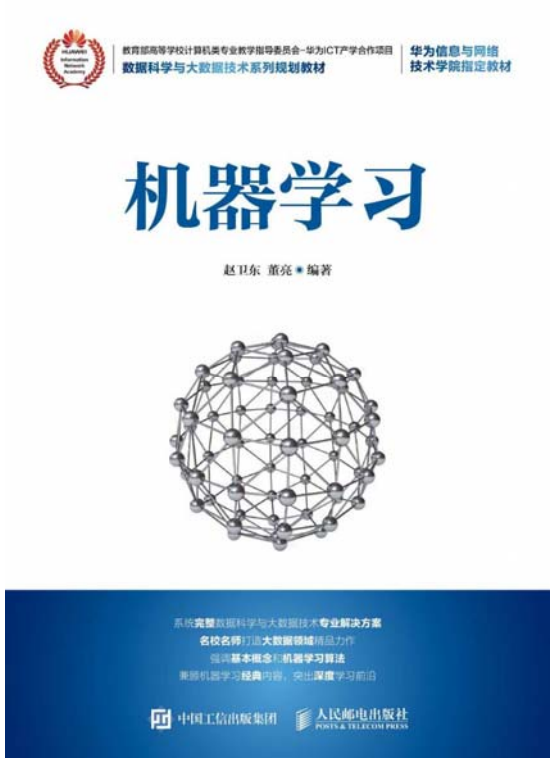


机器学习
回归分析

复旦大学 **赵卫东** 博士

wdzhao@fudan.edu.cn

章节结构

- 统计分析
  - 统计基础
  - 常见概率分布
  - 参数估计
  - 假设检验
  - 线性回归
  - Logistics回归

统计分析

- 统计学是研究如何搜集资料、整理资料和进行量化分析、推断的一门科学，在科学计算、工业和金融等领域有着重要应用，统计分析是机器学习的基本方法
- 与统计分析相关的基本概念有以下几个
  - 总体：根据定目的确定的所要 研究事物的全体
  - 样本：从总体中随机抽取的若干个体构成的集合
  - 推断：以样本所包含的信息为基础对总体的某些特征作出判断、预测和估计
  - 推断可靠性：对推断结果从概率上的确认，作为决策的重要依据
- 统计分析分为描述性统计和推断性统计，描述性统计是通过对样本进行整理、分析并就数据的分布情况获取有意义的信息，从而得到结论。推断统计又分为参数估计和假设检验，参数估计是对样本整体中某个数值进行估计，如推断总体平均数等，而假设检验是通过对所做的推断验证，从而进择行才方案

统计基础

- 输入空间、特征空间和输出空间
  - 向量空间模型包括输入空间、特征空间与输出空间，输入与输出所有的可能取值的集合分别称为输入空间与输出空间，每个具体的输入是一个实例， 通常由特征向量表示，所有特征向量存在的空间成为特征空间。输入变量用一般用xx表示，输出变量用y表示
- 联合概率分布
  - 在监督式学习中是假设输入与输出的变量x和y遵循联合概率分布$p（x，y）$，表示样本数据存在一定的规律，可以假定这个联合概率分布的存在，但是其分布是未知的，x和y具有联合概率分布的假设就是监督学习关于数据的基本假设
- 假设空间
  - 机器学习模型是由输入空间到输出空间的映射的集合，这个集合就是假设空间。假设空间确定了预测的范围。监督学习的目标是学习一个由输入到输出的映射规律，这个映射规律就是模型。监督学习的模型包括板率模型、非概率模型，前者由条件概率分布$p（y|x）$表示，后者由函数$y = f（x）$表示，模型确认之后就可以对具体的输入进行相应的输出预测

统计基础

- 均值、标准差、方差、协方差
  - 均值描述的是样本集合的平均值
  - 标准差描述是样本集合的各个样本点到均值的距离分布，描述的是样本集的分散程度
  - 在机器学习中的方差就是估计值与其期望值的统计方差。如果进行多次重复验证的过程，就会发现模型在训练集上的表现并不固定，会出现波动，这些波动越大，它的方差就越大
  - 协方差主要用来度量两个随机变量关系，如果结果为正值，则说明两者是正相关的；结果为负值，说明两者是负相关的；如果为0，就是统计上的“相互独立”
- 超参数
  - 超参数是机器学习算法的调优参数，常应用于估计模型参数的过程中，由用户直接指定，可以使用启发式方法来设置，并能依据给定的预测问题而调整
  - 超参数与模型参数不同，模型参数是学习算法拟合训练数据获得的参数，即这些参数是作为模型本司身的参数而存在的

统计基础

- 损失函数和风险函数
  - 损失函数是关于模型计算结果f（x）和样本实际目标结果Y的非负实值函数，记作L（y，f（x）），用它来解释模型在每个样本实例上的误差损失函数的值越小，说明预测值与实际值越接近，即模型的拟合效果越好
  - 损失函数主要包括以下几种： 0-1损失函数、平方损失函数、绝对损失函数、对数损失函数
- 训练误差

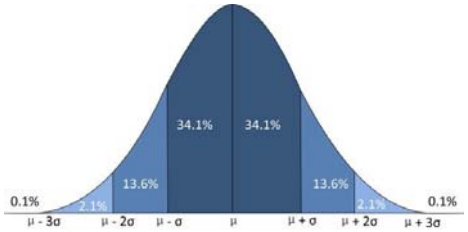
$$R_{exp} = \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y, f(x))$$

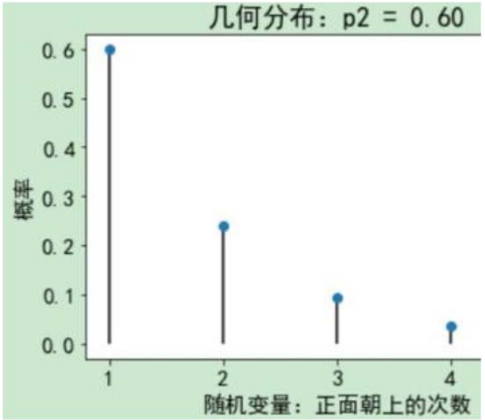
统计基础

- 正则化与交叉验证
  - L1正则化
  - L2正则化
  - HoldOut检验
  - 简单交叉检验
  - K折交叉检验
  - 留一交叉检验

常见概率分布

- 均匀分布
- 正态分布
- *t*分布
- 卡方分布
- F-分布
- 二项分布
- 0-1分布
- Poisson分布





概率分布表

分布名称	概率与密度函数 $p(x)$	数学期望	方差	图形
贝努里分布 两点分布	$p_k=\begin{cases}q, & k=0\\ p, & k=1\end{cases}$ $0<p<1,\ q=1-p$	p	pq	
二项分布 $b(k;n,p)$	$b(k;n,p)=\binom{n}{k}p^k q^{n-k}$ $k=0,1,\cdots,n$ $0<p<1,\ q=1-p$	np	npq	
泊松分布 $p(k;\lambda)$	$p(k,\lambda)=\frac{\lambda^k}{k!}e^{-\lambda}$, $\lambda>0$ $k=0,1,2,\cdots,n$	λ	λ	
指数分布	$p(x)=\begin{cases}\lambda e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$ $\lambda>0$，常数	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	
χ^2-分布	$p(x)=\begin{cases}\frac{1}{2^{n/2}\Gamma(\frac{n}{2})}x^{\frac{n}{2}}e^{-x/2}, & x\geq 0\\ 0, & x<0\end{cases}$ n正整数	n	$2n$	
Γ-分布	$p(x)=\begin{cases}\frac{\lambda^r}{\Gamma(r)}x^{r-1}e^{-\lambda x}, & x\geq 0\\ 0, & x<0\end{cases}$ $r>0$，$\lambda>0$常数	$r\lambda^{-1}$	$r\lambda^{-2}$	

参数估计

- 参数估计是用样本统计量去估计总体的参数，即根据样本数据选择统计量去推断总体的分布或数字特征
- 估计参数的目的，是希望用较少的参数去描述数据的总体分布，前提是要了解样本总体分布（如正态分布），这样就只需要估计其中参数的值。如果无法确认总体分布，那就要采用非参数估计的方法
- 参数估计是统计推断的种基本形式， 分为点估计和区间估计两部分。其中有多种方法，除了最基本的最小二乘法和极大似然法、贝叶斯估计、极大后验估计，还有矩估计、一致最小方差无偏估计、最小风险估计、最小二乘法、最小风险法和极小化极大熵法等

假设检验

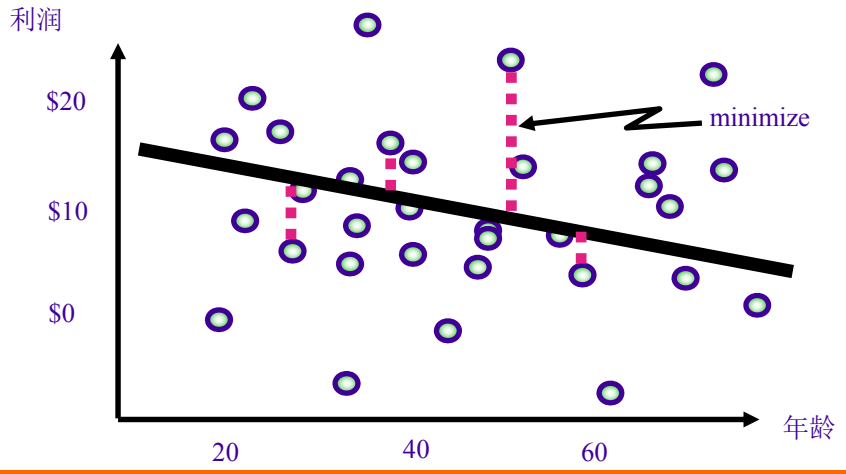
- 假设检验假设检验是先对总体的参数（或分布形式）提出某种假设，然后利用样本信息判断假设是否成立的过程。假设检验的基本思想是小概率反证法思想
- 假设检验包括原假设与备选假设。其中检验假设正确性的是原假设， 表明未知参数的看法。而备选假设通常反映研究者对参数可能数值对立的看法
- 假设检验的具体过程如下：首先所研究问题的总体做某种假设，记作HO；选取合适的统计量，这个统计量的选取要使得在假设HO成立时，其分布为已知；由实测的样本，计算出统计量的值，并根据预先给定的显著性水平进行检验，做出拒绝或接受假设HO的判断
- 常用的假设检验方法有u检验法、t检验法、x_2检验法（卡方检验）、F检验法、秩和检验等

假设检验

- 显著性检验是根据一定的理论或经验， 认为某一 假设 HO成立。例如，首先假设人的收入是服从F在分布的。当收集了一定的收入数据后。可以评价实际数据与理论假设HO >间的偏离，如果偏离达到了“显著”的程度就拒绝H10假设，这样的检验方法称为显著性检验
- 显著程度从中心的H0“非常显著”开始向外不断移动，当偏离达到某一较低显著的程度α（如0.05）时，再看H0假设，已经很难证明其正确了，这时就可以认为H0假设不成立，也就是被拒绝了， 就是它成立的概率不超过α，称α为显著性水平。这种假设检验的好处是不用考虑备择假设，只关心实验数据与理论之间拟合的程度，所以也称之为拟合优度检验

回归分析

- 分析一个变量与其他一个（或几个）变量之间的相关关系的统计方法就称为回归分析。常见的回归分析包括线性回归、多元回归、非线性回归、广义线性回归（对数回归、泊松回归）等。回归分析主要内容包括确定连续值变量之间的相关关系，建立回归模型，检验变量之间的相关程度，应用回归模型对变量进行预测等。

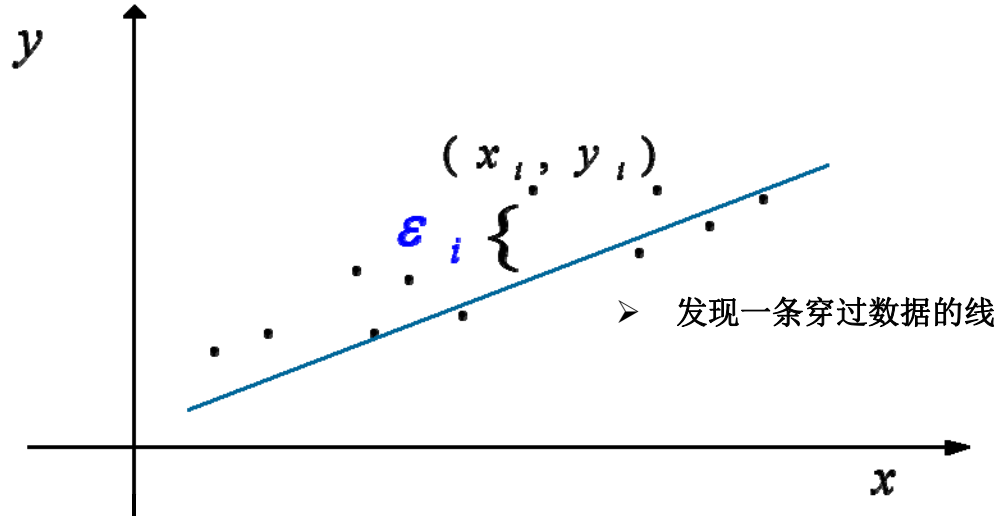


线性回归

- 线性回归是种通过拟合自变量与因变量之间最佳线性关系，来预测目标变量的方法
- 回归过程是给出一个样本集， 用函数拟合这 个样本集，使样本集与拟合函数间的误差最小
- 回归分析包括以下内容
  - 确定输入变量与目标变量间的回归模型，即变量间相关关系的数学表达式
  - 根据样本估计并检验回归模型及未知参数
  - 从众多的输入变量中，判断哪些变量对目标变量的影响是显著的
  - 根据输入变量的已知值来估计目标变量的平均值并给出预测精度
- 线性回归的类型包括简单线性回归和多元线性回归
  - 简单线性回归使用一个自变量，通过拟合最佳线性关系来预测因变量
  - 多元线性回归使用多个独立变量，通过拟合最佳线性关系来预测因变量

一元线性回归

- 一元线性回归是描述两个变量之间线性相关关系的最简单的回归模型，如下图。在散点图中两个变量呈线性关系。一元线性回归模型表示为$y=a+bx+\varepsilon$，其中a和b是系数，是随机变量。在这个线性模型中，自变量x是非随机变量。随机变量要求服从正态分布。



- 发现一条穿过数据的线，线上的点使对应数据点的方差最小。

一元线性回归

- 确定参数a和b（分别记作 和 $\hat{a}$）值的原则是使样本的回归直线同观察值的拟合状态最好，即使偏差| |较小。ε_i为此，可以采用最小二乘法计算。对应于每一个x_i，根据回归方程可以求出一个，它就是y_i的一个估计值。有n个观察值就有相应的n个偏差。为了计算方便，以偏差的平方和最小为标准确定回归模型：

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\frac{\partial Q}{\partial a} = 2\sum_{i=1}^{n} [y_i - (a + bx_i)] \cdot (-1) = 0$$

$$\frac{\partial Q}{\partial b} = 2\sum_{i=1}^{n} [y_i - (a + bx_i)] \cdot (-x_i) = 0$$

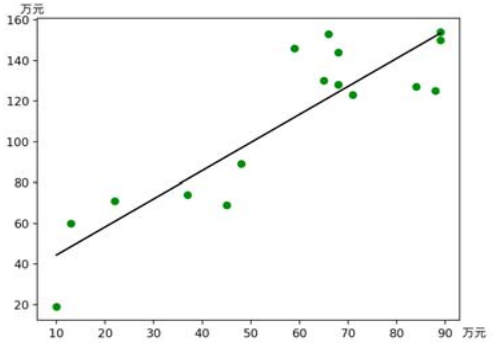
- 得到参数和的最小二乘估计：$\hat{b} = S_{xy} / S_{xx}$，$\hat{a} = \overline{y} - \hat{b}\overline{x}$ ，式中$\overline{x}, \overline{y}$分别是变量$x$，$y$的$n$个样本的平均值，$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ ，$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$。

线性回归示例

- 已知一个贸易公司某几个月的广告费用和销售额，如下表所示

广告费（万元）	10	13	22	37	45	48	59	65	66	68	68	71	84	88	89	89
销售额（万元）	19	60	71	74	69	89	146	130	153	144	128	123	127	125	154	150

- 可见随着广告费用的增加，公司的销售额也在增加，但是它们并非绝对的线性关系，而是趋向于平均，如下图所示



- 上述线性回归模型的公司为：y=1.38*x+30.6，其中x表示广告费用，y表示销售额，通过线性回归的公式就可以预测企业的销售额了

多元线性回归（1））

- 多元线性回归分析是研究一个变量y与多个其他变量$x_1, x_2, ..., x_k$之间关系的统计分析方法。假设因变量y与自变量$x_1, x_2, ..., x_k$之间有线性关系$y=\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k+u$，其中$\beta_0, \beta_1, \beta_2, ..., \beta_k$是回归系数，$u$为随机误差。上面的公式一般称为多元线性回归模型。由于可以利用已知样本数据进行估计。
- 设$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$是利用一组简单随机样本经计算得到的样本统计量，把它们作为未知参数$\beta_0, \beta_1, \beta_2, ..., \beta_k$的估计值，得到估计的回归方程$\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2+...+\hat{\beta}_k x_k$称为样本回归方程或经验回归方程，$\hat{y}$称为y的样本估计值或样本回归值。

多元线性回归（2）

- 设 $(x_{1i}, x_{2i}, \ldots, x_{ki}; y_i)$,其中 $i = 1, 2, \ldots, n$ 是对因变量y和自变量 $x_1, x_2, \ldots, x_k$ 的*n*次独立样本观测值，代入多元线性回归模型得到

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + u_i, i = 1, 2, \ldots, n$$

它是由*n*个方程组成的一个线性方程组：

$$\begin{cases} y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + u_1 \\ y_2 = \beta_0 + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + u_2 \\ \ldots\ldots \\ y_n = \beta_0 + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + u_n \end{cases}$$

多元线性回归（3）

- 表示成矩阵形式为$Y = X\beta + u$，其中

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{bmatrix}_{n\times 1} \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix}_{n\times(k+1)} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \cdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1} \quad u = \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{bmatrix}_{n\times 1}$$

- 这里*Y*是因变量样本观测值的$n\times(k+1)$阶列向量，*X*是自变量样本观测值的$n\times(k+1)$阶矩阵，它的每个元素x_{ij}都有两个下标，第一个下标*i*表示相应的列（第i个变量），第二个下标j表示相应的行（第j个观测值）。*X*的每一列表示一个自变量的*n*个观测值向量，β为未知参数的$(k+1)\times 1$阶列向量，*u*为随机误差项的$n\times 1$阶列向量。把样本数据代入$Y = X\beta + u$，得到$\hat{\beta} = (X'X)^{-1}X'Y$，式中$X'$表示*X*的转置，而$(X'X)^{-1}$表示$X'X$的逆操作。

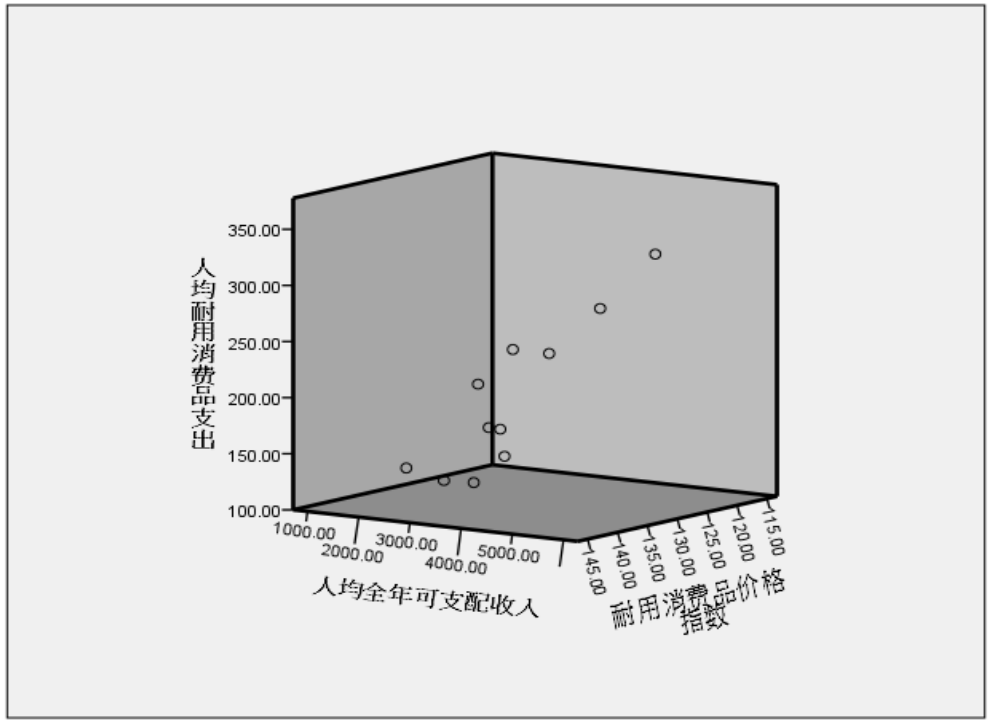
- 拟合优度检验和预测

多元线性回归案例(1)

•下表所示我国1988–1998年的城镇居民人均全年耐用消费品支出、人均全年可支配收入和耐用消费品价格指数的统计资料，试建立城镇居民人均全年耐用消费品支出y关于人均全年可支配收入x_1和耐用消费品价格指数x_2的回归模型。

年 份	人均耐用消费品支出 y	人均全年可支配收入x_1	耐用消费品价格指数x_2
1988	137.16	1181.4	115.96
1989	124.56	1375.7	133.35
1990	107.91	1510.2	128.21
1991	102.96	1700.6	124.85
1992	125.24	2026.6	122.49
1993	162.45	2577.4	129.86
1994	217.43	3496.2	139.52
1995	253.42	4283.0	140.44
1996	251.07	4838.9	139.12
1997	285.85	5160.3	133.35
1998	327.26	5425.1	126.39

多元线性回归案例(2)



$$R^2 = \frac{SS_R}{SS_T} = 0.948$$

$$\bar{R}^2 = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.929$$

估计的回归方程 $\hat{y} = 158.6251 + 0.0494x_1 - 0.9133x_2$

线性回归检验

- 一般使用R^2评价回归模型好坏

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

其中SS_T为总偏差平方，SS_R为回归平方和，SS_E为残差平方和

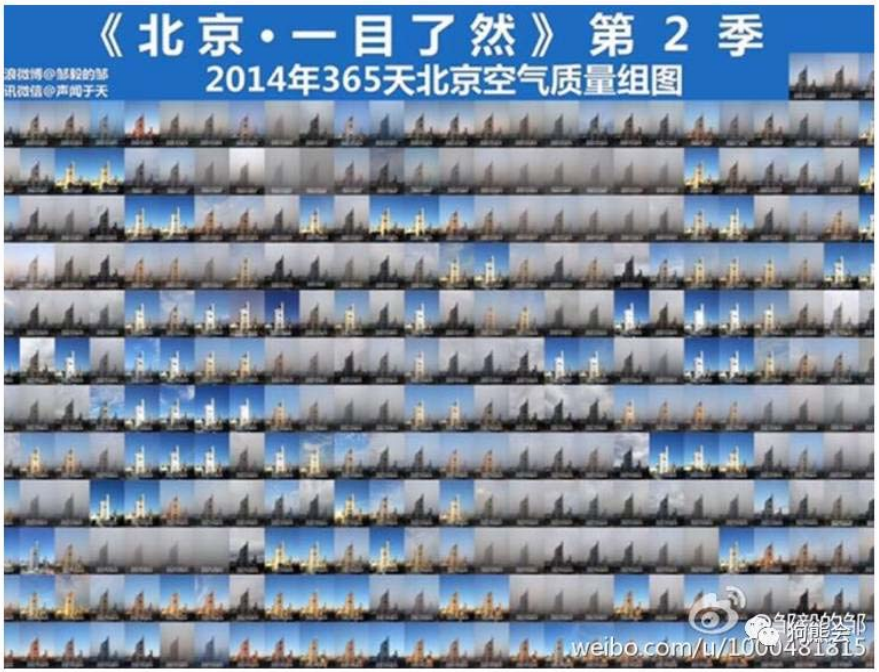
$$SS_T = \sum (y_i - \bar{y})^2 \qquad SS_E = \sum (y_i - \hat{y}_i)^2 \qquad SS_R = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

R^2判定系数度量一个线性回归方程的拟合程度，越接近1拟合程度越好。

- 多元回归的评价指标一般包括
  - 非标准化系数
  - 标准化系数
  - t检验和显著性水平
  - B的置信区间

预测城市PM2.5(1)

通过图片识别PM2.5



预测城市PM2.5(2)

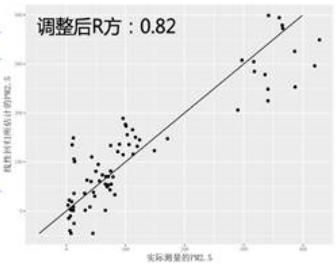
- 从衡量图像清晰程度的角度出发，对图像特征进行观察和分析，得到4个解释性变量：**灰度差分的方差、清晰度、饱和度、高频含量等**

预测城市PM2.5(3)

- 多元线性回归的拟合优度为0.82

线性回归：PM2.5

变量	系数估计值	P值
灰度差分的方差	18.273	0.144
清晰度	-70.447	<0.001 ***
饱和度的均值	-45.231	<0.001 ***
高频含量	-40.969	<0.001 ***



定序回归：污染等级

PM2.5值对应的空气质量等级：中国			
PM2.5值	0-50	50-100	100-150
空气质量等级	一级（优）	二级（良）	三级（轻度污染）
PM2.5值	150-200	200-300	300-500
空气质量等级	四级（中度污染）	五级（重度污染）	六级（严重污染）

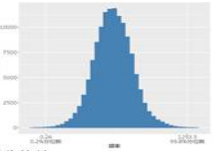
|预测等级-实际等级|	0(完全正确)	1	2	3
百分比 / %	68.1	30.1	0.0	1.4

预测集与训练集的划分——留一交叉验证法：
每次提取1个样本作为预测集，剩下的作为训练集进行对此样本的预测

解释性变量：高频含量

图像视为二维平面信号

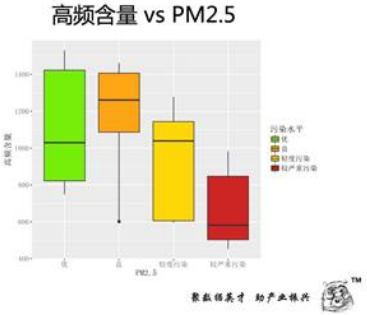
灰度图像进行二维傅里叶变换，得到图片的频率分布



高频含量 = 频率分布99.8%分位数 – 0.2%分位数

天气晴朗，细节丰富，图片高频含量会比较多；
雾霾严重，细节被掩盖，高频含量会显著减少。

解释性变量：高频含量

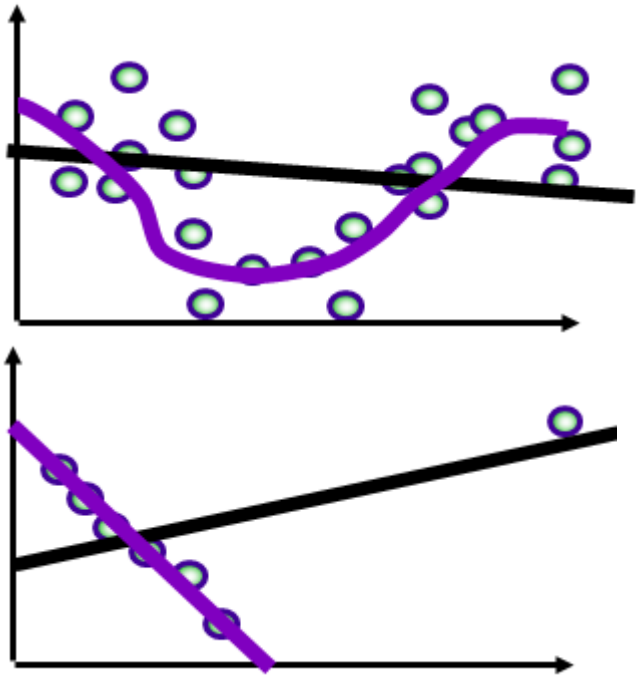


非线性模型

- 在统计学中，非线性回归是回归分析的一种形式， 非线性模型是由一个或多个自变量非线性组合
- 一些常见非线性模型
  - 阶跃函数
  - 分段函数
  - 样条曲线
  - 广义加性模型

非线性回归

- 事实上，现实中的大多数问题是非线性的，需要对变量进行变换，把非线性问题转换为线性问题解决。在线性回归问题中，变量一般是独立的。但很多情况下，高次多项式可以更好地反映变量之间的关系，这就需要引入非线性回归预测未知变量。
- 非线性回归的思想是通过对变量进行转换，把非线性模型转换为线性模型，然后按上述方法再求解线性模型求出其中参数后带入原非线性模型。例如对多项式回归$y=c_0+c_1x+c_2x^2+c_3x^3+c_4x^4$，先把此方程转换成线性方程，需要定义如下几个新变量：$x_1=x$，$x_2=x^2$，$x_3=x^3$，$x_4=x^4$。代入原先的多项式方程，得到$y=c_0+c_1x_1+c_2x_2+c_3x_3+c_4x_4$，多项式回归问题就转化为一个多元线性回归问题。
- 对于双曲线函数 $y=\frac{x}{ax+b}$，进行线性转换$y_1=1/y$，$x_1=1/x$，则有$y_1=a+b\,x_1$。
- 对于指数函数等比较复杂的非线性函数，需要通过更复杂的转换例如对 $y=\alpha x^{\beta}$ 可以做如下变换：$\ln y=\ln\alpha+\beta\ln x$，定义$y_1=\ln y, x_1=\ln x$，得到 $y_1=\ln\alpha+\beta x_1$。



耐热导线工厂质量管理数据分析

耐热导线工厂质量管理数据分析需求

（1）优化《杆材流转使用规定》

①通过数据分析缩短现有拉制铝单线、耐热铝合金单线选杆范围、拉制铝镁硅合金单线选杆范围，提高单丝、成品一次合格率。

②进一步论证单线线径与所选用杆强度之间的关系，即论证是否有必要按照单线线径的范围来选择相应的杆强度范围。

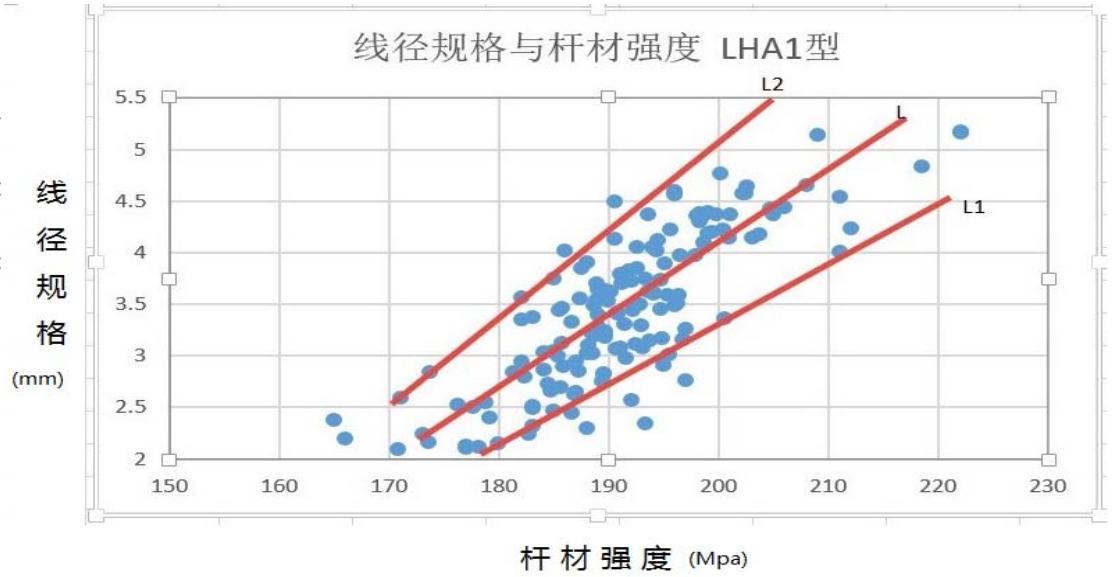
③建立单线强度与杆强度之间的对应关系或数学模型，例如单线强度在标准基础上提高10Mpa，则对应的杆强度需要提高多少比较合适。

（2）优化合金杆静置时间

通过合金杆下机测试、合金冷测、对应单丝检测日期等数据，分析合金杆材静置时间与单丝合格率影响，最佳静置时间带来的单丝最大的合格率。

（3）设备故障率分析

提供近两年设备运行情况明细以及汇总数据，分析每台机器的机台特性，确定机台的停机调整时机。



Logistics回归

- Logistic回归建立了一个多项式对数回归模型，用于预测二值变量的值(0或1)。相对于独立变量 $x_1, x_2, \ldots, x_n$ ，变量*y*等于1的概率定义如下：

$$p(y=1 \mid x_1, x_2, \ldots, x_n) = \frac{e^{-(a_1x_1+a_2x_2+\ldots+a_nx_n+\mu)}}{1+e^{-(a_1x_1+a_2x_2+\ldots+a_nx_n+\mu)}}$$

- Logistic回归在数据挖掘中很有用，特别是解决两类的数据概率打分问题，如顾客流失风险打分等。

Logistics回归

- 逻辑回归是一种预测分析， 解释因变量与一个或多个自变量之间的关与线性回归不同之处就是它的目标变量有几种类别，所以逻辑回归主要用于解决分类问题，与线性回归相比，它是用概率的方式，预测出来属于某一分类的概率值。如果超过50%，则属于某一分类。此外，它的可解释强，可控性高，并且训练速度快，特别是经过特征工程之后效果更好
- 按照逻辑回归的基本原理，求解过程可以分为以下三步
  - 找一个合适的预测分类函数，用来预测输入数据的分类结果，一 般表示为h函数，需要对数据有一定的了解或分析，然后确定函数的可能形式
  - 构造一个损失函数，该函数表示预测输出（h） 与训练数据类别（y）之间的偏差，一般是预测输出与实际类别的差，可对所有样本的Cost求R方值等作为评价标准， 记为$J(\theta)$函数
  - 找到$J(\theta)$函数的最小值，因为值越小表示预测函数越准确。求解损失函数的最小值是采用梯度下降法实现

应用logistic回归模型预测银行顾客是否拖欠贷款

- 根据历史数据识别银行拖欠顾客的特征，预测潜在信贷顾客是否拖欠贷款。这里选取700个信贷顾客的历史记录，其中21.5%是拖欠顾客。这里选择顾客性别（sex）、收入（income）、年龄（age)、edcation(文化程度)，employ(现单位工作年数)，debtinc(负债率)和creddebt（信用卡债务）等作为自变量，顾客拖欠贷款与否作为因变量：1代表拖欠，0代表正常。选择70%历史记录进行训练，剩下30%历史数据用于验证，建立一个预测因变量取1的概率的logistic回归模型，以对新的潜在顾客是否拖欠贷款进行预测。
- 影响顾客拖欠的自变量比较多，这里采用Forward/Backward方式用于剔除不重要的自变量，例如收入水平、文化程度和年龄等对顾客信用的影响不显著，拖欠概率的回归方程如下：

$$\ln\frac{p}{1-p} = -0.76 - 0.249employ - 0.069address + 0.08debtinc + 0.594creddebt$$

- 对模型进行显著性检验以及回归模型与样本数据的拟合程度以及模型预测精度进行评价，回归模型满足一定要求即可部署使用。从中可以发现拖欠贷款客户的特征：工作不稳定、住址经常变动、债务比率高、信用卡债务多的客户，拖欠贷款的概率较大。