上海交通大学
SHANGHAI JIAO TONG UNIVERSITY

机器学习内容分享

上海交通大学 殷奕珉

SJTU
饮 水 思 源 · 爱 国 荣 校

目 录

01

一元和多元线性回归理论

“回归”的由来?

- FrancisGalton，英国生物学家，他研究了父母身高与子女身高之间关系后得出，若父母身高高于平均大众身高，则其子女身高倾向于倒退生长，即会比其父母身高矮一些而更接近于大众平均身高。若父母身高小于平均身高，则其子女身高倾向于向上生长，以更接近于大众平均身高。此现象，被Galton称之为回归现象，即regression.

我们为什么使用回归分析?

- 回归分析估计了两个或多个变量之间的关系。
- 比如说，在当前的经济条件下，你要估计一家公司的销售额增长情况。现在，你有公司最新的数据，这些数据显示出销售额增长大约是经济增长的2.5倍。那么使用回归分析，我们就可以根据当前和过去的信息来预测未来公司的销售情况。
- 使用回归分析的好处良多。具体如下：
- 表明自变量和因变量之间的显著关系；多个自变量对一个因变量的影响强度。
- 回归分析也允许我们去比较那些衡量不同尺度的变量之间的相互影响，如价格变动与促销活动数量之间联系。这些有利于帮助市场研究人员，数据分析人员以及数据科学家排除并估计出一组最佳的变量，用来构建预测模型

什么是线性回归?

- 回归分析是一种统计工具，它利用两个或两个以上变量之间的关系，由一个或几个变量来预测另一个变量。
  - 自变量只有一个时，叫做一元线性回归

$$h(x) = b_0 + b_1 x$$

  - 自变量有多个时，叫做多元线性回归

$$h(x_1, x_2, \ldots, x_p) = b_0 + b_1 x_1 + \ldots + b_p x_p$$

一元线性回归

- 1、一元线性回归模型：$y = \beta_0 + \beta_1 x + \varepsilon$

  （1）模型的特点y 是 x 的线性函数(部分)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,是不能由 x 和 y 之间的线性关系所解释的变异性

  （4）β_0 和 β_1 称为模型的参数

- 一元线性回归模型的假定

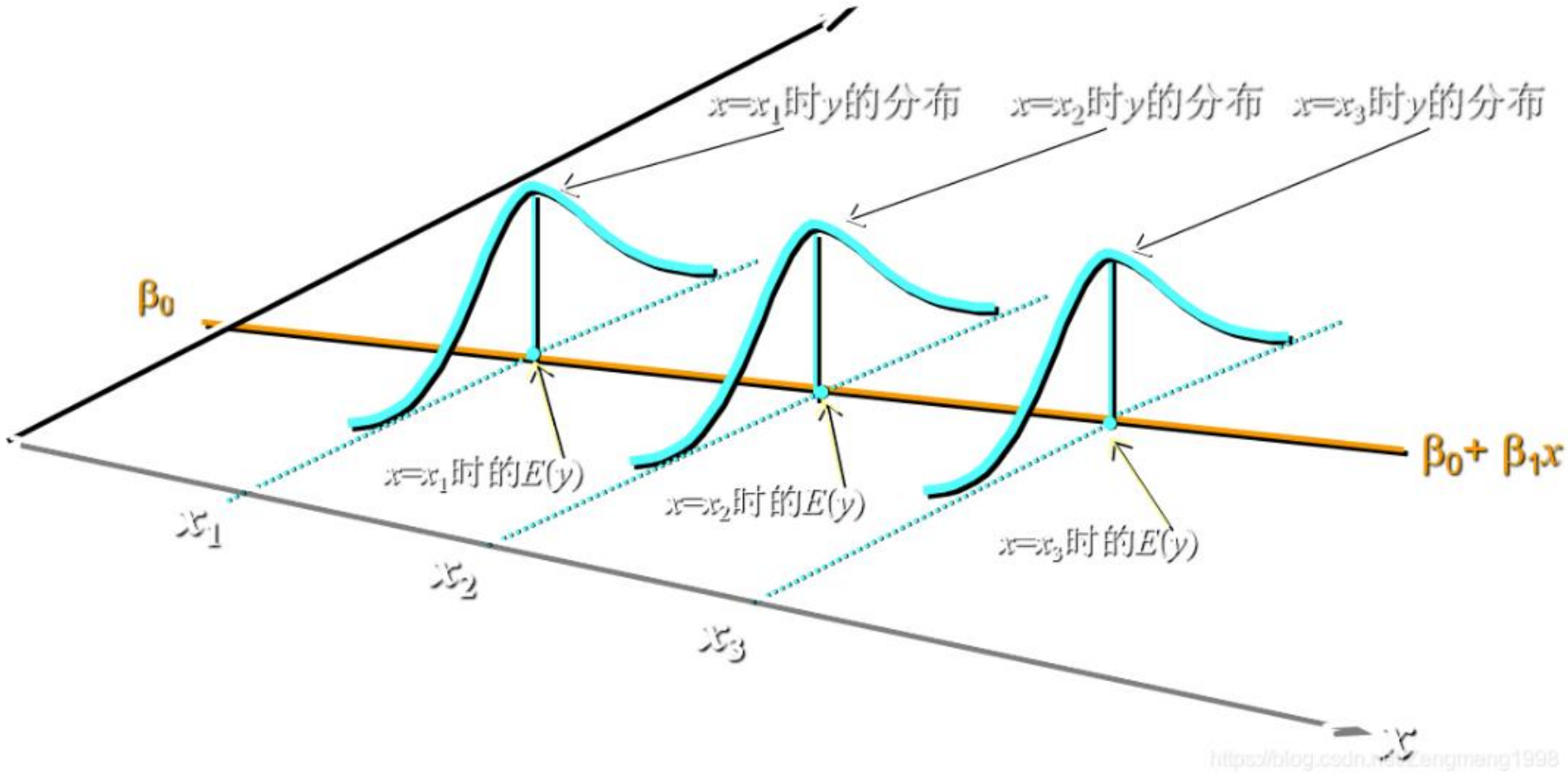
  （1）因变量x与自变量y之间具有线性关系；

  （2）在重复抽样中，自变量x的取值是固定的，即假定x是非随机的

  （3）误差项ε是一个期望值为0的随机变量，既有：$E(y) = \beta_0 + \beta_1 x$

  （4）误差项ε是一个服从正态分布的随机变量，且相互独立。即ε~N(0 ,σ2 ）（关于随机误差的理解）

一元线性回归

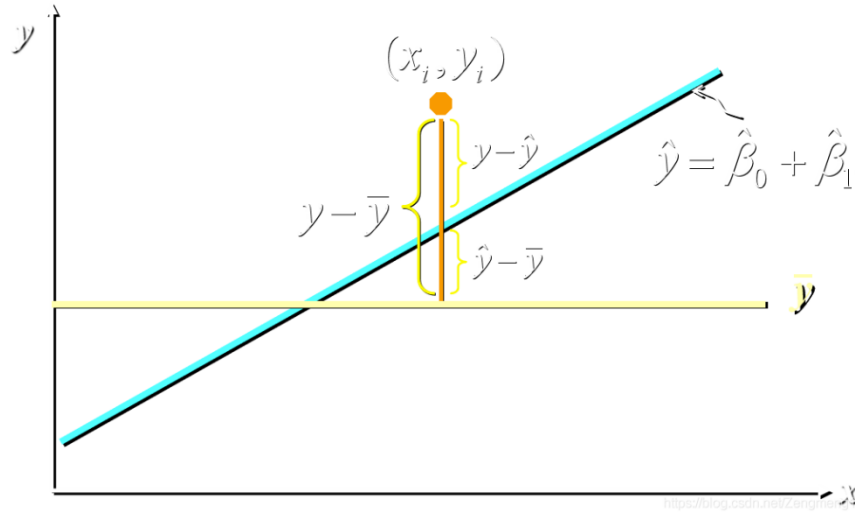

x=x1时y的分布
x=x2时y的分布
x=x3时y的分布
β0
β0+ β1x
x=x1时的E(y)
x=x2时的E(y)
x=x3时的E(y)
x1
x2
x3
x

一元线性回归

- 2、一元线性回归方程：$E(y) = \beta_0 + \beta_1 x$
  （1）β_0是回归直线在 y 轴上的截距，是当 x=0 时 y 的期望值
  （2）β_1是直线的斜率，称为回归系数，表示当 x 每变动一个单位时，y 的平均变动值
- 3、估计的回归方程
  （1）作用：用样本统计量$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 代替回归方程中的未知参数 β_0 和 β_1 就得到了估计的回归方程
  （2）估计的回归方程：$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- 用最小二乘法估计回归方程的参数：
  （1）本质：使得$\sum_{i=1}^{n}(y_i - \hat{y})^2 = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = \min$，求法是分别对$\hat{\beta}_0$ 和$\hat{\beta}_1$求偏导；
  （2）求解公式：求偏导
  $$\begin{cases} \left.\frac{\partial Q}{\partial \beta_0}\right|_{\beta_0=\hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = 0 \\ \left.\frac{\partial Q}{\partial \beta_1}\right|_{\beta_1-\hat{\beta}_1} = -2\sum_{i=1}^{n} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 = 0 \end{cases}$$
  解得：$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$　　$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

一元线性回归：回归直线的拟合优度检验

- 误差分解;



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

（1）总平方和SST=$\sum_{i=1}^{n}(y_i-\bar{y})^2$;反映因变量的 n 个观察值与其均值的总误差

（2）回归平方和SSR=$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$;反映自变量 x 的变化对因变量 y 取值变化的影响，或者说，是由于 x 与 y 之间的线性关系引起的 y 的取值变化，也称为可解释的平方和

（3）残差平方和SSE=$\sum_{i=1}^{n}(y_i-\hat{y})^2$。反映除 x 以外的其他因素对 y 取值的影响，也称为不可解释的平方和或剩余平方和

- 判定系数R^2的计算

（1）计算公式：$R^2=\frac{SSR}{SST}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}$;

（2）反映回归直线的拟合程度;

（3）取值范围在 [0 , 1] 之间;

（4）判定系数等于相关系数的平方，即$R^2=r^2$

- 标准估计误差的计算:

（1）计算公式：$s_e=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}=\sqrt{\frac{SSE}{n-2}}=\sqrt{MSE}$;

（2）实际观察值与回归估计值误差平方和的均方根;

（3）对误差项ε的标准差σ2 的估计，是在排除了x对y的线性影响后，y随机波动大小的一个估计量。

一元线性回归：显著性检验

- 线性关系检验

  （1）检验自变量与因变量之间的线性关系是否显著

  （2）计算：将回归均方(MSR)同残差均方(MSE)加以比较，应用F检验来分析二者之间的差别是否显著，回归平方和SSR除以相应的自由度(自变量的个数k)，残差平方和SSE除以相应的自由度(n-k-1)。计算公式为：$\mathrm{F} = \frac{\mathrm{SSR}/1}{\mathrm{SSE}/(\mathrm{n}-2)} = \frac{\mathrm{MSR}}{\mathrm{MSE}} \sim \mathrm{F}(1, \mathrm{n}-2)$

- 回归系数检验

  （1）目的：检验 x 与 y 之间是否具有线性关系，或者说，检验自变量 x 对因变量 y 的影响是否显著；

  （2）理论基础是回归系数$\hat{\beta}_1$ 的抽样分布

  （3）t检验统计量计算公式：$\mathrm{t} = \frac{\hat{\beta}_1}{s_{\hat{\beta}}} \sim \mathrm{t}(\mathrm{n}-2)$（等价于相关系数的显著性检验）

一元线性回归：残差

*残差

（1）因变量的观测值与根据估计的回归方程求出的预测值之差，用e表示公式为：$e_i = y_i - \hat{y}_i$

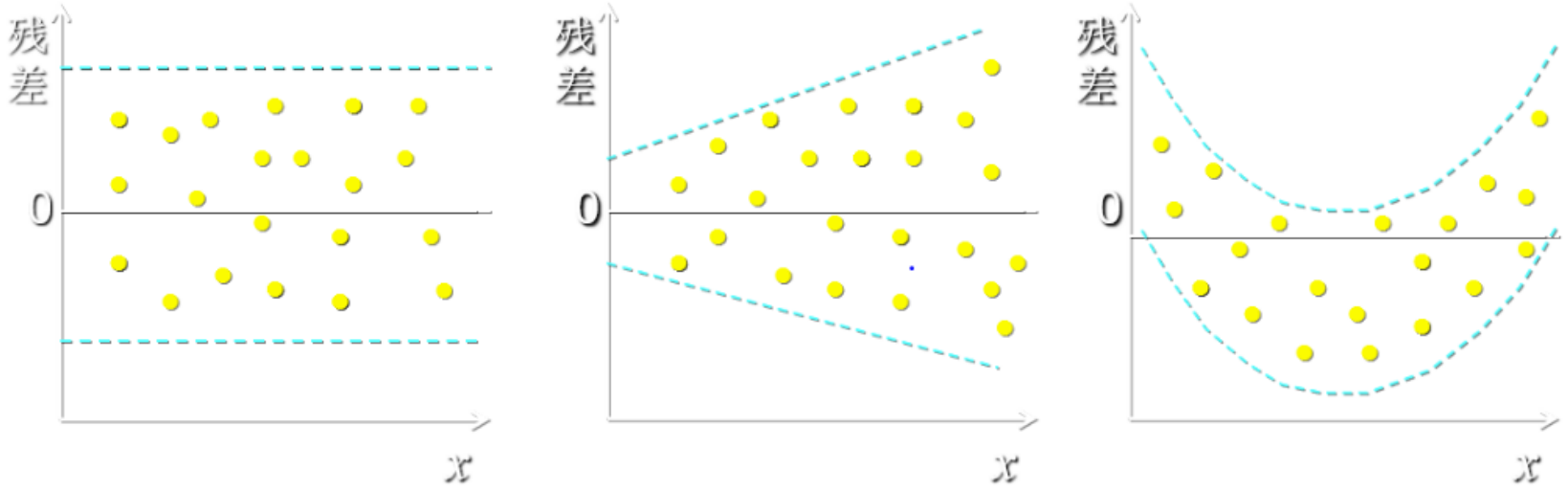
（2）反映了用估计的回归方程去预测而引起的误差；

（3）作用：可用于确定有关误差项的假定是否成立 。

- 残差图像

  （1）好坏判别：判断误差项ε是否符合假定（均值为零的正态分布）

  （2）一般图像：



(a)满意模式　　(b)非常数方差　　(c)模型不合适

一元线性回归：标准化残差

- 标准化残差的计算：（残差除以它的标准差）：$z_{e_i} = \frac{e_i}{s_e} = \frac{y_i - \hat{y}_i}{s_e}$ 即：$z_i = \frac{y_i - \hat{y}_i}{s_e \sqrt{1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}\right)}}$

  （1）作用：用以直观地判断误差项服从正态分布这一假定是否成立；

  （2）若假定成立，标准化残差的分布也应服从正态分布；

  （3）在标准化残差图中，大约有95%的标准化残差在-2到+2之间。

多元线性回归

多元线性回归模型通常用来描述变量y和x之间的随机线性关系，即：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \xi$$

式中，$x_1, \cdots, x_k$ 是非随机的变量；y 是随机的因变量；$\beta_0, \cdots, \beta_k$ 是回归系数；ξ 是随机误差项。

如果对y和x进行了x次观测，得到n组观察值yi,x1i,…,xki(i=1,2,…,n),他们满足一下关系式：

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + \xi$$

式中，$x_1, \cdots, x_k$ 是非随机的变量；y 是随机的因变量；$\beta_0, \cdots, \beta_k$ 是回归系数；ξ 是随机误差项。

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \xi_i$$

多元线性回归

用矩阵表示：

$$y=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad x=\begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \quad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \xi=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

这时模型可以写作：

$$y=X\beta+\xi$$

如果模型满足条件（1）$E_\varepsilon=0$ （2）$\mathrm{Var}(\varepsilon)=\sigma^2 I$ （3）$x_1,\ldots,x_k$互不相关，则称模型为普通线性模型。如果模型的随机误差服从正态分布，即$\varepsilon\sim N(0,\sigma^2 I)$，则称模型为普通正太线性回归模型。

多元线性回归

1.2模型参数的检验

在正态假定下，如果X是列满秩的，则普通线性回归模型的参数最小二乘估计为：

$$\hat{\beta} = (x^T x)^{-1} x^T y$$

于是y的估计值为：

$$\hat{y} = x\hat{\beta}$$

记残差向量为 $e = y - \hat{y} = y - X\hat{\beta}$，则随机误差方差 σ^2 的最小二乘估计为：

$$\hat{\sigma}^2 = \frac{e^T e}{n-k-1}$$

得到回归模型参数的估计值后，需要对回归方程和回归系数进行显著性检验

多元线性回归

（1）回归方程的显著性检验

原假设$H_0: \beta_1 = \cdots = \beta_k = 0$,备择假设$H_1: \beta_1, \ldots, \beta_k$不全为 0，当假设成立时，检验统计量

$$F = \frac{SSR / k}{SSE / (n-k-1)} \sim F(k, n-k-1)$$

式中，$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ 是回归平方和；$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ 是残差平方和。对于给定的

显著性水平α，检验的拒绝域$F > F_\alpha(k, n-k-1)$。

多元线性回归

(2) 回归系数的显著性检验

原假设$H_0: \beta_j = 0$,备择假设$H_1: \beta_j \neq 0 (j = 0,1,\ldots,k)$，当原假设成立时，检验统计量

$$F_j = \frac{SSE_j - SSE}{SSE/(n-k-1)} \sim F(1, n-k-1)$$

式中，SSE_j是去掉x_j后的残差平方和。对于给定的显著水平α，检验的拒绝域为

$F_j > F_\alpha(1, n-k-1)$。

也可以用检验统计量

$$t_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{c_{jj}}} \sim t(n-k-1)$$

式中，c_{jj}是$c = (x^T x)^{-1}$对角线上第 j(j=0,1,…,k)个元素。对于给定的显著性水平α，检验的拒绝域为$|t_j| > t_{\alpha/2}(n-k-1)$。

线性回归的python实现

- 1. Simple Linear Regression and Hypothesis Testing
- 2. Multiple Linear Regression

回归算法

- Linear Regression线性回归
- Logistic Regression逻辑回归
- Polynomial Regression多项式回归
- Stepwise Regression逐步回归
- Ridge Regression岭回归
- Lasso Regression套索回归

参考网址:

https://cloud.tencent.com/developer/article/1102103

https://zhuanlan.zhihu.com/p/483694937

分类算法

- 真假新闻识别
- 水果分类

常用分类算法

- Bayes 朴素贝叶斯
- Decision Tree 决策树
- SVM 支持向量机
- KNN K邻近
- Logistic Regression 逻辑回归
- 神经网络
- Ensemble Learning

参考网站

- https://blog.csdn.net/china1000/article/details/48597469
- https://zhuanlan.zhihu.com/p/82114104

LR算法详解

- 见文件：机器学习-LR算法-yym

THANK YOU