11 | 基础线性回归：一元与多元

2018-06-28 王天一

机器学习40讲 进入课程 >

讲述：王天一
时长 23:34 大小 9.51M

从今天开始，专栏将进入统计机器学习模块。虽然统计机器学习中千姿百态的模型让人眼花缭乱，但究其本原，它们都来源于最原始的**线性回归**（linear regression）。

在我看来，**线性模型最大的优点不是便于计算，而是便于解释。**它能以简洁明了的方式清晰体现出输入的变化如何导致输出的变化。正所谓“一生二，二生三，三生万物”，将不同的改进方式融入线性模型的基本思想中，就可以得到各种巧夺天工的复杂方法。

在第一季“人工智能基础课”专栏中，我介绍了线性回归的原理，证明了当噪声满足正态分布时，基于最小二乘法（least squares）的线性回归和最大似然估计是等价的。

《机器学习 | 简约而不简单：线性回归》

这次我们换个角度，来看看**最小二乘法的几何意义**。之前，线性回归的数学表达式被写成 $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = \sum_{i=0}^{n} w_i \cdot x_i$。但在讨论几何意义时，这个表达式要被改写成

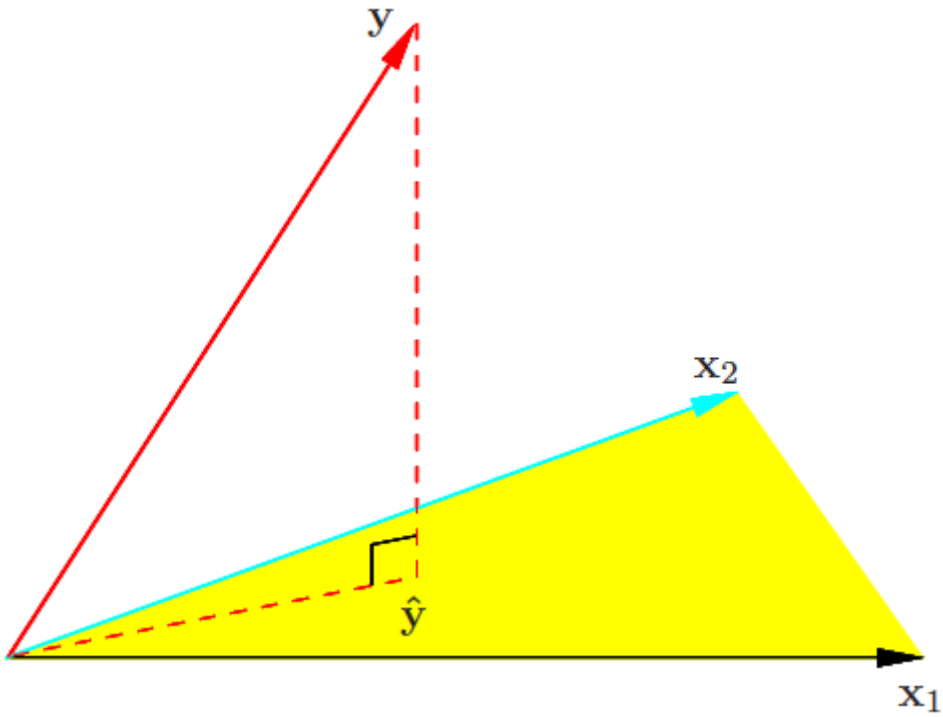
$$f(\mathbf{x}) = 1 \cdot \beta_0 + \sum_{j=1}^{n} x_j \cdot \beta_j = \mathbf{x}^T\boldsymbol{\beta}$$

可别小看这个简单的写法变化，从列向量 $\mathbf{x}$ 到行向量 $\mathbf{x}^T$ 的改变就像矩阵的左乘和右乘一样具有不同的意义。

当输出被写成 $\mathbf{w}^T\mathbf{x}$ 时，其背后的寓意是每个包含若干输入属性和一个输出结果的样本都被视为一个整体，误差分散在不同的样本点上；而当输出被写成 $\mathbf{x}^T\boldsymbol{\beta}$ 时，其背后的寓意是**每个单独属性在所有样本点上的取值被视为一个整体，误差分散在每个不同的属性上**。但横看成岭侧成峰，观察角度的变化不会给观察对象本身造成改变，数据本身是没有变化的。

假设数据集中共有 N 个样本，那么 $\mathbf{x}^T$ 就变成了 $N \times (n+1)$ 维的数据矩阵 $\mathbf{X}$，它的每一行表示的都是同一个样本的不同属性，每一列则表示不同样本中的相同属性。如果待拟合数据的特性完美到任意两个属性都线性无关的话，$\mathbf{X}$ 就可以看成一个由它的所有列向量所张成的空间。

一般来说，属性的数目 n 会远远小于数据的数目 N，因此 $\mathbf{X}$ 张成的是 N 维空间之内的**n维生成子空间**，或者叫**n 维超平面**。这个超平面的每一个维度都对应着数据集的一个列向量。理想条件下，输出 $\mathbf{y}$ 作为属性的线性组合，也应该出现在由数据属性构成的超平面上。但受噪声的影响，真正的 $\mathbf{y}$ 是超平面之外的一个点，这时就要退而求其次，在超平面上找到离 $\mathbf{y}$ 最近的点作为最佳的近似。



最小二乘的几何意义（图片来自 Elements of Statistical Learning，图 3.2）

在上图中，黄色区域表示由所有属性张成的超平面；黑色向量 x_1 和天蓝色向量 x_2 表示输入属性；红色实线 y 表示真实输出，水平的红色虚线 $\hat{y}$ 表示数据的最优估计值（属性的线性组合）；垂直的红色虚线表示 y 与 $\hat{y}$ 的残差，它与超平面正交。

根据几何知识不难得出，要找的最佳近似 $\hat{\mathbf{y}}$ 就是 $\mathbf{y}$ 在超平面上的投影，而最佳近似所对应的系数 $\hat{\boldsymbol{\beta}}$ 就是线性回归的解，点 $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$ 和 $\mathbf{y}$ 之间的距离就是估计误差，也叫**残差**（residual），它就是最小二乘法最小化的对象，其表达式是 $||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2$。对参数 $\boldsymbol{\beta}$ 求导不难得到，能够使均方误差最小化的参数 $\hat{\boldsymbol{\beta}}$ 应该满足

$$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\mathbf{X} = 0$$

这个式子说明了最小二乘法的几何意义：**计算高维空间上的输出结果在由所有属性共同定义的低维空间上的正交投影**（orthogonal projection）。投影操作意味着残差不会在数据维度上遗留任何分量，这种基于误差和数据正交性的最优解也经常出现在信号处理当中。

在实际应用中，如何解释线性回归的结果呢？下面这个例子可以说明。

眼下世界杯正进行得如火如荼。如果爱好足球，你一定不会对数据网站 WhoScored 感到陌生，它的一大特色是会在每场比赛结束后根据球员表现给出评分，0 分最低，10 分最

高。虽然这个评分系统能够直观体现谁踢得好谁踢得差，但关于其专业性的质疑却从未停止。那么 WhoScored 的评分到底准不准呢？我们不妨用线性回归做个测试。

如果 WhoScored 的评分足够合理，那球员的评分就应该和球队的成绩呈现出正相关，而线性又是正相关最直接的描述。为了验证球队赛季积分 y 和所有球员的赛季平均评分 x 之间是否存在线性关系，我从网站上复制了 2017~18 赛季英超联赛的相关数据，这个包含 20 个样本的小数据集就是训练集。在拟合数据时，我使用的第三方库是 StatsModels，之所以选择这个库是因为它能够给出更多统计意义上的结论，这些结论对于理解线性回归模型大有裨益。

R	Team	Pts
1	Manchester City	100
2	Manchester United	81
3	Tottenham	77
4	Liverpool	75
5	Chelsea	70
6	Arsenal	63
7	Burnley	54
8	Everton	49
9	Leicester	47
10	Newcastle United	44
11	Crystal Palace	44
12	Bournemouth	44
13	West Ham	42
14	Watford	41
15	Brighton	40
16	Huddersfield	37
17	Southampton	36
18	Swansea	33
19	Stoke	33
20	West Bromwich Albion	31

R	Team	Rating
1	Manchester City	7.15
2	Liverpool	6.99
3	Manchester United	6.98
4	Tottenham	6.95
5	Chelsea	6.94
6	Arsenal	6.92
7	Crystal Palace	6.84
8	Burnley	6.78
9	Leicester	6.74
10	Newcastle United	6.72
11	West Ham	6.72
12	Stoke	6.70
13	Brighton	6.69
14	Southampton	6.67
15	Everton	6.67
16	Watford	6.66
17	Bournemouth	6.65
18	West Bromwich Albion	6.63
19	Huddersfield	6.61
20	Swansea	6.59

2017~18 赛季英超联赛积分榜与评分榜（图片来自 whoscored.com）

在模型拟合之前有必要对输入数据做一点处理，那就是将因变量从球队的赛季总积分转换成场均积分。在足球联赛中，一场比赛的胜 / 平 / 负分别对应 3/1/0 分，因此计算场均积分可以看成是某种意义上的归一化，使数据在 [0, 3] 这个一致的较小尺度上得到更加直观的比较。

在使用 StatsModels 拟合模型时，首先要用 add_constant 函数在每个输入数据的后面添加一个 1，借此把常数项纳入模型之中；接下来就可以调用 OLS，也就是普通最小二乘法（ordinary least squares）作为拟合对象，计算线性模型的参数；最后使用 fit 函数获取拟合结果。要查看拟合模型的统计特性，只需打印出模型的 summary。下图就是对英超数据集的拟合结果。

OLS Regression Results

Dep. Variable:	Points	R-squared:	0.905
Model:	OLS	Adj. R-squared:	0.900
Method:	Least Squares	F-statistic:	172.2
Date:	Thu, 21 Jun 2018	Prob (F-statistic):	1.18e-10
Time:	15:38:52	Log-Likelihood:	9.3960
No. Observations:	20	AIC:	-14.79
Df Residuals:	18	BIC:	-12.80
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-19.4345	1.586	-12.256	0.000	-22.766	-16.103
Ratings	3.0685	0.234	13.123	0.000	2.577	3.560

Omnibus:	3.586	Durbin-Watson:	2.238
Prob(Omnibus):	0.166	Jarque-Bera (JB):	1.886
Skew:	-0.713	Prob(JB):	0.389
Kurtosis:	3.481	Cond. No.	308.

英超数据集上的简单线性回归拟合结果

可以看到，模型拟合最核心的结果显示在第二排：coef 表示的是参数的**估计值**，也就是通过最小二乘计算出的权重系数。拟合结果告诉我们，球队场均积分 y 和球员平均评分 x 之间的关系可以近似表示为回归式 $y = 3.0685x - 19.4345$，这说明如果所有球员共同努力将平均评分拉高 0.1 的话，球队在每场比赛中就能平均多得 0.306 分。

右侧 std err 表示的是参数估计的**标准误**（standard error），虽然最小二乘得到的是无偏估计量，意味着估计结果中不存在系统误差，但每一个特定的估计值结果依然会在真实值的附近波动，标准误度量的就是估计值偏离真实值的平均程度。

最后两列 [0.025 0.975] 给出了 95% 置信区间：每个参数真实值落在这个区间内的可能性是 95%。对于线性回归而言，置信下界和上界分别是估计值减去和加上二倍的标准误，也就是 coef$\pm 2\times$std err。

置信区间告诉我们，平均评分拉高 0.1 并不意味着球队每场一定能多得 0.306 分，但多得的分数基本在 0.258 到 0.356 之间。如果用 2016-17 赛季的数据作为训练数据的话，这个

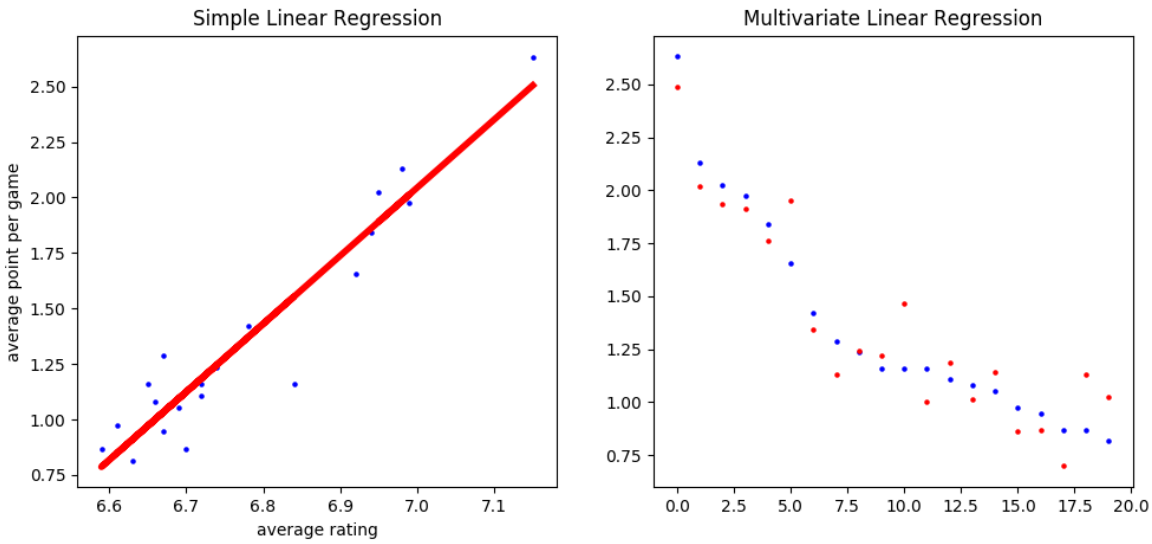
数据的计算结果就变成了 0.33——也落在置信区间之内，这也验证的估计结果的波动性。

中间两列中的 t 和 P>|t|都是统计学中的关键指标，它们评估的是拟合结果的统计学意义。t 代表 t 统计量（t-statistic），表示了参数的估计值和原始假设值之间的偏离程度。在线性回归中通常会假设待拟合的参数值为 0，此时的 t 统计量就等于估计值除以标准误。当数据中的噪声满足正态分布时，t 统计量就满足 t 分布，其绝对值越大意味着参数等于 0 的可能性越小，拟合的结果也就越可信。

P>|t|表示的则是统计学中争议最大的指标——p **值**。p 值（p-value）是在当原假设为真时，数据等于观测值或比观测值更为极端的概率。简单地说，p 值表示的是数据与一个给定模型不匹配的程度，p 值越小，说明数据和原假设的模型越不匹配，也就和计算出的模型越匹配。在这个例子里，原假设认为待估计的参数等于 0，而接近于 0 的 p 值就意味着计算出的参数值得信任。

看完第二排再来看第一排，也就是对模型拟合数据的程度的评价，重要的指标在右侧一列。R-squared 表示的是 R^2 统计量，也叫作**决定系数**（coefficient of determination），这个取值在 [0, 1] 之间的数量表示的是输出的变化中能被输入的变化所解释的部分所占的比例。在这个例子里，$R^2 = 0.905$ 意味着回归模型能够通过 x 的变化解释大约 91% 的 y 的变化，这表明回归模型具有良好的准确性，回归后依然不能解释的 9% 就来源于噪声。

R^2 统计量具有单调递增的特性，即使在模型中再添加一些和输出无关的属性，计算出来的 R^2 也不会下降。Adj. R-squared 就是校正版的 R^2 统计量。当模型中增加的变量没有统计学意义时，多余的不相关属性会使校正决定系数下降。校正决定系数体现出的是正则化的思想，它在数值上小于未校正的 R^2 统计量。



英超数据集上简单线性回归（左）和多元线性回归（右）的拟合结果

上图给出了英超数据集上简单线性回归和多元线性回归的拟合结果，其中蓝点为数据点，红点为预测点。在简单回归中，大部分数据点集中在拟合直线附近，一个明显的异常点是中游球队水晶宫（Crystal Palace）。

回到英超数据集的例子，图形结果和数值指标都表明线性回归能够较好地拟合两者之间的关系，这说明 WhoScored 的评分系统是值得信任的。但这个例子只是线性回归的一个特例，它特殊在输出的因变量只与单个的输入自变量存在线性关系，这种模型被称为**简单线性回归**（simple linear regression）。更一般的情况是因变量由多个自变量共同决定，对这些自变量同时建模就是**多元线性回归**（multivariate linear regression）。

与简单线性回归一样，多元线性回归中的参数也要用最小二乘法来估计。还是以积分和评分的关系为例，在简单线性回归中，自变量是所有球员在所有比赛中评分的均值，但是球场上不同位置的球员发挥的作用也不一样。为了进一步分析不同位置球员对球队表现的影响，就要将单个自变量替换成不同位置球员（门将 / 后卫 / 中场 / 前锋）在整个赛季中的平均评分，再使用多元回归进行拟合。

在这个实例中，多元回归的属性，也就是自变量被设置为每队每个位置上出场时间较多的球员的赛季平均评分的均值，所有选中球员的出场时间都在 1000 分钟以上。

利用 OLS 模型可以得到多元回归的结果，可如果对结果加以分析，就会发现一个有趣的现象：一方面，多元模型的校正决定系数是 0.876，意味着所有位置评分共同解释了输出结果

的大部分变化，这也可以从预测值与真实值的散点图上观察出来；可另一方面，只有后卫评分和前锋评分的 p 值低于 0.05，似乎球队的战绩只取决于这两个位置的表现。

OLS Regression Results

Dep. Variable:	Points	R-squared:	0.902
Model:	OLS	Adj. R-squared:	0.876
Method:	Least Squares	F-statistic:	34.57
Date:	Thu, 21 Jun 2018	Prob (F-statistic):	2.09e-07
Time:	15:38:53	Log-Likelihood:	9.0610
No. Observations:	20	AIC:	-8.122
Df Residuals:	15	BIC:	-3.143
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-15.5306	2.366	-6.563	0.000	-20.574	-10.487
GK	-0.2112	0.454	-0.466	0.648	-1.178	0.756
DF	1.5722	0.545	2.885	0.011	0.411	2.734
MF	0.5510	0.354	1.556	0.141	-0.204	1.306
FW	0.5337	0.237	2.249	0.040	0.028	1.040

Omnibus:	3.397	Durbin-Watson:	1.989
Prob(Omnibus):	0.183	Jarque-Bera (JB):	2.804
Skew:	-0.861	Prob(JB):	0.246
Kurtosis:	2.369	Cond. No.	823.

英超数据集上的多元线性回归拟合结果

看起来校正决定系数和 p 值给出了自相矛盾的解释，这时就需要观察另外一个重要的指标：F **统计量**。

F 统计量（F-statistic）主要应用在多元回归中，它检验的原假设是所有待估计的参数都等于 0，这意味着只要有一个参数不等于 0，原假设就被推翻。F 统计量越大意味着原假设成立的概率越低，理想的 F 值应该在百千量级。可在上面的多元回归中，F 统计量仅为 34.57，这就支持了 p 值的结论：估计出的参数的统计学意义并不明显。

英超数据集在统计上的非显著性可能源自过小的样本数导致的过拟合，也可能源自不同属性之间的共线性（collinearity）。可在更广泛的意义上，它揭示的却是多元线性回归无法回避的一个本质问题：**模型虽然具有足够的精确性，却缺乏关于精确性的合理解释。**

假定数据共有 10 个属性，如果只保留 10 个属性中的 5 个用于拟合的话，肯定会有不止一个 5 元属性组能够得到彼此接近的优良性能，可对不同 5 元组的解读方式却会大相径庭。这种现象，就是统计学家莱奥·布雷曼口中的“罗生门”（Rashomon）。

《罗生门》是日本导演黑泽明的作品，取材于日本作家芥川龙之介的小说《草莽中》。一名武士在竹林中被杀，不同当事人的供词既是不同程度上的事实，也是不同角度下的谎言。布雷曼用这个词来描述最优模型的多重性，以及由此造成的统计建模的艰难处境：当不同的多元线性模型性能相近，却公说公有理婆说婆有理时，到底应该如何选择？

将“罗生门”深挖一步，就是机器学习和统计学在认识论上的差异：统计学讲究的是“知其然，知其所以然”，它不仅要找出数据之间的关联性，还要挖出背后的因果性，给计算出的结果赋予令人信服的解释才是统计的核心。

相比之下，机器学习只看重结果，只要模型能够对未知数据做出精确的预测，那这个模型能不能讲得清楚根本不是事儿。四十年前那句名言说得好：不管白猫黑猫，抓住耗子就是好猫。这句话用在机器学习上再合适不过了。

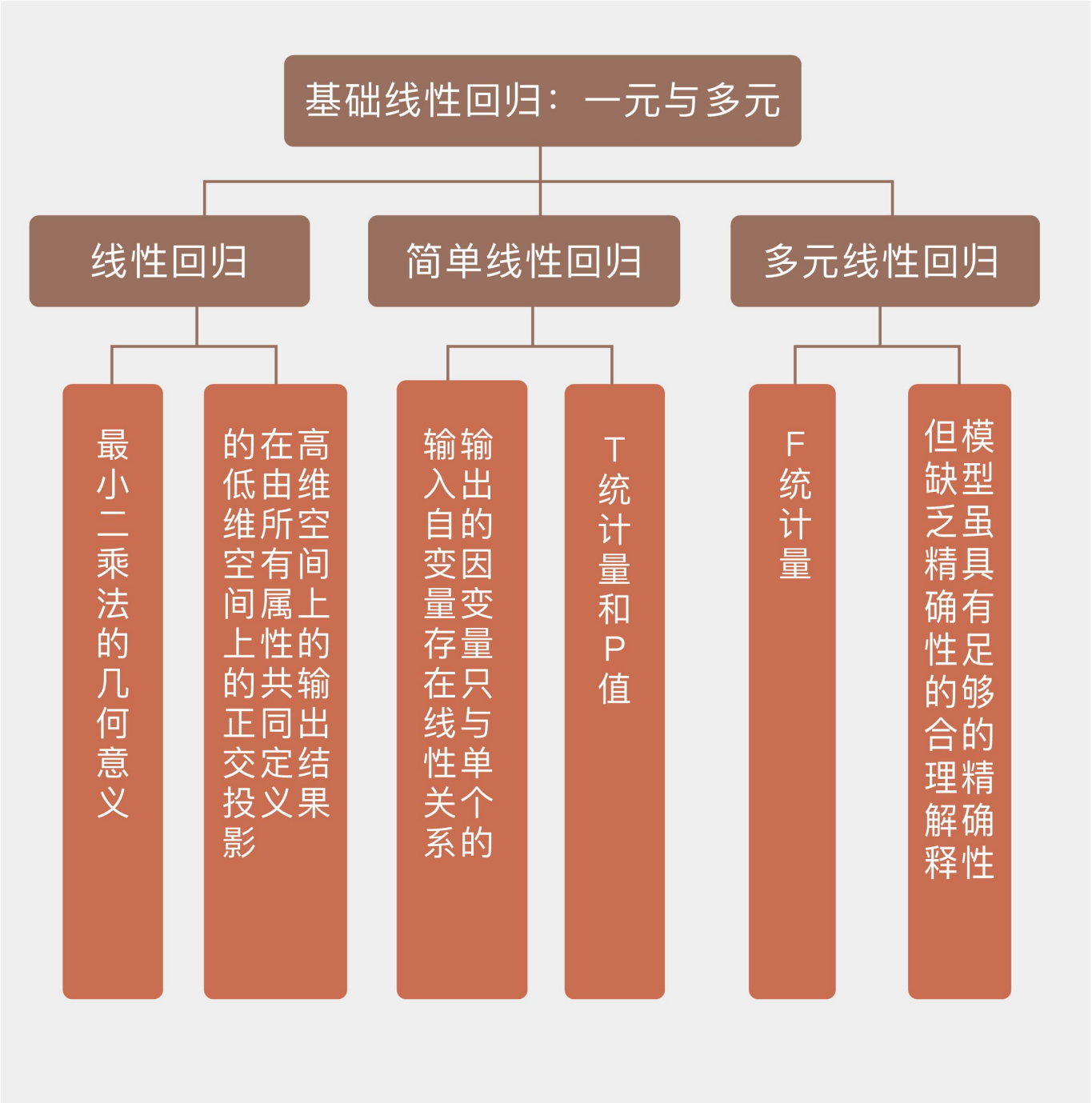
今天我向你介绍了基于最小二乘法的线性回归模型的理解以及从统计学角度的阐释，其要点如下：

> 线性回归拟合的是高维空间上的输出结果在由所有属性共同定义的低维空间上的正交投影；
>
> 简单线性回归的统计意义可以用 t 统计量和 p 值等指标描述；
>
> 多元线性回归的统计意义可以用 F 统计量描述，但回归结果可能缺乏对模型的解释能力；
>
> 机器学习与统计学的区别在于机器学习重于预测，统计学则重于解释。

本篇中的例子只以 2017~18 赛季英超联赛的数据作为训练数据集。如果使用不同赛季的数据训练的话，你就会发现每次拟合出来的系数都不一样。这样的事实会让你如何看待估计出的系数的准确性呢？

欢迎发表你的观点。

注：本文中的数据及代码可在下面地址查看。 https://github.com/tywang89/mlin40


基础线性回归：一元与多元
线性回归
简单线性回归
多元线性回归
最小二乘法的几何意义
高维空间上的输出结果在由所有属性共同定义的低维空间上的正交投影
输出的因变量只与单个的输入自变量存在线性关系
T统计量和P值
F统计量
模型虽具有足够的精确性但缺乏精确性的合理解释

上一篇　10 | 特征预处理

下一篇　12 | 正则化处理：收缩方法与边际化

精选留言 (6)

写留言

paradox
2018-08-16

1

老师
x.T 就变成了 N×(n+1)，每一行都是一个样本，那么x.T*β不也是一个样本作为一个整体么?
实在想不通，谢谢指点

作者回复: 正因为一个样本就是一个整体，所以要放在属性形成的空间里观察。

王大伟

2018-09-27

请问老师，标准误是如何计算的?

展开

作者回复: 样本的标准误约等于样本真正的标准差除以根号n，也就是样本容量的平方根。

BGu

2018-08-08

您好，您在多元回归例子中看了F stats 的数值大小，但是否应该用f stats的p值得出结论?

作者回复: 应该看，但我认为当F本身已经很小时，再看F的p值没什么意义。

林彦

2018-07-02

估计出的系数是观察数据的统计值。在做了数据分布的假设后，有较大的概率这些系数能让某个特定赛季的观测到的真实数据的某种误差最小，但系数并不是一组完全确定不变的值，它会收到训练数据的影响。(1)由线性回归假设得到的估计值和真实值之间的误差在不同赛季的数据是可变的，为了使某个赛季的的计算误差最小，计算出来的系数会不同；(2)不同赛季的数据中的噪声是不同的，也会影响计算出来的最优系数。...

展开

作者回复: 关键是你说的(2)，也就是噪声的问题。
一般的假设是观测结果是数据和噪声的叠加，每个数据集上的噪声都不一样，所以不同数据集上计算出的结果大差不差，但都是在真实值附近波动，不会和真实值吻合，这体现的就是前面所说的“样本内误差”的思想。
但长远来看，如果估计量本身是无偏的，那么在统计意义上，估计值就是准确的，不存在系统误差。在不同的数据集上估计，再取平均，估计的次数越多，均值就会越接近真值。
但是在这个例子里，结果的不同也不全是噪声的原因。毕竟每个赛季有升降级的球队，每个球队的人员也会有变化，可能不同赛季的数据不满足同分布的条件。

itzzy

2018-06-28

老师github上代码能加些注释吗？感谢！

展开

作者回复: 我这个编辑器不能输入汉字，所以索性英文注释也没加。所有代码基本上都是导入数据-调用功能类-画图的流程，如果哪里有问题可以把数据打印出来，或者查阅sklearn的文档。

我心飞扬

2018-06-28

当输出被写成 wTxwTx{\bf w} ^ T {\bf ...

极客时间版权所有: https://time.geekbang.org/column/article/9789?device=geekTime.android

...

展开

作者回复: 一旦模型参数定了，误差也就固定了，关键是怎么分解它。w^T x相当于把误差分散到样本上，误差是每个样本到计算出的超平面的距离；x^T \beta相当于把误差分散到属性上，在计算出来的超平面上做分解。