干货：7个常用的回归分析法！

中国统计网 2019-07-03

什么是回归分析?

回归分析是一种预测性的建模技术，它研究的是**因变量（目标）和自变量（预测器）之间的关系。**这种技术通常用于预测分析，时间序列模型以及发现变量之间的因果关系。例如，司机的鲁莽驾驶与道路交通事故数量之间的关系，最好的研究方法就是回归。

回归分析是建模和分析数据的重要工具。在这里，我们使用曲线/线来拟合这些数据点，在这种方式下，从曲线或线到数据点的距离差异最小。我会在接下来的部分详细解释这一点。

我们为什么使用回归分析?

如上所述，回归分析估计了两个或多个变量之间的关系。下面，让我们举一个简单的例子来理解它：

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。

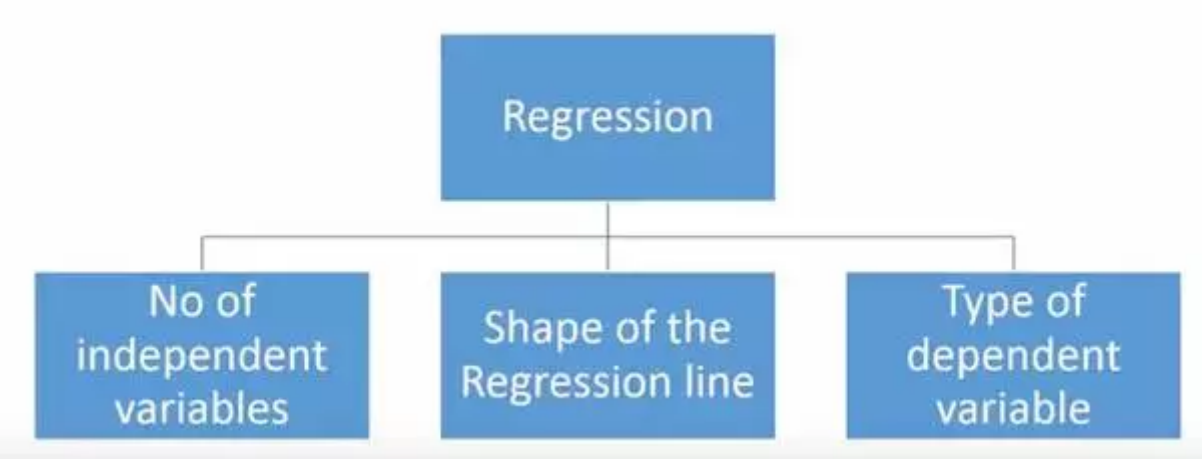
使用回归分析的好处良多。具体如下：

1. **它表明自变量和因变量之间的显著关系；**
2. **它表明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。

有多少种回归技术？

有各种各样的回归技术用于预测。这些技术主要有三个度量（自变量的个数，因变量的类型以及回归线的形状）。我们将在下面的部分详细讨论它们。



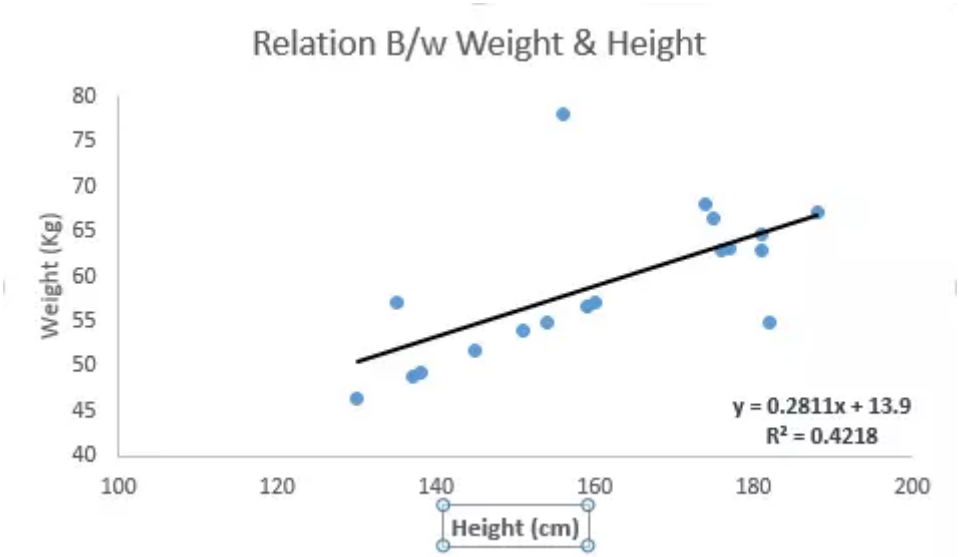
对于那些有创意的人，如果你觉得有必要使用上面这些参数的一个组合，你甚至可以创造出一个没有被使用过的回归模型。但在你开始之前，先了解如下最常用的回归方法：

01 Linear Regression线性回归

它是最为人熟知的建模技术之一。线性回归通常是人们在学习预测模型时首选的技术之一。在这种技术中，因变量是连续的，**自变量可以是连续的也可以是离散的，回归线的性质是线性的。**

线性回归使用最佳的拟合直线（也就是回归线）在因变量（Y）和一个或多个自变量（X）之间建立一种关系。

用一个方程式来表示它，即Y=a+b*X+e，其中a表示截距，b表示直线的斜率，e是误差项。这个方程可以根据给定的预测变量（s）来预测目标变量的值。

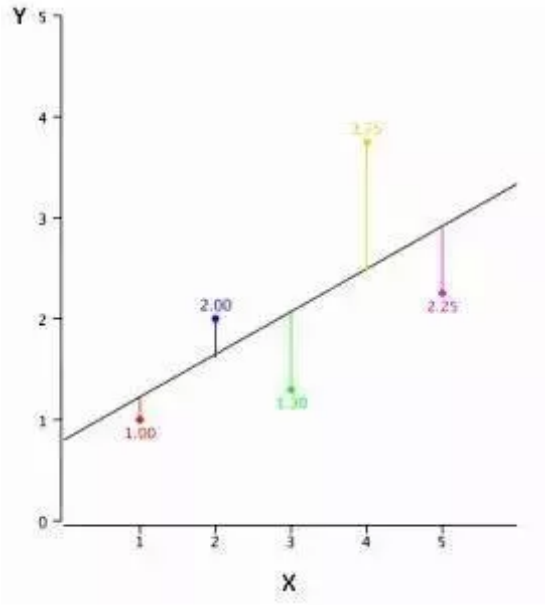


一元线性回归和多元线性回归的区别在于，多元线性回归有（>1）个自变量，而一元线性回归通常只有1个自变量。现在的问题是**“我们如何得到一个最佳的拟合线呢？”**。

如何获得最佳拟合线（a和b的值）？

这个问题可以使用最小二乘法轻松地完成。最小二乘法也是用于拟合回归线最常用的方法。对于观测数据，它通过最小化每个数据点到线的垂直偏差平方和来计算最佳拟合线。因为在相加时，偏差先平方，所以正值和负值没有抵消。

$$\min_{w} ||Xw - y||_2^2$$



我们可以使用R-square指标来评估模型性能。想了解这些指标的详细信息，可以阅读：模型性能指标 Part 1,Part 2.

> **要点：**
>
> **1.自变量与因变量之间必须有线性关系**
>
> **2.多元回归存在多重共线性，自相关性和异方差性。**
>
> **3.线性回归对异常值非常敏感。它会严重影响回归线，最终影响预测值。**
>
> **4.多重共线性会增加系数估计值的方差，使得在模型轻微变化下，估计非常敏感。结果就是系数估计值不稳定**
>
> **5.在多个自变量的情况下，我们可以使用向前选择法，向后剔除法和逐步筛选法来选择最重要的自变量。**

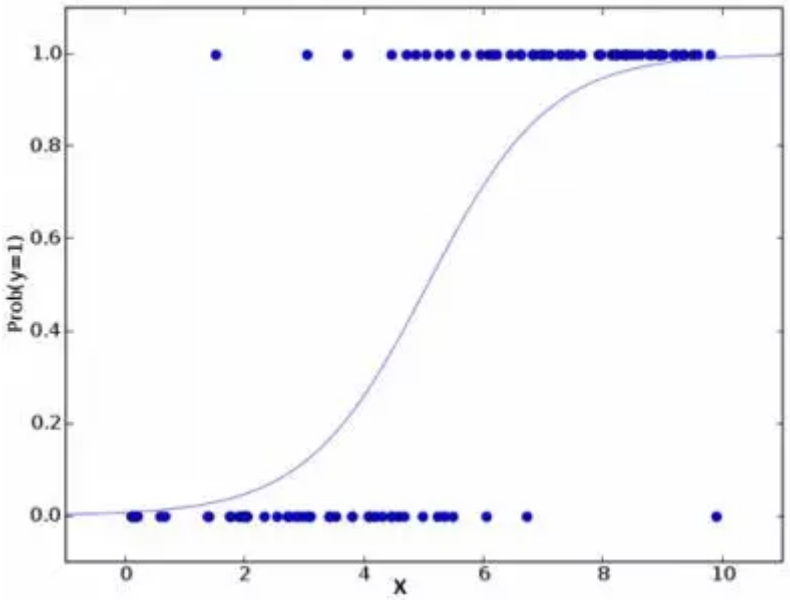
02 Logistic Regression逻辑回归

逻辑回归是用来计算“事件=Success”和“事件=Failure”的概率。当因变量的类型属于二元（1 / 0，真/假，是/否）变量时，我们就应该使用逻辑回归。这里，Y的值从0到1，它可以用下方程表示。

odds= p/ (1-p) = probability of event occurrence / probability of not event occurrenceln(odds) = ln(p/(1-p))logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

上述式子中，p表述具有某个特征的概率。你应该会问这样一个问题：“我们为什么要在公式中使用对数log呢？”。

因为在这里我们使用的是的二项分布（因变量），我们需要选择一个对于这个分布最佳的连结函数。它就是Logit函数。在上述方程中，通过观测样本的极大似然估计值来选择参数，**而不是最小化平方和误差（如在普通回归使用的）。**



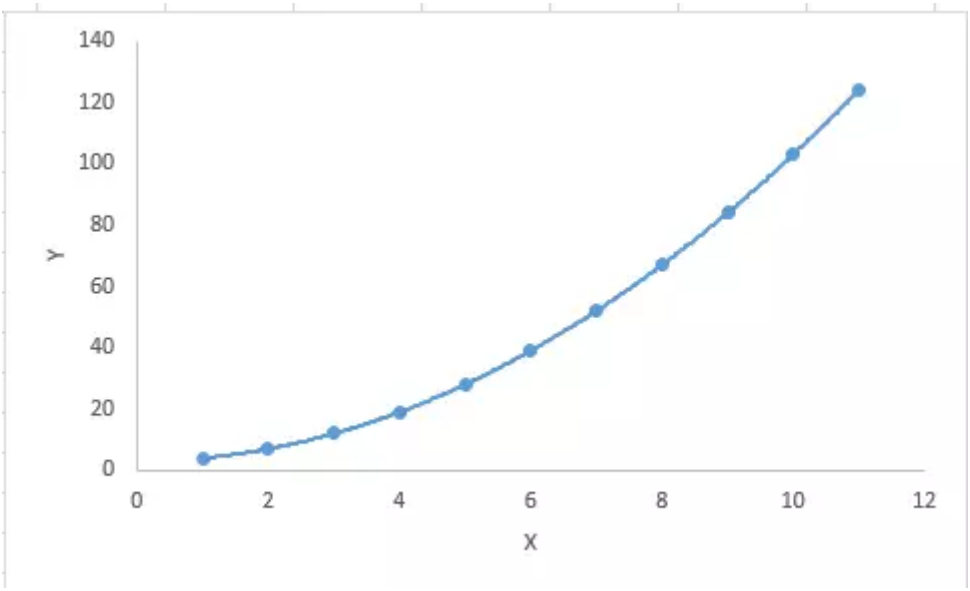
> **要点：**
>
> **1.它广泛的用于分类问题。**
>
> **2.逻辑回归不要求自变量和因变量是线性关系。它可以处理各种类型的关系，因为它对预测的相对风险指数OR使用了一个非线性的log转换。**
>
> **3.为了避免过拟合和欠拟合，我们应该包括所有重要的变量。有一个很好的方法来确保这种情况，就是使用逐步筛选方法来估计逻辑回归。**
>
> **4.它需要大的样本量，因为在样本数量较少的情况下，极大似然估计的效果比普通的最小二乘法差。**
>
> **5.自变量不应该相互关联的，即不具有多重共线性。然而，在分析和建模中，我们可以选择包含分类变量相互作用的影响。**
>
> **6.如果因变量的值是定序变量，则称它为序逻辑回归。**
>
> **7.如果因变量是多类的话，则称它为多元逻辑回归。**

03 Polynomial Regression多项式回归

对于一个回归方程，如果自变量的指数大于1，那么它就是多项式回归方程。如下方程所示：

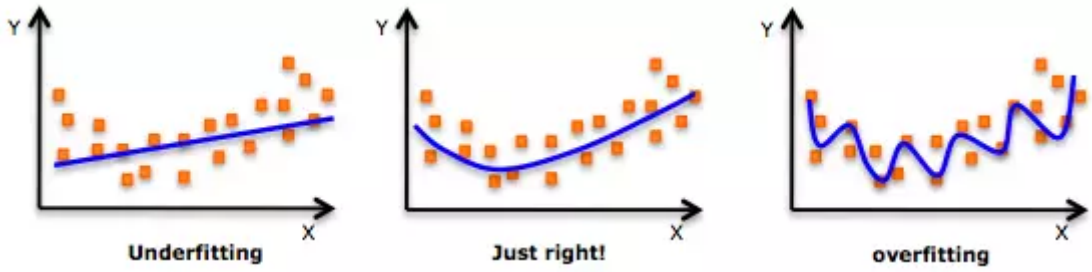
y=a+b*x^2

在这种回归技术中，最佳拟合线不是直线。而是一个用于拟合数据点的曲线。



> **重点：**
>
> 虽然会有一个诱导可以拟合一个高次多项式并得到较低的错误，但这可能会导致过拟合。你需要经常画出关系图来查看拟合情况，并且专注于保证拟合合理，既没有过拟合又没有欠拟合。

下面是一个图例，可以帮助理解：



明显地向两端寻找曲线点，看看这些形状和趋势是否有意义。更高次的多项式最后可能产生怪异的推断结果。

04 Stepwise Regression逐步回归

在处理多个自变量时，我们可以使用这种形式的回归。在这种技术中，自变量的选择是在一个自动的过程中完成的，其中包括非人为操作。

这一壮举是通过观察统计的值，如R-square，t-stats和AIC指标，来识别重要的变量。**逐步回归通过同时添加/删除基于指定标准的协变量来拟合模型。**

下面列出了一些最常用的逐步回归方法：

- 标准逐步回归法做两件事情。即增加和删除每个步骤所需的预测。
- 向前选择法从模型中最显著的预测开始，然后为每一步添加变量。
- 向后剔除法与模型的所有预测同时开始，然后在每一步消除最小显着性的变量。

这种建模技术的目的是使用最少的预测变量数来最大化预测能力。这也是处理高维数据集的方法之一。

05 Ridge Regression岭回归

岭回归分析是一种用于存在多重共线性（自变量高度相关）数据的技术。在多重共线性情况下，尽管最小二乘法（OLS）对每个变量很公平，但它们的差异很大，使得观测值偏移并远离真实值。岭回归通过给回归估计上增加一个偏差度，来降低标准误差。

上面，我们看到了线性回归方程。还记得吗？它可以表示为：

y=a+b*x这个方程也有一个误差项。完整的方程是：
y=a+b*x+e (error term), [error term is the value needed to correct for a prediction error between the observed and predicted value]
=> y=a+y= a+ b1x1+ b2x2+....+e, for multiple independent variables.

在一个线性方程中，预测误差可以分解为2个子分量。一个是偏差，一个是方差。预测错误可能会由这两个分量或者这两个中的任何一个造成。在这里，我们将讨论由方差所造成的有关误差。

岭回归通过收缩参数λ（lambda）解决多重共线性问题。看下面的公式

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_2^2}_{\text{Penalty}}$$

在这个公式中，有两个组成部分。第一个是最小二乘项，另一个是β2（β-平方）的λ倍，其中β是相关系数。为了收缩参数把它添加到最小二乘项中以得到一个非常低的方差。

> **要点：**
>
> 1.除常数项以外，这种回归的假设与最小二乘回归类似；
>
> 2.它收缩了相关系数的值，但没有达到零，这表明它没有特征选择功能
>
> 3.这是一个正则化方法，并且使用的是L2正则化。

06 Lasso Regression套索回归

它类似于岭回归，Lasso （Least Absolute Shrinkage and Selection Operator）也会惩罚回归系数的绝对值大小。此外，它能够**减少变化程度并提高线性回归模型的精度。**看看下面的公式：

$$= \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \underbrace{\|y - X\beta\|_2^2}_{\text{Loss}} + \lambda \underbrace{\|\beta\|_1}_{\text{Penalty}}$$

Lasso 回归与Ridge回归有一点不同，它使用的惩罚函数是绝对值，而不是平方。这导致惩罚（或等于约束估计的绝对值之和）值使一些参数估计结果等于零。使用惩罚值越大，进一步估计会使得缩小值趋近于零。这将导致我们要从给定的n个变量中选择变量。

> **要点：**
>
> 1.除常数项以外，这种回归的假设与最小二乘回归类似；
>
> 2.它收缩系数接近零（等于零），这确实有助于特征选择；
>
> 3.这是一个正则化方法，使用的是L1正则化；
>
> 如果预测的一组变量是高度相关的，Lasso 会选出其中一个变量并且将其它的收缩为零。

07 ElasticNet回归

ElasticNet是Lasso和Ridge回归技术的混合体。它使用L1来训练并且L2优先作为正则化矩阵。当有多个相关的特征时，ElasticNet是很有用的。Lasso 会随机挑选他们其中的一个，**而ElasticNet则会选择两个。**

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1).$$

Lasso和Ridge之间的实际的优点是，它允许ElasticNet继承循环状态下Ridge的一些稳定性。

> **要点：**

> 1.在高度相关变量的情况下，它会产生群体效应；
>
> 2.选择变量的数目没有限制；
>
> 3.它可以承受双重收缩。

除了这7个最常用的回归技术，你也可以看看其他模型，如Bayesian、Ecological和Robust回归。

如何正确选择回归模型?

当你只知道一个或两个技术时，生活往往很简单。如果结果是连续的，就使用线性回归。如果是二元的，就使用逻辑回归！然而，在我们的处理中，**可选择的越多，选择正确的一个就越难。**类似的情况下也发生在回归模型中。

在多类回归模型中，基于自变量和因变量的类型，数据的维数以及数据的其它基本特征的情况下，选择最合适的技术非常重要。以下是你要选择正确的回归模型的关键因素：

1.数据探索是构建预测模型的必然组成部分。在选择合适的模型时，比如识别变量的关系和影响时，它应该首选的一步。

2.比较适合于不同模型的优点，我们可以分析不同的指标参数，如统计意义的参数，R-square，Adjusted R-square，AIC，BIC以及误差项，另一个是Mallows' Cp准则。这个主要是通过**将模型与所有可能的子模型进行对比**（或谨慎选择他们），检查在你的模型中可能出现的偏差。

3.交叉验证是评估预测模型最好额方法。在这里，**将你的数据集分成两份（一份做训练和一份做验证）。**使用观测值和预测值之间的一个简单均方差来衡量你的预测精度。

4.如果你的数据集是多个混合变量，那么你就不应该选择自动模型选择方法，因为你应该不想在同一时间把所有变量放在同一个模型中。

5.它也将取决于你的目的。可能会出现这样的情况，一个不太强大的模型与具有高度统计学意义的模型相比，更易于实现。

6.回归正则化方法（Lasso，Ridge和ElasticNet）在高维和数据集变量之间多重共线性情况下运行良好。

End.

来源：经管之家