回归分析的五个基本假设

翻译 Noob_daniel 2017-07-25 17:30:38 71166 收藏 178

分类专栏： 统计学方法 文章标签： 统计学 预测 数学 数据

回归分析的五个基本假设

> 最近读到一篇很棒的文章，介绍了回归分析的五个基本假设，假设失效的影响及检验方法，现总结归纳如下。为己乃梳理巩固，亦期能有助于各位。

综述

回归分析是一种统计学上分析数据的方法，目的在于了解两个或多个变量间是否相关、相关方向与强度，并建立数学模型。以便通过观察特定变量（自变量），来预测研究者感兴趣的变量（因变量）。

总的来说，回归分析是一种参数化方法，即为了达到分析目的，需要设定一些“自然的”假设。如果目标数据集不满足这些假设，回归分析的结果就会出现偏差。因此想要进行成功的回归分析，我们就必须先证实这些假设。

回归分析的五个基本假设

1. **线性性 & 可加性**

   > 假设因变量为Y，自变量为X_1，X_2，则回归分析的默认假设为
   > $Y = b + a_1X_1 + a_2X_2 + \varepsilon$。
   > 线性性：X_1每变动一个单位，Y相应变动a_1个单位，与X_1的绝对数值大小无关。
   > 可加性：X_1对Y的影响是独立于其他自变量（如X_2）的。

2. **误差项（ε）之间应相互独立。**

   > 若不满足这一特性，我们称模型具有**自相关性**（Autocorrelation）。

3. **自变量（X_1，X_2）之间应相互独立。**

   > 若不满足这一特性，我们称模型具有**多重共线性性**（Multicollinearity）。

4. **误差项（ε）的方差应为常数。**

   > 若满足这一特性，我们称模型具有**同方差性**（Homoskedasticity），若不满足，则为**异方差性**（Heteroskedasticity）。

5. **误差项（ε）应呈正态分布。**

假设失效的影响

1. **线性性 & 可加性**

   > 若事实上变量之间的关系加性（如含有$X_1 \cdot X_2$项

点赞57 评论12 分享 收藏178 举报 关注 一键三连

> 可能导致很大的**泛化误差**（generalization error）

2. **自相关性（Autocorrelation）**

> 自相关性经常发生于时间序列数据集上，后项会受到前项的影响。当自相关性发生的时候，我们测得的标准差往往会**偏小**，进而会导致置信区间**变窄**。
> 假设没有自相关性的情况下，自变量X的系数为15.02而标准差为2.08。假设同一样本是有自相关性的，测得的标准差可能会只有1.20，所以置信区间也会从$(12.94, 17.10)$缩小到$(13.82, 16.22)$。

3. **多重共线性性（Multicollinearity）**

> 如果我们发现本应相互独立的自变量们出现了一定程度（甚至高度）的相关性，那我们就很难得知自变量与因变量之间真正的关系了。
> 当多重共线性性出现的时候，变量之间的联动关系会导致我们测得的标准差**偏大**，置信区间**变宽**。
> 采用岭回归，Lasso回归或弹性网（ElasticNet）回归可以一定程度上减少方差，解决多重共线性性问题。因为这些方法，在最小二乘法的基础上，加入了一个与回归系数的模有关的惩罚项，可以收缩模型的系数。
> 岭回归：$= argmin_{\beta \in \mathbb{R}^p}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2)$
> Lasso回归：$= argmin_{\beta \in \mathbb{R}^p}(\|y - X\beta\|_2^2 + \lambda\|\beta\|_1)$
> 弹性网回归：$= argmin_{\beta \in \mathbb{R}^p}(\|y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2)$
>
> $$where \|Z\|_p = \left(\sum_{i=1}^{N} |Z_i|^p\right)^{(1/p)}$$

4. **异方差性（Heteroskedasticity）**

> 异方差性的出现意味着误差项的方差不恒定，这常常出现在有异常值（Outlier）的数据集上，如果使用标准的回归模型，这些异常值的重要性往往被高估。在这种情况下，标准差和置信区间不一定会变大还是变小。

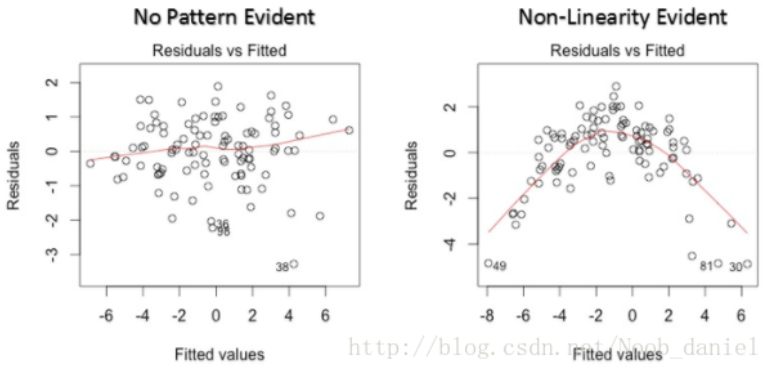
5. **误差项（ε）应呈正态分布**

> 如果误差项不呈正态分布，意味着置信区间会变得很不稳定，我们往往需要重点关注一些异常的点（误差较大但出现频率较高），来得到更好的模型。

假设检验方法

1. **线性性 & 可加性**
   观察残差（Residual）/估计值（Fitted Value，$\hat{Y}$）图。

点赞57 评论12 分享 收藏178 举报 关注 一键三连



http://blog.csdn.net/Noob_daniel

相较于图一（残差随机分布），图二的残差明显呈现了某种二次型趋势，说明回归模型没有抓住数据的某些非线性特征。
为了克服非线性性的影响，我们可以对**自变量**做一些非线性变换，如 $\log(X), \sqrt{X}, X^2 .. etc$

2. **自相关性（Autocorrelation）**
   观察杜宾-瓦特森统计量（Durbin-Watson Statistic）

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

该统计量的值落在$(0, 4)$内，$DW = 2$意味着没有自相关性，$0 < DW < 2$表明残差间有正的相关性，$2 < DW < 4$表明残差间有负的相关性。
经验上，如果$DW < 1$或$DW > 3$，则自相关性已经达到了需要示警的水平。如果事先给定了检验的方向（正/负相关性）和置信度α，也可以根据假设检验的思路进行对应计算。

3. **多重共线性性（Multicollinearity）**
   首先，可以通过观察自变量的散点图（Scatter Plot）来进行初步判断。
   然后，针对可能存在多重共线性性的变量，我们观察其方差膨胀系数（VIF–Variance Inflation Factor）
   假设回归模型为:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

对于变量X_j，可证得，其估计系数β_j的方差为:

$$\hat{var}(\hat{\beta}_j) = \frac{s^2}{(n-1)\hat{var}(X_j)} \cdot \frac{1}{1 - R_j^2}$$

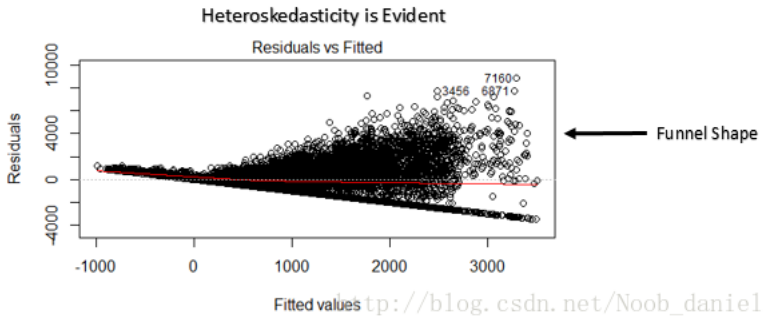
其中唯一与其它自变量有关的值是R_j^2，R_j^2是X_j关于其它自变量回归的残差:

$$X_j = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \cdots + \beta_k X_k + \varepsilon$$

$\frac{1}{1-R_j^2}$便称作VIF，若$VIF < 3$，说明该变量基本不存在多重共线性性问题，若$VIF > 10$，说明问题比较严重。

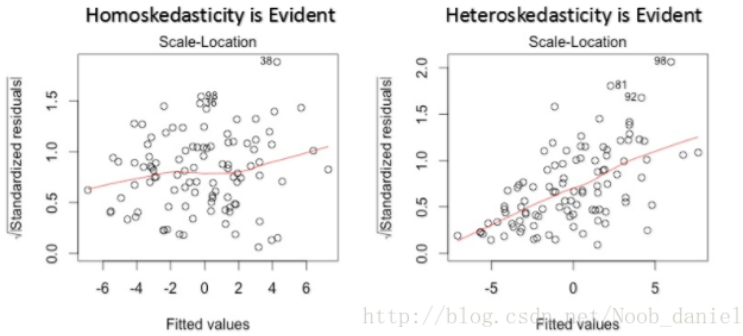
4. **异方差性（Heteroskedasticity）**
   观察残差（Residual）/估计值（Fitted Value，$\hat{Y}$）图。

点赞57 评论12 分享 收藏178 举报 关注 一键三连



若该图呈现如上图所示的“漏斗形”，即随着$\hat{Y}$的变化，残差有规律的变大或变小，则说明存在明显的异方差性。

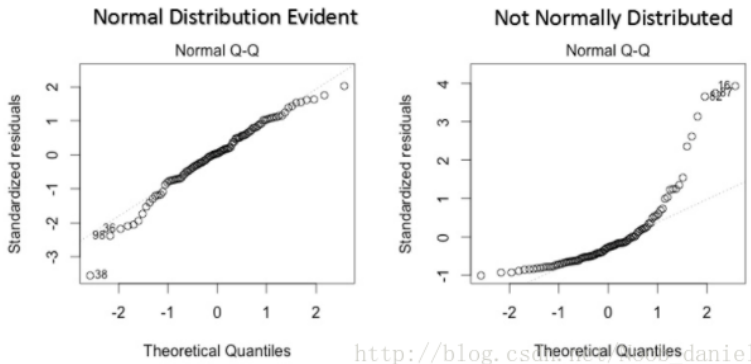
或观察残差的标准差（$\sqrt{StandardizedResidual}$）/估计值图(Scale Location Plot)。



也可以看出，异方差数据集呈现出明显的趋势性。

为了克服异方差性的影响，我们可以对**因变量**做一些非线性变换，如$\log(Y), \sqrt{Y}..etc$

5. **误差项（ε）应呈正态分布**
   方法一：观察Q-Q Plot (quantile-quantile plot)



如果误差项满足正态分布，Q-Q Plot里的散点会近似的落在一条直线上。若不满足正态分布，则散点会偏离该直线。

方法二：进行正态检验–如Kolmogorov-Smirnov检验、Shapiro–Wilk检验

点赞57 评论12 分享 收藏178 举报 关注 一键三连

总结

如果走在了错误的道路上，就算健步如飞，也只会渐行渐远。回归分析是久经考验的很有效的分析手段，但在使用的过程中，我们一定要时刻注意数据集是否满足建模的基本假设，是否需要调整。上述的图表在用R做回归时都会自动生成，更好的理解和观察它们会帮助我们更好地运用回归分析！

有诗云：

悟以往之不谏，知来者之可追。实迷途其未远，觉今是而昨非。

舟遥遥以轻扬，风飘飘而吹衣。问征夫以前路，恨晨光之熹微。

–《归去来兮辞》

线性回归的五个基本假设 Matrix-yang的博客 1万+
回归分析的五个基本假设 最近读到一篇很棒的文章，介绍了回归分析的五个基本假设，假设失效的…

R语言与回归分析几个假设的检验 beta 3万+
一、从线性回归的假设说起 对于线性回归而言，若要求回归估计有一些良好性质比如无偏性，…

优质评论可以帮助作者获得更高权重 评论

小鸡杂毛的女： 博主您好，我想问一下， 逻辑回归是假设数据服从伯努利分布,通过极大似然函数的方法,运用梯度上升/下降法来求解参数,从而实现数据的二分类。但是为什么西瓜书上说逻辑回归的优点之一是直接对分类可能性进行建模，而无需事先假设数据分布？ 2 年前 回复 1

weixin_42663919： 不好意思，VIF好像解释有点问题。“检验模型解释能力的检验统计指标为R^2（样本可决系数）”（参考https://mp.weixin.qq.com/s?src=11&timestamp=1605241767&ver=2703&signature=SyIBZtWhEDgmw-YNd3YPMsyXWAPAh056ZUlEOboVcJd7q

登录 查看 12 条热评

相关推荐

回归分析的假设条件_不争而善胜的博客 3-19
数据什么样就能扔进回归分析回归分析

机器学习---最小二乘线性回归模型的5个基本假设(Machin... 3-14
在之前的文章《机器学习---线性回归(Machine Learning Linear Regression)》中说到,使用最小二乘…

[译]回归分析的基本假设 GG的专栏 1万+
原文地址：《Going Deeper into Regression Analysis with Assumptions, Plots & Solutions》 …

机器学习---最小二乘线性回归模型的5个基本假设（Machin… weixin_34368949的博客 663
在之前的文章《机器学习---线性回归（Machine Learning Linear Regression）》中说到，使用最小…

回归分析-(多元)线性回归分析基础(Linear Regression)_c_pump... 3-15
[2]实用回归分析,第二版(何晓群) [3]计量经济学,第三版(庞皓) [4](可参考)回归分析的五个基本假设htt…

【回归】回归分析中的假定:什么假定?为什么要满足?(为... 2-26
泛化误差《回归分析的五个基本假设》,翻译自《Going Deeper into Regression Analysis with As…

SPSS篇—回归分析 小白数据营的博客 4万+
之前跟大家介绍了一款做数据分析的利器—SPSS，不知道大家对这个软件的熟悉程度有没有提高…

OLS的基本假定 Eliza的学习记录 1万+
小样本OLS的基本假定 对于小样本来说，为了得到样本统计量关于总体的BLUE（最佳线性无偏估…

回归分析详解及matlab实现 我已经不能按照自己最初的意愿去生活了 15万+
回归分析方法 想要资源的请关注公众号： 在一起的足球自动获取资源和数十种经典算法，帮助各…

残差分析（残差原理与标准化残差
1、残差分析定义 在回归模型 中，假定

点赞57 评论12 分享 收藏178 举报 关注 一键三连