简介

回归分析：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，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$Y = b + a_1 * X_1 + a_2 * X_2 + \epsilon$。 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不满足线性性（如含有x_1^2, x_1^3项），或不满足可加性（如含有$X_1 \cdot X_2$项），则模型将无法很好的描述变量之间的关系，极有可能导致很大的**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：$\operatorname{argmin}_{\beta\in\mathbb{R}^p}\left(\|y-X\beta\|_2^2+\lambda\|\beta\|_2^2\right)$

   Lasso回归：$\operatorname{argmin}_{\beta\in\mathbb{R}^p}\left(\|y-X\beta\|_2^2+\lambda\|\beta\|_1\right)$

   弹性网回归：$\operatorname{argmin}_{\beta\in\mathbb{R}^p}\left(\|y-X\beta\|_2^2+\lambda_1\|\beta\|_1+\lambda_2\|\beta\|_2^2\right)$

4. **异方差性（Heteroskedasticity）**

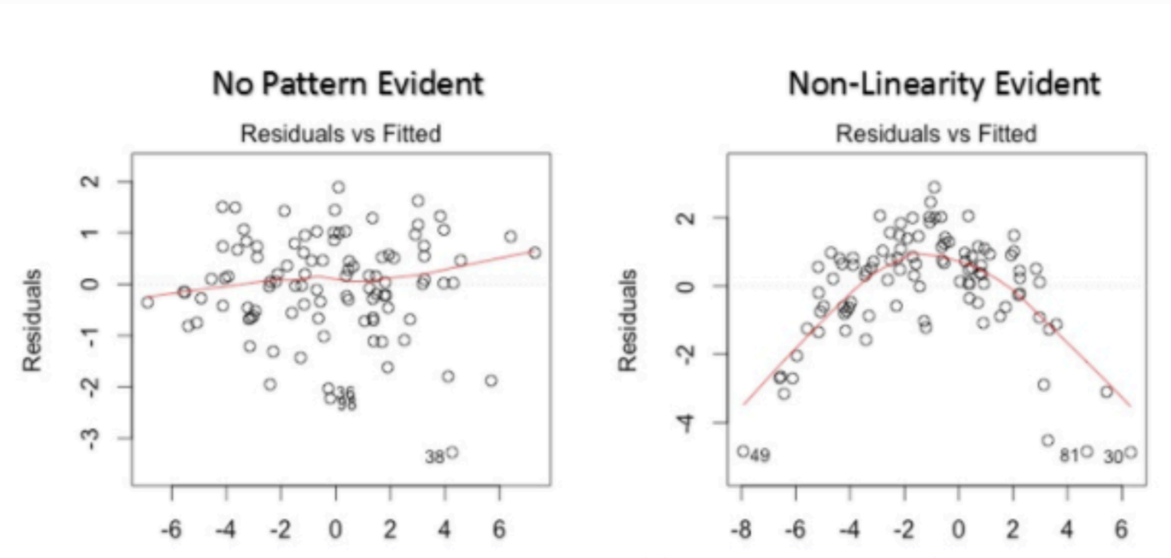
   异方差性的出现意味着误差项的方差不恒定，这常常出现在有异常值（Outlier）的数据集上，如果使用标准的回归模型，这些异常值的重要性往往被高估。在这种情况下，标准差和置信区间不一定会变大还是变小。

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**

   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$log(X), \sqrt{X}, X^2\ldots$.

2. **自相关性（Autocorrelation）**

   观察杜宾-瓦特森统计量:

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

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。

假设回归模型为：

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\varepsilon$$

对于变量X_j,可证得，其估计系数β_j的方差为:

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

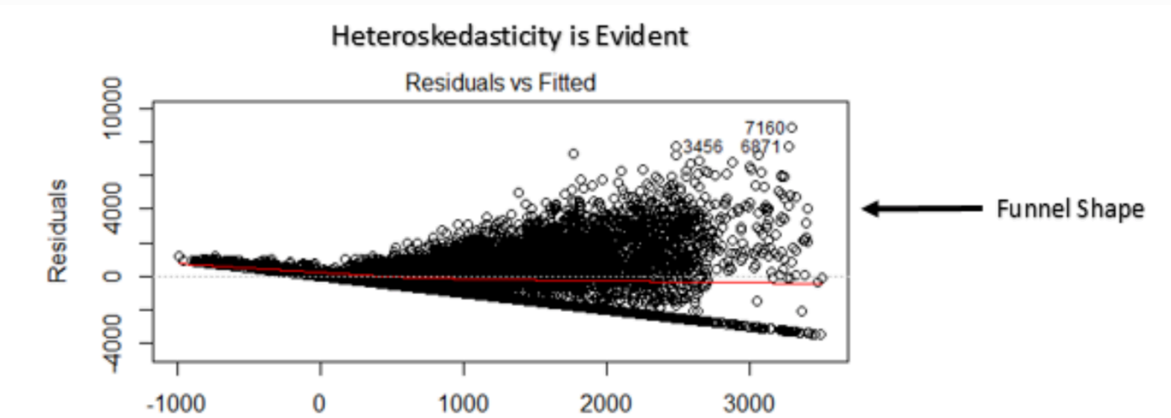
其中唯一与其它自变量有关的值是R_j^2,R_j^2是X_j关于其它自变量回归的残差：

$$X_j=\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_{j-1}X_{j-1}+\beta_{j+1}X_{j+1}+\cdots+\beta_kX_k+\varepsilon$$

$\frac{1}{1-R_j^2}$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，Ŷ）：

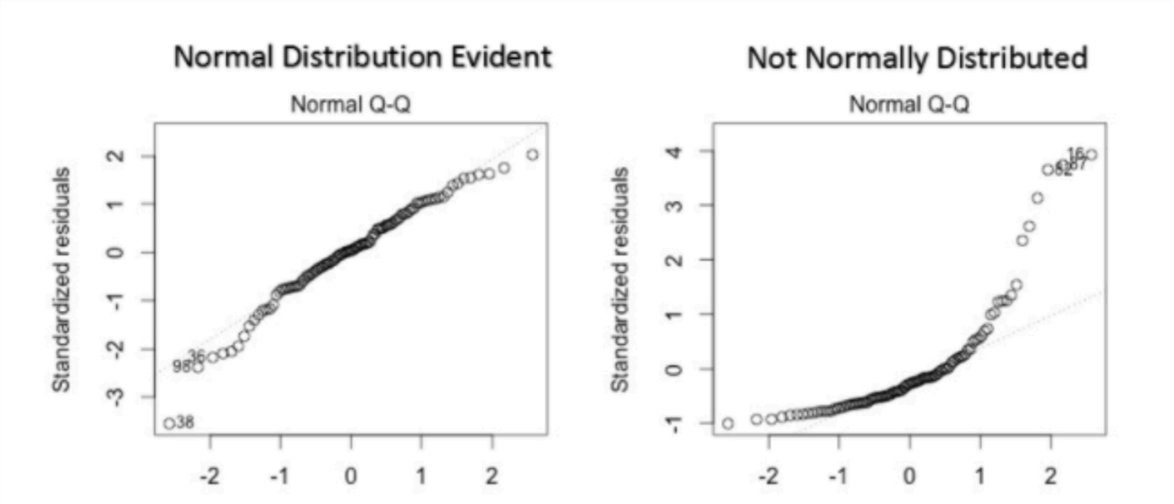


若该图呈现如上图所示的“漏斗形”，即随着$y^{\star}$的变化，残差有规律的变大或变小，则说明存在明显的异方差性。

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$log(Y),\sqrt{Y}$

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.