**24. 一元线性回归**

回归分析是研究一个或多个变量（因变量）与另一些变量（自变量）之间关系的统计方法。主要思想是用最小二乘法原理拟合因变量与自变量间的最佳回归模型（得到确定的表达式关系）。其作用是对因变量做解释、控制、或预测。

**回归与拟合的区别：**

拟合侧重于调整曲线的参数，使得与数据相符；而回归重在研究两个变量或多个变量之间的关系。它可以用拟合的手法来研究两个变量的关系，以及出现的误差。

**回归分析的步骤：**

（1）获取自变量和因变量的观测值；

（2）绘制散点图，并对异常数据做修正；

（3）写出带未知参数的回归方程；

（4）确定回归方程中参数值；

（5）假设检验，判断回归方程的拟合优度；

（6）进行解释、控制、或预测。

**一、一元线性回归模型为**

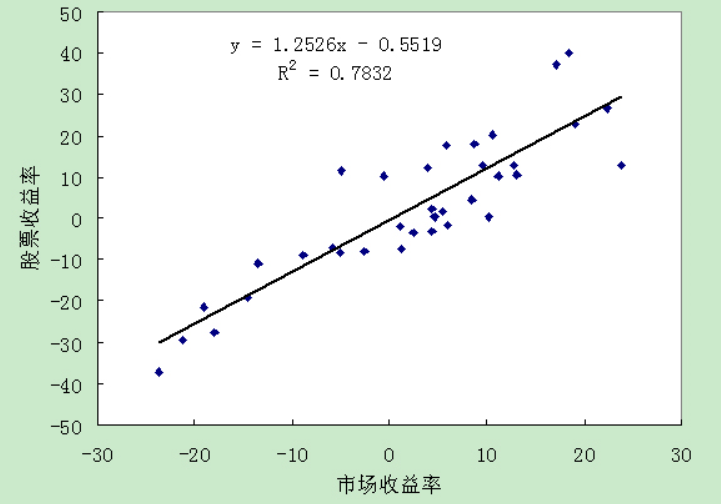
Y=𝛽0+𝛽1X+ε

其中 X是自变量，Y是因变量，𝛽0,𝛽1是待求的未知参数，𝛽0也称为截距；ε是随机误差项，也称为残差，通常要求ε满足：

① ε的均值为0；

② ε的方差为𝜎2；

③ 协方差COV(εi, εj)=0，当i≠j时。即对所有的i≠j, εi与εj 互不相关。



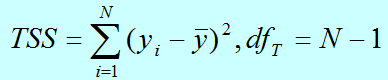
**二、**用**最小二乘法原理**，得到最佳拟合效果的值：

， 

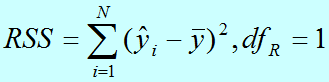
**三、假设检验**

1. 拟合优度检验

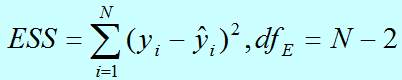
总偏差平方和及其自由度：



回归平方和及其自由度：



残差平方和及其自由度：



TSS=RSS+ESS



通常可以认为当R2大于0.9时，所得到的回归直线拟合得较好，而当R2小于0.5时，所得到的回归直线很难说明变量之间的依赖关系。

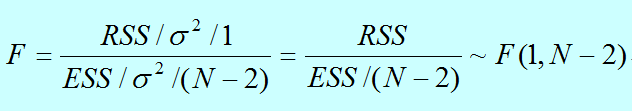
2. 回归方程参数的检验

回归方程反应了因变量Y随自变量X变化而变化的规律，若𝛽1=0，则Y不随X变化，此时回归方程无意义。所以，要做如下假设检验：

H0: 𝛽1=0, H1: 𝛽1≠0；

（1） F检验

若𝛽1=0为真，则回归平方和RSS与残差平方和ESS/(N-2)都是𝜎2的无偏估计，因而采用*F*统计量：

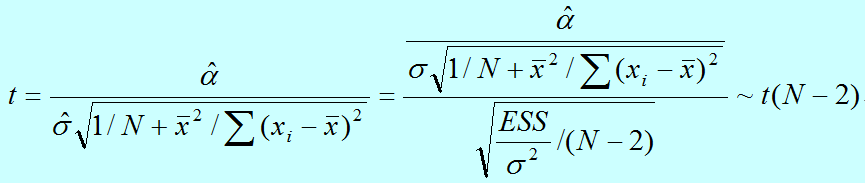


来检验原假设β1=0是否为真。

（2）T检验

对H0: 𝛽1=0的T检验与F检验是等价的（t2=F）。

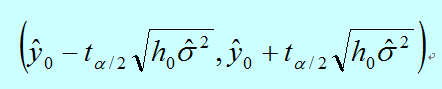
对H0: 𝛽0=0的T检验，若𝛽0=0为真，t统计量为：



3. 用回归方程做预测

得到回归方程后，预测X=x0处的Y值.

的预测区间为：



其中tα/2的自由度为N-2.

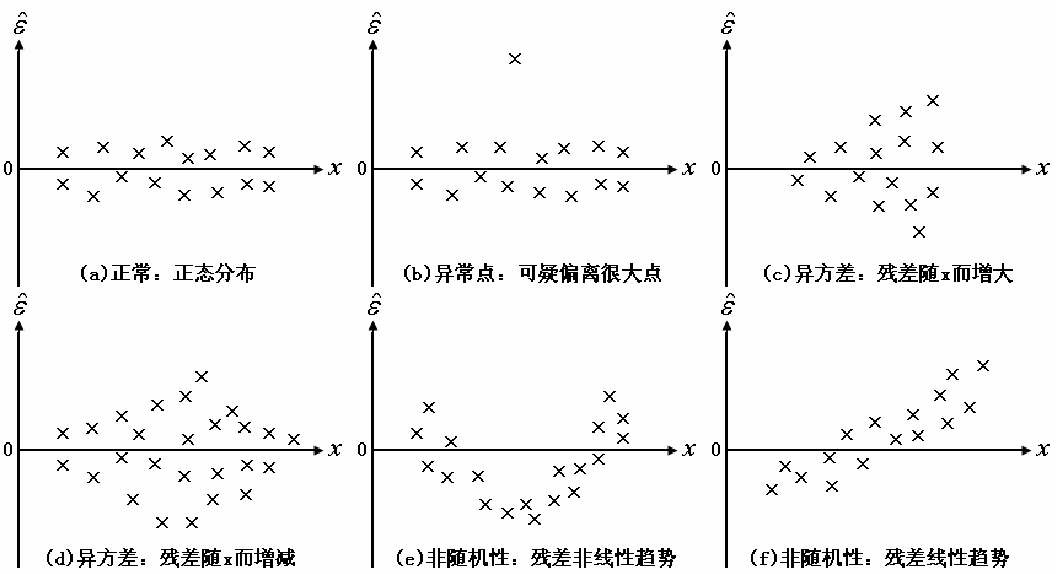
SAS中是用model语句中的clm选项来计算预测区间的。

4. 回归诊断

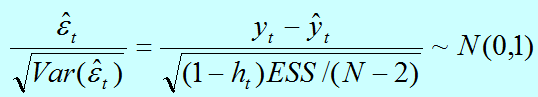
（1）残差图分析

残差图就是以残差为纵坐标，某一个合适的自变量为横坐标的散点图。

回归模型中总是假定误差项是独立的正态分布随机变量，且均值为零和方差相等为𝜎2. 如果模型适合于观察到的数据，那么残差作为误差的无偏估计，应基本反映误差的假设特征。即残差图应该在零点附近对称地密布，越远离零点的地方就疏散（在形象上似有正态趋势），则认为模型与数据拟合得很好。



若残差图呈现如图（a）所示的形式，则认为建立的回归模型正确，更进一步再诊断“学生化残差”是否具有正态性：



图（b）表明数据有异常点，应处理掉它重新做回归分析（在SAS的REG回归过程步中用来度量异常点影响大小的统计量是COOKD统计量）；

图（c）残差随x的增大而增大，图（d）残差随x的增大而先增后减，都属于异方差。此时应该考虑在回归之前对数据y或x进行变换，实现方差稳定后再拟合回归模型。原则上，当误差方差变化不太快时取变换；当误差方差变化较快时取变换log y或ln y；当误差方差变化很快时取变换1/y；还有其他变换，如著名的Box-Cox幂变换.

图（e）（f）表示选用回归模型是错误的。

（2）共线性

回归分析中很容易发生模型中两个或两个以上的自变量高度相关，从而引起最小二乘估计可能很不精确（称为共线性问题）。

在实际中最常见的问题是一些重要的自变量很可能由于在假设检验中*t*值不显著而被不恰当地剔除了。共线性诊断问题就是要找出哪些变量间存在共线性关系。SAS的REG过程步提供了特征值法、条件指数*collin*和方差膨胀因子*vif.*

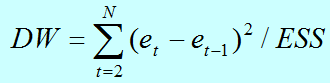
（3）误差的独立性

回归分析之前，要检验误差的独立性。若误差项不独立，那么回归模型的许多处理，包括误差项估计、假设检验等都将没有推导依据。

由于残差是误差的合理估计，因此检验统计量通常是建立在残差的基础上。检验误差独立性的最常用方法，是对残差的一阶自相关性进行Durbin-Watson检验。

H0: 误差项是相互独立的; H1: 误差项是相关的

检验统计量：



*DW*接近于0，表示残差中存在正自相关；如果*DW*接近于4，表示残差中存在负自相关；如果DW接近于2，表示残差独立性。在给定显著水平α下，我们可以查Durbin-Watson表得到不能拒绝独立性原假设的区间。

**四、PROC REG过程步**

**1. 基本语法**

前面三种回归分析都可以用PROC REG过程步实现。

基本语法：

PROC REG data = 数据集;

MODEL *因变量 = 自变量列表 </可选项>*;

说明：MODEL语句用来指定因变量和自变量；

常用的输出可选项：

STB——输出标准化偏回归系数矩阵

CORRB——输出参数估计矩阵

COLLINOINT——对自变量进行共线性分析

P——输出个体观测值、预测值及残差 （R/CLM/CLI包含P）

R——输出每个个体观测值、残差及标准误差

CLM——输出因变量均值95%的置信界限的上下限

CLI——对各预测值输出95%的置信界限的上下限

MSE——要求输出随机扰动项方差𝜎2的估计

与残差分析有关的可选项

 VIF——输出变量间相关性的方差膨胀系数，VIF越大，说明由于共线性存在，使方差变大；

COLLIN——输出条件数，它表示最大的特征值与每个自变量特征值之比的平方根。一般情况下，条件数越大越可能存在共线性；

TOL——表示共线性水平的容许值，TOL越小说明其可用别的自变量解释的部分多，自然可能与别的自变量存在共线性关系；

DW——输出Durbin-Watson统计量；

influence——对异常点进行诊断，对每一观测点输出统计量（Cook’s D > 50%, defits/debetas > 2说明该点影响较大）。

**2. 绘制回归分析的图形**

在PROC REG过程步加入绘图选项语句即可。

基本语法：

PROC REG data = 数据集PLOTS *=* (*图形类型*);

可选的绘图类型：

FITPLOT——带回归线、置信预测带的散点图；

RESIDUALS——自变量的残差图；

DIAGNOSTICS——诊断图（包括下面各图）；

COOKSD——Cook's D统计量图；

OBSERVEDBYPREDICTED——根据预测值的因变量图；

QQPLOT ——检验残差正态性的QQ图；

RESIDUALBYPREDICTED——根据预测值的残差图；

RESIDUALHISTOGRAM——残差的直方图；

RFPLOT——残差拟合图；

RSTUDENTBYLEVERAGE——杠杆比率的学生化残差图；

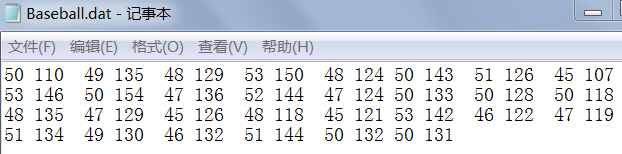
RSTUDENTBYPREDICTED——预测值的学生化残差图；

注：残差图(RESIDUALS)和诊断图(DIAGNOSTICS)是自动生成的，根据模型也有其它默认的图形输出；若只绘制指定的图形需要加上ONLY:

**PROC** **REG** data = 数据集 PLOTS(ONLY) = (图形类型);

**例1 （线性回归）**

30名儿童棒球选手的数据（C:\MyRawData\Baseball.dat），变量包括击球高度、球飞出的距离：



读入数据，以球飞出的距离为因变量，击球高度为自变量做线性回归分析。

**代码：**

**data** hits;

infile 'c:\MyRawData\Baseball.dat';

input Height Distance @@;

**run**;

**proc** **reg** data = hits PLOTS(ONLY) = (DIAGNOSTICS FITPLOT);

model Distance = Height /r clm cli dw;

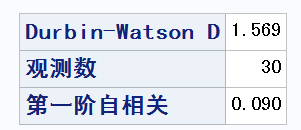
title 'Results of Regression Analysis';

**run**;

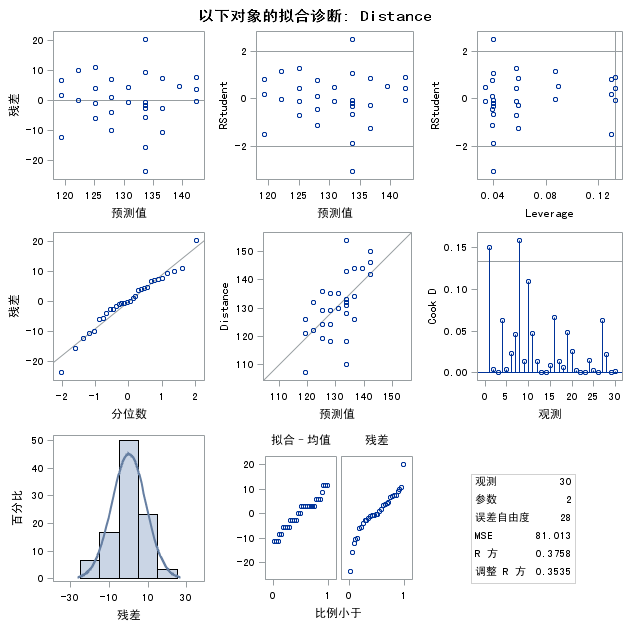
运行结果：

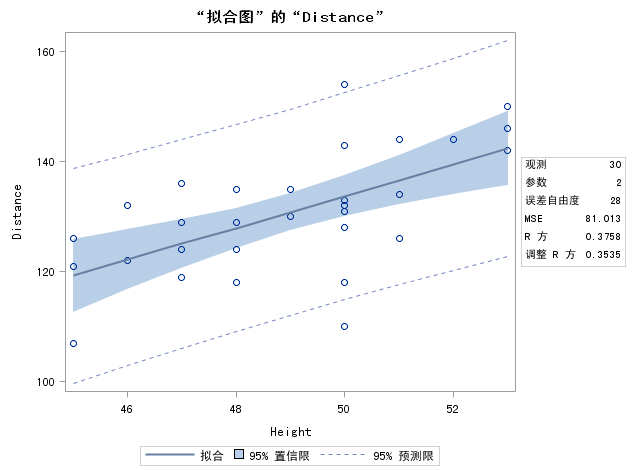












结果说明：

（1）只有一个自变量的，故其自由度为DF=1，因变量的自由度为N-1=29；误差的自由度为N-1-1=28；

（2）离差平方和(TSS)=回归平方和(RSS)+残差平方和(ESS)，即

3633.86667= 1365.50831+ 2268.35836

离差平方和平均值为

MSR=RSS/df=1365.50831/1= 1365.5083，

误差项的离差平方和平均值为

MSE=ESS/df= 2268.35836/28= 81.01280；

（3）假设检验H0：回归系数β=0.

F(1,28)=MSR/MSE=1365.5083/81.01280=16.86

P值=0.0003<α=0.05，说明斜率β不为0（与后面的t检验结果相同）；

（4）均方根误差：Root MSE= MSE开根号= 9.00071；

变异系数（或称方差系数）：

Coeff Var=均方根误差/因变量均值×100=6.88479；

（5）拟合优度R2=0.3758，校正的R2=0.3535；说明线性回归关系并不强，球飞出的距离可能还和选手的年龄、经验有关；

（6）参数估计表给出了截距和斜率的估计值，得到回归方程：

Distance = -11.00859 + (2.89466 \* Height)

斜率2.89466的t检验P值=0.0003<α=0.05，说明该结果在显著水平α下有统计学意义；但截距的t检验结果并不显著。

（7） Durbin-Watson D 值=1.569，说明误差基本上是独立的（靠近2说明误差基本上是独立的，小于2说明是正相关）；

（8）COOKD统计量大于50%，则认为该观测是异常点。