

R语言回归

回归检验

正态性检验

$$\mu \sim N(0, \sigma^2)$$

JB检验 (Jarque-Bera检验) :

```
library(tseries)
jarque.bera.test(lm$residuals) # Jarque-Bera检验, lm为前面定义的线性模型,
residuals为模型的残差
```

适用样本量 $n \geq 30$

W检验 (Shapiro-Wilk检验) :

```
shapiro.test(lm$residuals)
```

适用样本量 $3 \leq n \leq 50$

判断通过: 看 P 值

P 值高于0.05, 认为数据具有近似服从均值为0的正态分布。

解决办法 (未通过检验)

Box-Cox变换

```
library(MASS)
bc = boxcox(y~x1+x2+x3+x4, data=data, lambda=seq(-5, 5, 0.01))
# lambda取值为[-5, 5]上步长为0.01的值
lambda = bc$x[which.max(bc$y)] # 找到y最大的lambda值
y_bc = (data$y^lambda-1)/lambda # lambda不为0, 若为0则直接log(y)
lm_bc = lm(y_bc~x1+x2+x3+x4, data=data)
summary(lm_bc)
```

变换后再进行正态性检验, 查看是否 $p \geq 0.05$

异方差性检验

bptest

```
library(lmtest)
result = bptest(lm)
result
```

判断通过：看 P 值

P 值高于0.05，认为模型误差项满足齐性假设（同方差）。

解决办法（未通过检验）

加权最小二乘：

核心：为更大的项，赋予更小的权重（通常取 $W = x_i^{-2}$ ）

```
e = resid(lm) # 提取回归残差
abse = abs(e) # 取绝对值
cor.test(data$x1, abse, method="spearman") # 计算各自变量和残差绝对值的等级相关系数
cor.test(data$x2, abse, method="spearman")
# ...
cor.test(data$x4, abse, method="spearman")
w = data$xi^(-2) # xi为系数绝对值最大者，作为权重参考，若p值更小，可选其它变量
model_wls = lm(y~x1+x2+x3+x4, data=data, weights=w)
summary(model_wls)
```

变换后再进行异方差性检验，查看是否 $p \geq 0.05$

自相关性检验

各残差 μ_i 存在相关性。

dwtest

```
result = dwtest(lm)
result
```

前提 $n \geq 15$ ，数据和时间有关。

查看Durbin-Waston检验表，当 DW 值介于 $(d_U, 4 - d_U)$ 之间，认为无自相关性。

解决办法（未通过检验）

差分法

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \mu_i \quad (1)$$

$$y_{i-1} = \beta_0 + \beta_1 x_{1(i-1)} + \cdots + \beta_n x_{n(i-1)} + \mu_{i-1} \quad (2)$$

(1) - (2), 得

$$(y_i - y_{i-1}) = \beta_1(x_{1i} - x_{1(i-1)}) + \cdots + \beta_n(x_{ni} - x_{n(i-1)}) + (\mu_i - \mu_{i-1})$$

随后对新方程进行检验, 若不通过, 继续差分。

```
d1m = lm(diff(y)~diff(x1)+diff(x2)+diff(x3)+diff(x4) - 1, data=data) # 不包含常数项, 需要-1
```

多重共线性检验

回归方程中, 某自变量 x_i 和其余自变量存在高度关联。

方差扩大因子法 (vif)

```
result = vif(lm)
result
```

判断通过: 看各变量对应值是否 < 10 , 若有变量大于等于10, 则存在多重共线性。

非线性回归

多项式回归（交叉项, 幂次项）

$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_1^2 + b_4 x_2^2 + b_5 x_1 x_2$$

```
lm = lm(y~x1+x2 + I(x1^2) + I(x2^2) + I(x1 * x2))
```

一元非线性回归

```
nlm = nls(formula, data, start)
# formula: 回归方程形式
# start: 参数初始值
```

定性变量（范畴变量）

数据处理

```
x_1 = ifelse(data$x1 == "A", 1, 0) # 设置虚拟变量x_1, 当x1为A时取1, 否则取0
x_2 = ifelse(data$x1 == "B", 1, 0)
```