加权线性回归 R 教程

小狗

目录

1 加权线性回归（Weighted Linear Regression） 1

1.1 1. 放宽线性模型的假设 1
1.2 2. 加权线性回归的实现 2
1.3 3. 普通最小二乘法（OLS）回归 3
1.4 4. 加权最小二乘法（WLS）回归 4
1.5 5. 加权残差与模型诊断 6
1.6 6. 帽子矩阵（Hat Matrix）与杠杆值 7
1.7 7. 总结 8

1 加权线性回归（Weighted Linear Regression）

1.1 1. 放宽线性模型的假设

1.1.1 1.1 常规线性回归模型的假设

在线性回归模型中，假设观测值 Y_i 与预测变量 x_i 的关系是线性的，模型形式如下：

$$Y_i = x_i^T \beta + \epsilon_i$$

其中：- Y_i：第 i 个观测值的响应变量。- x_i：第 i 个观测值的预测变量（解释变量）的向量。- β：回归系数向量，表示预测变量对响应变量的影响。- ϵ_i：误差项，表示实际观测值与模型预测值之间的差异。

误差项 ϵ_i 的假设：1. **零均值：** $E(\epsilon_i) = 0$，即误差的期望值为零。2. **同方差性：** $\mathrm{Var}(\epsilon_i) = \sigma^2$，所有观测点的误差方差相同。3. **独立性：**各观测点的误差项相互独立。4. **正态分布：** ϵ_i 服从正态分布 $N(0, \sigma^2)$。

1.1.2 1.2 异方差性问题

在某些应用中，观测点的误差方差可能不相等，这种现象称为**异方差性**。例如，某些数据点可能更可靠，应当赋予它们更大的权重。

问题：如果继续使用普通最小二乘法（OLS）不考虑异方差性，结果可能会被方差较大的数据点影响，导致估计的不准确。

为了解决这一问题，**加权线性回归（WLS）**引入了权重，使得更可靠的观测值在模型中占据更大权重。

1.2 2. 加权线性回归的实现

在加权线性回归中，权重 w_i 是已知的、固定的，用来反映第 i 个观测点的重要性。模型的方差公式可以写成：

$$\mathrm{Var}(Y_i|x_i) = \frac{\phi}{w_i}$$

权重 w_i 越大，观测点的方差越小，表明该观测值对回归模型更有信息量。

1.2.1 2.1 数据集准备

我们以风力发电机裂缝数据为例，假设每组发电机组运行一段时间后记录了裂缝比例。每组的数据权重（即发电机数量）不相同。

首先，生成数据集：

```r
# yi 表示每组风力发电机出现裂缝的比例
yi <- c(0.00, 0.08, 0.06, 0.10, 0.17, 0.23, 0.21, 0.46, 0.65, 0.52, 0.58)

# mi 表示每组发电机数量（作为权重）
mi <- c(39, 53, 33, 73, 30, 39, 42, 13, 34, 40, 36)

# 创建数据框
data <- data.frame(yi, mi)

# 打印数据集
print(data)
```

```
##      yi mi
## 1  0.00 39
## 2  0.08 53
## 3  0.06 33
## 4  0.10 73
## 5  0.17 30
## 6  0.23 39
## 7  0.21 42
```

```
## 8  0.46 13
## 9  0.65 34
## 10 0.52 40
## 11 0.58 36
```

1.3 3. 普通最小二乘法（OLS）回归

1.3.1 3.1 普通最小二乘法的解释

普通最小二乘法（OLS）的基本思想是最小化观测值与模型预测值之间的误差平方和。对于 OLS 回归，假设所有观测点的误差方差相等。

OLS 模型：

$$S(\beta) = \sum_{i=1}^{n} (y_i - x_i^T \beta)^2$$

其中：- y_i：第 i 个观测值。- $x_i^T \beta$：模型预测值。

OLS 估计通过最小化误差平方和 $S(\beta)$ 来估计参数 β。

1.3.2 3.2 OLS 模型的实现

在 R 中，我们可以通过 `lm()` 函数实现普通最小二乘法回归。由于数据集中只有一个常数项（即截距项），我们拟合一个常数模型。

```r
# OLS 模型拟合
ols_model <- lm(yi ~ 1, data = data)

# 输出 OLS 模型结果
summary(ols_model)
```

```
##
## Call:
## lm(formula = yi ~ 1, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.27818 -0.18818 -0.06818  0.21182  0.37182
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.27818    0.06978   3.987  0.00257 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2314 on 10 degrees of freedom
```

1.3.3 3.3 可视化 OLS 拟合结果

我们可以绘制 OLS 模型的拟合结果，将数据点和 OLS 拟合的常数（截距）显示出来：

```
# 绘制 OLS 模型拟合结果
plot(data$yi, main="OLS 模型拟合", xlab=" 样本编号", ylab=" 裂缝比例", pch=19, col="blue")
abline(h=mean(data$yi), col="red", lwd=2, lty=2) # OLS 拟合水平线
legend("topleft", legend=c(" 观测点", "OLS 拟合"), col=c("blue", "red"), pch=c(19, NA), lty=c(NA, 2))
```

1.4 4. 加权最小二乘法（WLS）回归

1.4.1 4.1 加权最小二乘法的解释

加权最小二乘法（WLS）通过为每个观测点分配一个权重 w_i，解决异方差性问题。权重越大，观测点的方差越小，对模型的贡献越大。WLS 的目标是最小化加权的误差平方和：

$$S(\beta) = \sum_{i=1}^{n} w_i (y_i - x_i^T \beta)^2$$

1.4.2 4.2 WLS 模型的实现

在 R 中，我们可以通过 `lm()` 函数中的 `weights` 参数实现加权最小二乘法。我们将发电机数量 m_i 作为每个组的权重，权重较大的观测点将对拟合结果有更大影响。

```
# WLS 模型拟合
wls_model <- lm(yi ~ 1, data = data, weights = mi)

# 输出 WLS 模型结果
summary(wls_model)
```

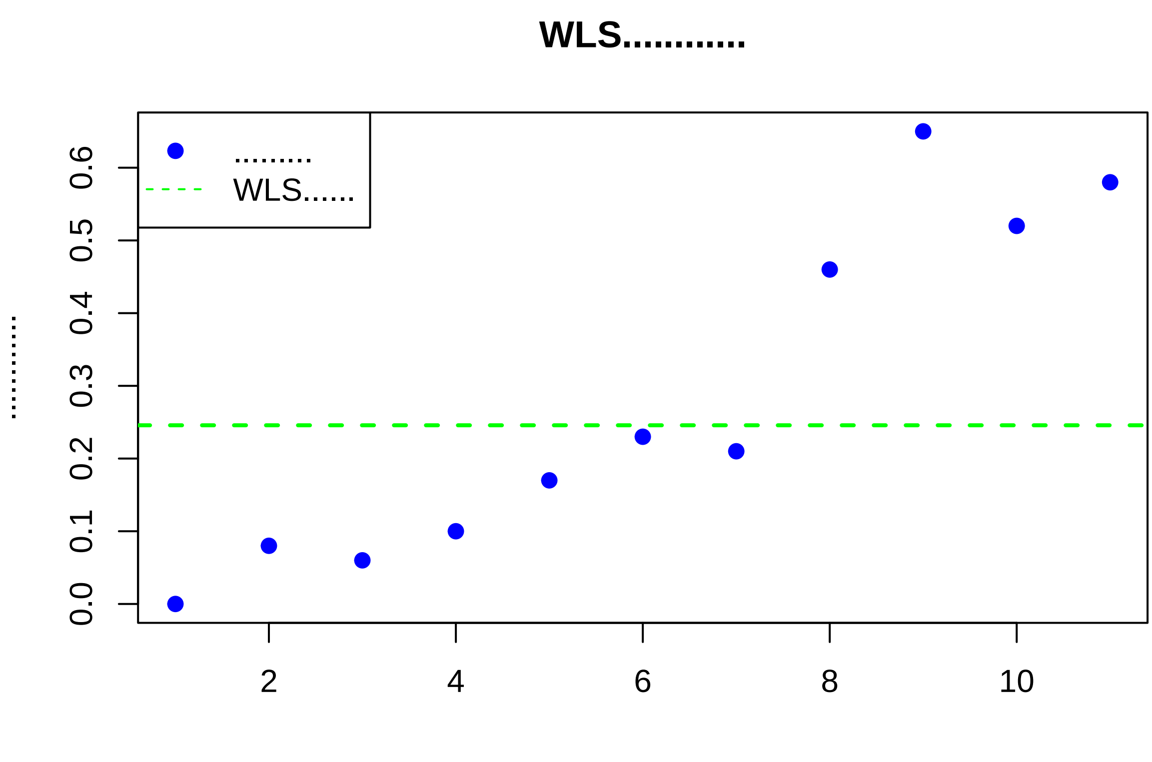
```
##
## Call:
## lm(formula = yi ~ 1, data = data, weights = mi)
```

```
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5348 -1.1370 -0.2318  1.2534  2.3571
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.2458     0.0679   3.619   0.0047 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.411 on 10 degrees of freedom
```

1.4.3 4.3 可视化 WLS 拟合结果

我们同样可以将 WLS 模型的拟合结果进行可视化，观察不同权重对拟合结果的影响。

```r
# 绘制 WLS 模型拟合结果
plot(data$yi, main="WLS 模型拟合", xlab=" 样本编号", ylab=" 裂缝比例", pch=19, col="blue")
wls_fit <- rep(sum(yi * mi) / sum(mi), length(yi)) # 计算加权平均
abline(h=wls_fit[1], col="green", lwd=2, lty=2) # WLS 拟合水平线
legend("topleft", legend=c(" 观测点", "WLS 拟合"), col=c("blue", "green"), pch=c(19, NA), lty=c(NA, 2))
```



1.5 5. 加权残差与模型诊断

1.5.1 5.1 加权残差的解释

加权残差 r_i^* 是每个观测点的实际值与模型预测值之间的差异，残差的大小可以反映模型的拟合效果。公式如下：

$$r_i^* = \sqrt{w_i}(y_i - \hat{\mu})$$

1.5.2 5.2 计算加权残差

在 R 中，我们可以通过 `resid()` 函数提取 WLS 模型的加权残差。

```r
# 提取加权残差
wls_residuals <- resid(wls_model)

# 打印加权残差
print(wls_residuals)
```

```
##           1           2           3           4           5           6           7           8
## -0.24576389 -0.16576389 -0.18576389 -0.14576389 -0.07576389 -0.01576389 -0.03576389  0.21423611  0.4
```

1.5.3 5.3 标准化残差与杠杆值

标准化残差可以将残差标准化，使其易于比较。通过 `hatvalues()` 函数计算杠杆值，可以识别对模型影响较大的观测点。

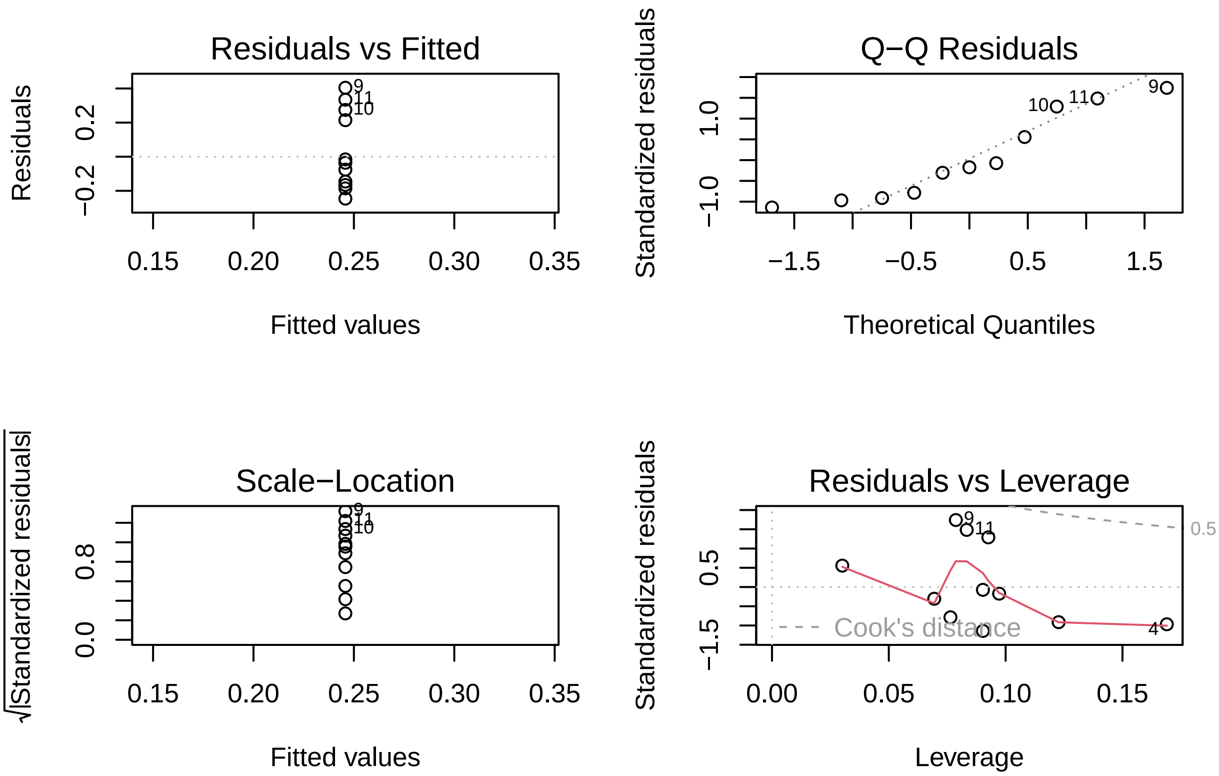
```r
# 计算杠杆值（hat values）和标准化残差
wls_hat_values <- hatvalues(wls_model)
wls_standardized_residuals <- wls_residuals / sqrt(1 - wls_hat_values)

# 打印标准化残差
print(wls_standardized_residuals)
```

```
##           1           2           3           4           5           6           7           8
## -0.25766989 -0.17697511 -0.19329326 -0.15989858 -0.07854004 -0.01652757 -0.03764041  0.21753420  0.4
```

1.5.4 5.4 回归诊断

R 提供了丰富的诊断工具，用于评估模型的拟合质量。通过 `plot()` 函数，我们可以生成回归诊断图。

```r
# 生成回归诊断图
par(mfrow=c(2,2))
plot(wls_model)
```



1.6 6. 帽子矩阵（Hat Matrix）与杠杆值

1.6.1 6.1 帽子矩阵的解释

帽子矩阵 H 是将观测值 Y 投影到模型拟合值 $\hat{Y}$ 上的矩阵。加权线性回归的帽子矩阵定义为：

$$H = X(X^TWX)^{-1}X^TW$$

其中 W 是对角权重矩阵，帽子矩阵的对角元素称为**杠杆值（leverage）**，用于衡量观测点对模型拟合的影响。

1.6.2 6.2 杠杆值的计算

通过 `hatvalues()` 函数，我们可以计算每个观测点的杠杆值，并识别那些对模型拟合影响较大的点。

```r
# 计算杠杆值
leverage <- hatvalues(wls_model)
```

```r
# 打印杠杆值
print(leverage)
```

```
##          1          2          3          4          5          6          7          8          9
## 0.09027778 0.12268519 0.07638889 0.16898148 0.06944444 0.09027778 0.09722222 0.03009259 0.07870370 0
```

1.7 7. 总结

通过本 R 教程，我们详细介绍了普通最小二乘法（OLS）和加权最小二乘法（WLS）的基本概念及其在 R 中的实现。我们展示了如何计算残差、标准化残差和杠杆值，解释了加权线性回归模型中各个观测点对拟合的影响。

加权线性回归是一种强大的工具，特别适用于解决数据中存在异方差性的问题。通过为不同观测点赋予不同权重，WLS 能够提高模型的拟合效果，确保估计的可靠性。