

# 评价方法

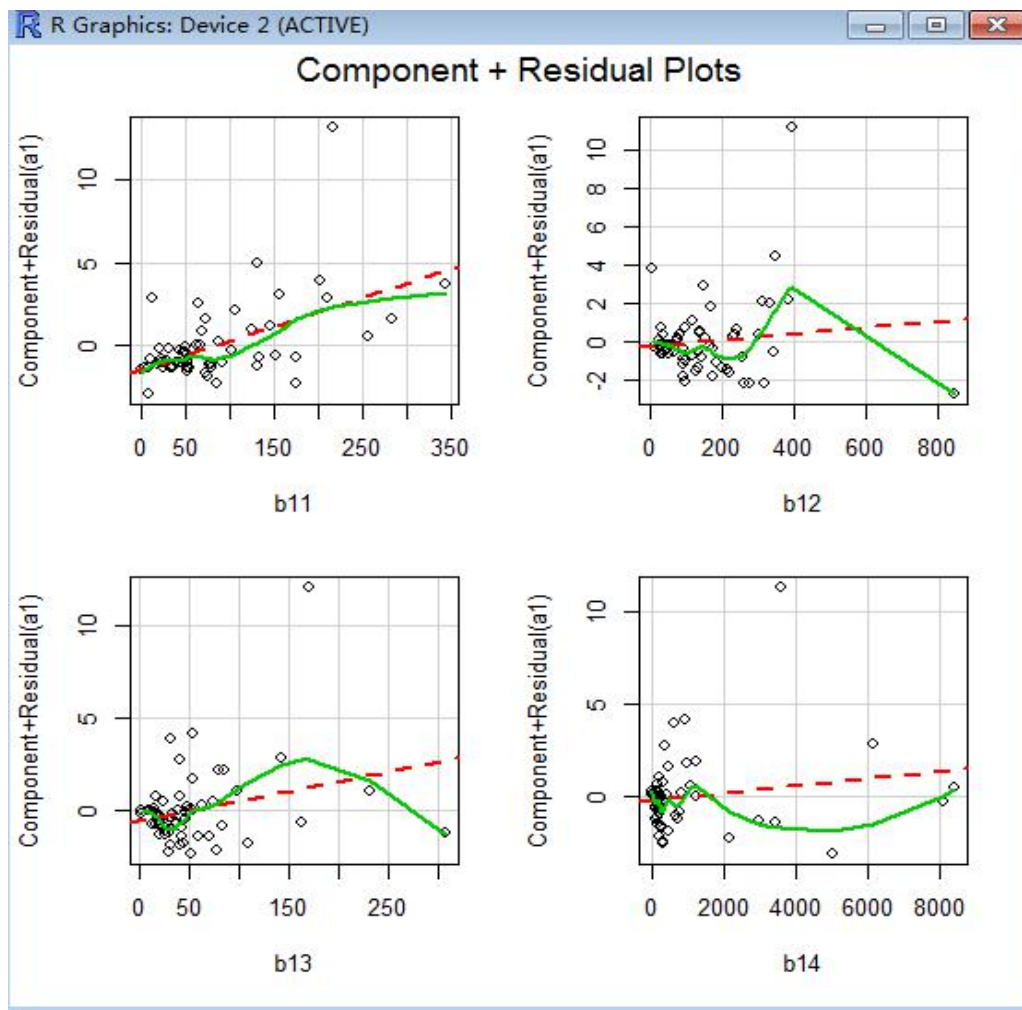
## 一、残差分析

残差= $y - \hat{y}$ ，其中  $y$  是初始值， $\hat{y}$  是计算值。

残差分析 (residual analysis) 就是通过残差所提供的信息，分析出数据的可靠性、周期性或其它干扰。用于分析模型的假定正确与否的方法。

所谓残差是指观测值与预测值 (拟合值) 之间的差，即是实际观察值与回归估计值的差。在回归分析中，测定值与按回归方程预测的值之差，以  $\delta$  表示。**残差  $\delta$  遵从正态分布  $N(0, \sigma^2)$** 。**找到异常值的一个快速方法是使用标准化残差**。若某一实验点的标准化残差落在  $(-2, 2)$  区间以外，可在 95% 置信度将其判为异常实验点，不参与回归线拟合。显然，有多少对数据，就有多少个残差。

- 在 R 中加载 car 包可以使用 crplots() 函数来获得线性模型的**成分残差图** (component plus residual Plots)，来判断拟合模型的线性性
- 看两条线能否接近，如果接近，则说明可以进行线性拟合，即模型是线性的！



## ①直方图

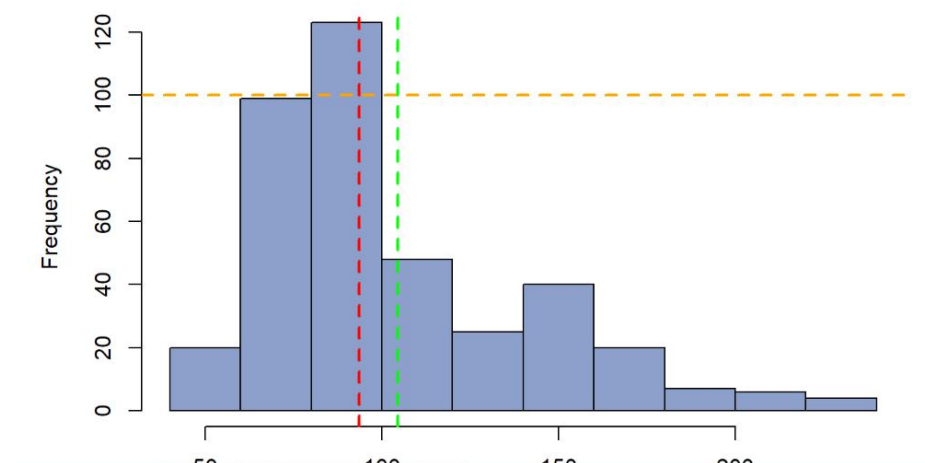
直方图(Histogram)，又称质量分布图，是一种统计报告图。

`hist(mydata$Middle_Price)`

## 4.3 直方图

直方图可以让我们对数据的分布有一个初步的认识。直方图把变量均等的分成若干区间，然后再计算每一区是每一区间内的数据点的频率。

```
> attach(auto)
> hist(horsepower, col = miscolores[3], xlab = "Horsepower", main = "")
> abline(v = median(horsepower), col = "red", lty = 2, lwd = 2)
> abline(v = mean(horsepower), col = "green", lty = 2, lwd = 2)
> abline(h = 100, col = "orange", lty = 2, lwd = 2)
```

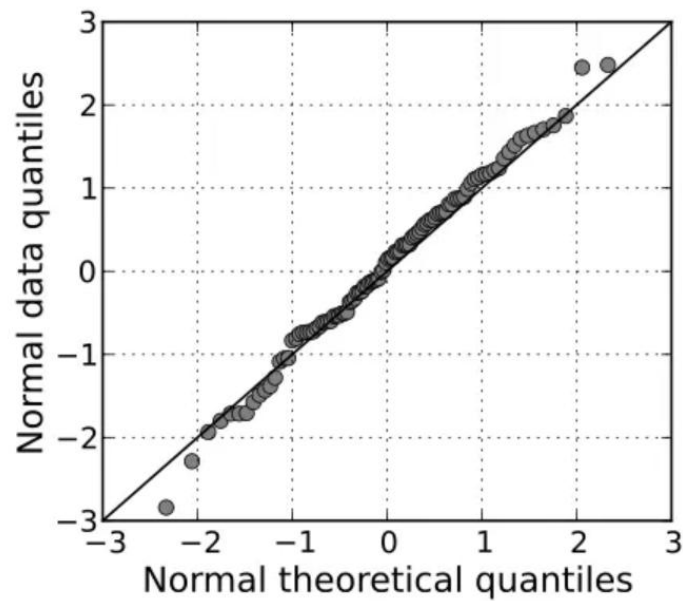


## ②Q-Qplot

QQPlot 图是用于直观验证一组数据是否来自某个分布，或者验证某两组数据是否来自同一（族）分布。在教学和软件中常用的是检验数据是否来自于正态分布。用于验证数据。

若是检验两组数据是否来自同一个分布函数  $F(x)$ ，则直接将两组数据的各自的理论分位点当作横纵坐标，然后看是否在一条直线的附近。此种方法对于两组数据数量不一致的时候，需要用插值法，将数据少的那组数据通过插值的方法补齐。在 R 中 `qqplot` 就在进行这件事情。或者将两个 QQ 图放在一起，将两组数据用不同的颜色标识，看两组数据是否离得很近。

x值是通过分为点区间中间值和反函数求得理论值，y值是实际值。



qq plot

如果两个分布比较相似，那么 qq 图上的点就大致分布在  $y=x$  的线上。如果这两个分布线性相关，那么 qq 图上的点就大致分布在一条线上，但不一定在  $y=x$  上。用 qq 图比较两组数据被认为是一种非参数的方法比较他们的可能分布情况。对于上述这些情况，qq 图是比普通的技术比较柱状图更有说服力的方法，但同时也需要更多地方式来解释。Q-Q 图通常用于将数据集与理论模型进行比较。这可以**提供对图形的“适合度”的评估**，而不是简化为数字摘要。

## 标准化残差-相对于其标准偏差的比例残差

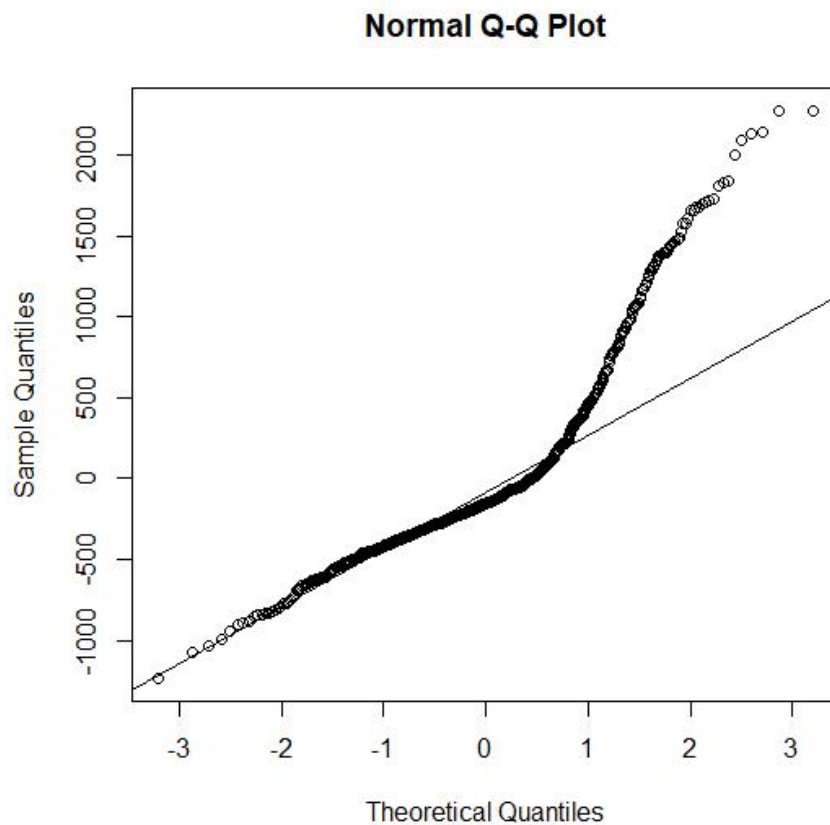
```
residueStandard<-rstandard(lmfit)
```

```
df[residueStandard>3,]
```

- 绘制残差柱状图；
- 添加一个分位数图，其中有一条线穿过，即第一个和第三个分位数。

```
hist(lmfit$residuals)
```

```
qqnorm(lmfit$residuals);qqline(lmfit$residuals)
```



图偏离了正常值（正常值用直线表示）。

但这种非黑即白的信息一般是不够的。因此，我们应该检查**偏态**和**峰度**，以了解分布的分散性。

首先，我们将计算偏态；我们将使用一个简单的高尔顿偏态（Galton's skewness）公式。

残差是否相互关联。

对于一个模型来说，为了解释观测值的所有变化，残差必须随机发生，并且彼此不相关。

**Durbin-Watson** 测试允许检验残差彼此之间的独立性。

```
dwt(lm(df$days.windspeed~df$days.atemp+df$days.hum),method='resample', alternative='two.sided')
```

```
lag Autocorrelation D-W Statistic p-value
```

```
1 0.2616792 1.475681 0
```

```
Alternative hypothesis: rho != 0
```

**独家 | 手把手教你用 R 语言做回归后的残差分析（附代码）**

<https://blog.csdn.net/tMb8Z9Vdm66wH68VX1/article/details/89702486>

## 二、误差指标：MSE，MAPE，MAD

MSE：均方误差

MAPE(mean absolute percentage error) 平均百分比误差

$$MAPE = \sum_{t=1}^n \left| \frac{observed_t - predicted_t}{observed_t} \right| \times \frac{100}{n}$$

mape 取决于数据中数的大小，比如你有两个数据，一个是 100， 一个是 1，你的预测分别是 101 和 2， 误差都是 1，但是 mape 却一个大一个小。

- ❖ mape 只能用来进行对对不同模型同一组数据的评估，比如，对同一组数据，模型 a 给出 mape 比模型 b 给出的 mape 小，这样的话结论是模型 a 会比较好。但是如果我只说 mape=10%，是不能判断这个模型好还是不好的。

MAD 没找到？

## 三、模型稳定性分析

1) Bootstrapping 验证系数分布情况

- ❖ Bootstrapping 算法，指的就是利用有限的样本资料经由多次重复抽样，重新建立起足以代表母体样本分布的新样本。bootstrapping 的运用基于很多统计学假设，因此采样的准确性会影响假设的成立与否。
- ❖ Bootstrapping 从字面意思翻译是拔靴法，从其内容翻译又叫自助法，是一种再抽样的统计方法。一种新的增广样本的统计方法，就是 Bootstrap 方法，为解决小子样试验评估问题提供了很好的思路。
- ❖ 如果不知道总体分布，那么，对总体分布的最好猜测便是由数据提供的分布。自助法的要点是：①假定观察值便是总体；②由这一假定的总体抽取样本，即再抽样。由原始数据经过再抽样所获得的与原始数据集含量相等的样本称为再抽样样本(resamples)或自助样本(bootstrapsamples)。