L13-回归模型

郑盼盼

2024-12-25

目录

13.1 函数模型与回归模型

函数模型

$$y = f(x \,|\, \theta)$$

函数模型是简述因变量 y 和自变量 x 之间关系的一种模型，其中 θ 是模型参数，当确定模型参数后，这种模型的 x 能够唯一确定因变量 y。

回归模型

$$\begin{cases} Y = f(x \,|\, \theta) + \varepsilon \\ \mathbb{E}(\varepsilon) = 0, D(\varepsilon) = \sigma^2 \end{cases}$$

是描述**响应变量** Y 和**解释变量** x 之间关系的另一种模型，其中 θ 是模型参数，当确定模型参数后，这种模型的响应变量只能用 $f(x \,|\, \theta)$ 所近似。

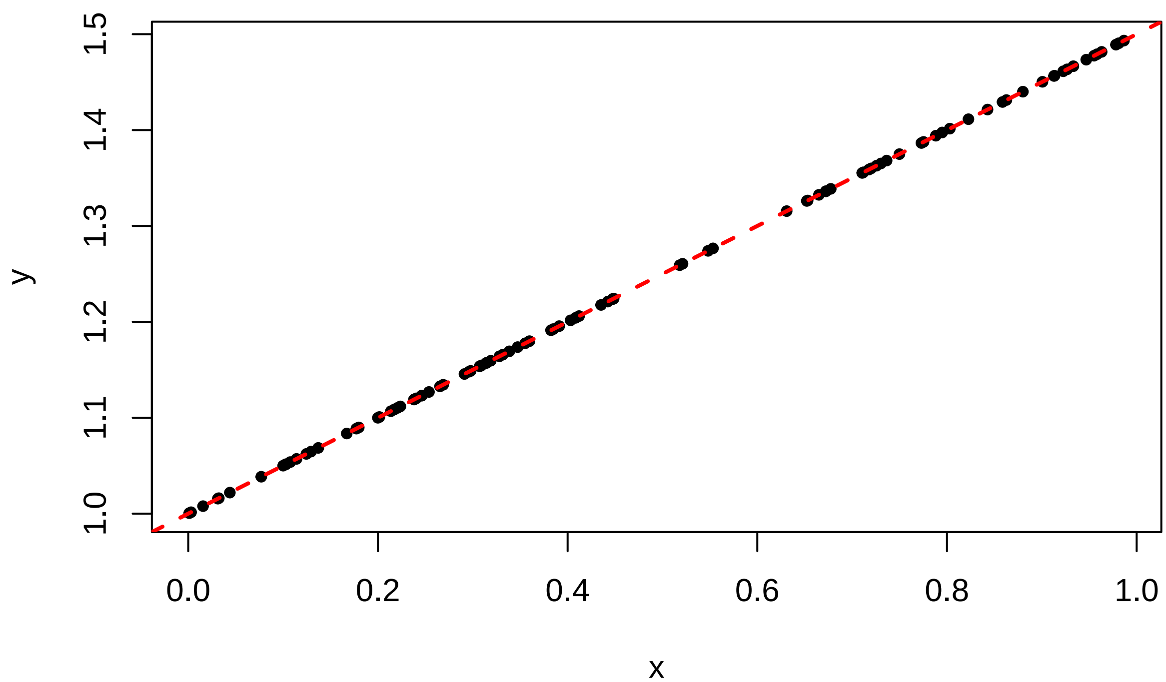
通过散点图判断是哪种模型

通常可以通过样本观测数据 (x_i, y_i) 的散点图来判断数据来自哪种模型：函数模型的样本点都在函数曲线上；回归模型的样本点分布在回归曲线周围。

函数模型

$$y = 0.5x + 1$$

```r
a <- 1; b <- 0.5
x <- runif(100)
y <- a + b*x
plot(x, y, type="p", pch=20)
abline(a,b, lwd=2, lty=2, col="red")
```

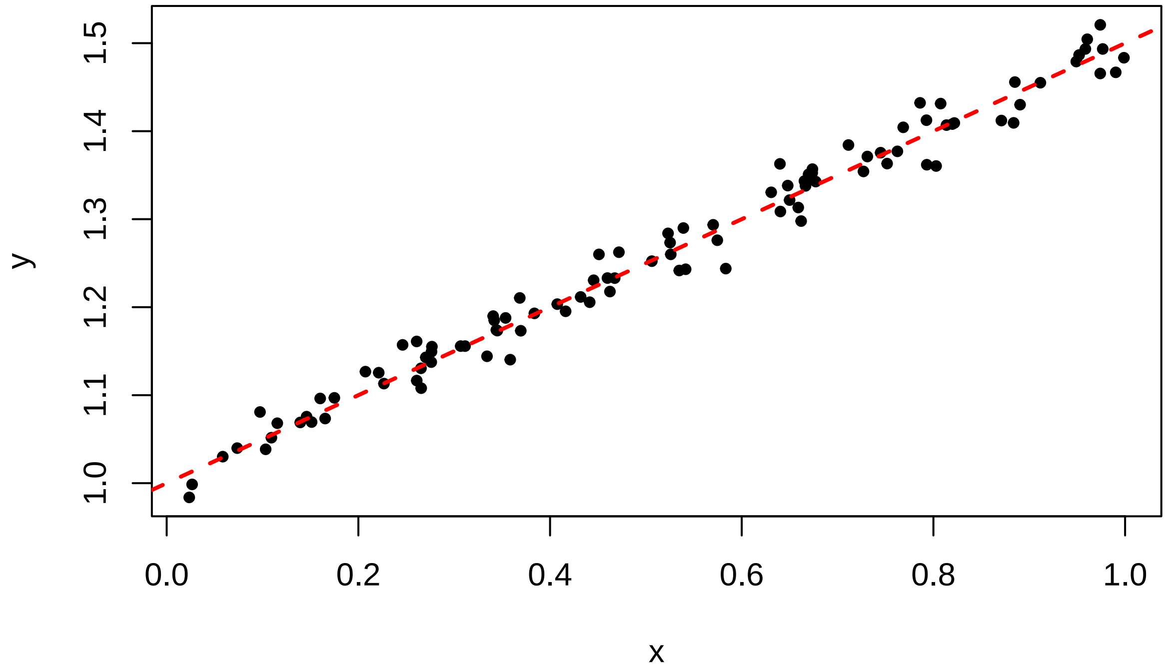


回归模型

$$\begin{cases} y = 0.5x + 1 + \varepsilon \\ \mathbb{E}(\varepsilon) = 0,\ D(\varepsilon) = \sigma^2 \end{cases}$$

```r
a <- 1; b <- 0.5
x <- runif(100)
y <- a + b*x +
    rnorm(100, 0, 0.02)
```

```
plot(x, y, type="p", pch=20)
abline(a,b, lwd=2, lty=2, col="red")
```



13.2 线性回归模型的参数估计

对于身高 h 和体重 w，我们可以使用如下的线性模型进行拟合

$$W = a + bh + \varepsilon$$

13.2.1 模型拟合

- `lm(formula, data=)` 使用 R 语言建立线性模型：
  - `formula` 的形式为 响应变量 ~ 解释变量，即 ~ 左侧为因变量，右侧为自变量，进行回归。
- 建立模型后（`model <- lm()`）
  1. 我们可以用 `summary(model)` 的方式来查看线性回归模型拟合的参数结果。可用 `summary(model)$r.squared` 的方式获得模型的 R^2
  2. 可直接用 `coef(model)` 来查看各参数的估计值
  3. 利用 `model$residuals` 来查看所有的残差，并可以使用 `sum(model$residuals ^ 2)` 来计算残差平方和。
  4. 利用 `predict(model)` 可直接计算得到回归模型估计值 $\hat{Y} = f(x|\hat{\theta})$
- `poly(x, m, raw=TRUE)` 用于 m 阶多项式拟合：即拟合 $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m$

```
h <- runif(1000, min = 160, max = 195) # 模拟身高
w <- (h-100) * 0.9 + rnorm(1000, 0, 3) # 模拟体重
tmpData <- data.frame(h = h, w = w)
tmpLm <- lm(w ~ h, data=tmpData) # 使用 lm 进行线性模型的回归拟合
summary(tmpLm) # 总结模型信息
```

```
##
## Call:
## lm(formula = w ~ h, data = tmpData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7131 -2.2007 -0.0471  2.2180 10.9910
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -90.396293   1.728495  -52.30   <2e-16 ***
## h             0.901819   0.009723   92.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.102 on 998 degrees of freedom
## Multiple R-squared:  0.8961, Adjusted R-squared:  0.8959
## F-statistic:  8603 on 1 and 998 DF,  p-value: < 2.2e-16
```

此外，我们可以使用下面的模型进行拟合（a, b 和 c 为模型的参数）

$$W = a + bh + ch^2 + \varepsilon$$

```
tmpLm2 <- lm(w ~ h + I(h^2), data=tmpData) # 使用 lm 进行线性模型的回归拟合
summary(tmpLm2) # 总结模型信息
```

```
##
## Call:
## lm(formula = w ~ h + I(h^2), data = tmpData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6652 -2.1756 -0.0852  2.2229 10.8440
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.359e+02  3.364e+01  -4.040 5.76e-05 ***
## h            1.416e+00  3.795e-01   3.730 0.000202 ***
## I(h^2)      -1.446e-03  1.068e-03  -1.355 0.175856
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.101 on 997 degrees of freedom
## Multiple R-squared:  0.8962, Adjusted R-squared:  0.896
## F-statistic:  4306 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
RSS1 = sum(tmpLm$residuals ^ 2)
RSS2 = sum(tmpLm2$residuals ^ 2)
cat(" 模型 1 的残差为 ", RSS1, "\n模型 2 的残差为 ", RSS2)
```

```
## 模型1的残差为  9602.291
## 模型2的残差为  9584.651
```

13.2.2 AIC 准则

随着模型变得复杂，对于观测值的拟合效果会越来越好。但残差平方和过小会导致过拟合的出现，使得模型的泛化能力变差；因此，我们可以根据 AIC 准则挑选合适的模型：挑选使得

$$\text{AIC} = n \log\left(Q(\hat{\theta})\right) + 2k,$$

最小的模型，其中

- k 为模型参数的个数
- n 为样本量

```
1000 * log(sum(tmpLm$residuals^2)) - 2 * length(tmpLm$coefficients)
```

```
## [1] 9165.757
```

```
1000 * log(sum(tmpLm2$residuals^2)) - 2 * length(tmpLm2$coefficients)
```

```
## [1] 9161.918
```

13.2.3 多项式拟合例

已有观测数据 x 和 Y 尝试将 Y 作为响应变量，x 作为解释变量，进行线性回归

```
# 已有观测数据 x 和 Y
x<-seq(0.20,6.00,by=0.2)
Y<-c(-1.56,5.33,0.54,6.99,0.62,4.66,7.87,13.26,12.82,10.56,
     7.66,13.85,21.94,18.53,28.46,36.26,35.90,39.70,
     45.35,54.08,60.16,52.95,64.51,72.06,73.68,93.60,
     91.76,89.83,104.98,111.50)
myData <- data.frame(x = x, Y = Y)
```

- 模型 1：线性回归模型

$$Y = a + bx$$

```
myS1<-lm(Y~x, data=myData) # 拟合一元线性模型
coef(myS1) # 输出参数估计
```

```
## (Intercept)           x
##   -19.80547    19.05392
```

- 模型 2：没有常数项的二阶多项式回归模型

$$Y = bx + cx^2$$

```
myS2<-lm(Y~-1+poly(x,2,raw=T), data=myData)# 拟合没有截距项的二阶曲线
coef(myS2) # 输出参数估计
```

```
## poly(x, 2, raw = T)1 poly(x, 2, raw = T)2
##            0.1038142            3.0784018
```

- 模型 3：仅有二次项的回归模型

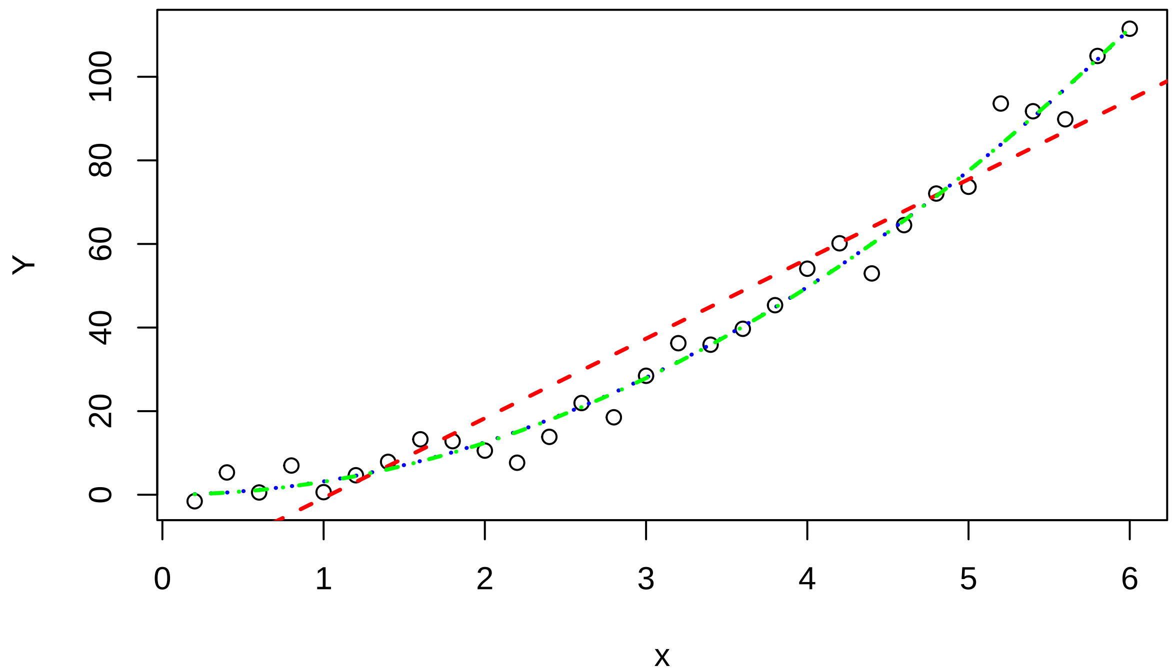
$$Y = cx^2$$

```
myS3<-lm(Y~0+I(x^2))# 拟合仅包含二次项的模型
coef(myS3) # 输出参数估计
```

```
##   I(x^2)
## 3.099683
```

绘制图像，观察三种模型的拟合效果：

```
plot(x, Y, type="p")
abline(myS1, col="red", lty=2, lwd=2)
lines(x,predict(myS2), col="blue", lty=3, lwd=2)
lines(x,predict(myS3), col="green", lty=4, lwd=2)
```




```
myS1<-lm(Y~x)    # 拟合线性模型
coef(myS1)       # 输出参数估计
```

```
## (Intercept)           x
##   -19.80547    19.05392
```

```
sum((myS1$residuals)^2) # 计算残差平方和
```

```
## [1] 2688.365
```

```
summary(myS1)$r.squared # 获得模型 1 的 R 方
```

```
## [1] 0.9238999
```

```
cat(" 模型拟合参数:",coef(myS1),"\n",
    " 拟合残差平方和:",sum((myS1$residuals)^2),"\n",
    "R 方:",summary(myS1)$r.squared)
```

```
## 模型拟合参数: -19.80547 19.05392
```

```
##  拟合残差平方和: 2688.365
##  R方: 0.9238999
```