统计学导论：R语言实验12

回归模型

主讲人：郑盼盼

12.0 Outline

1. 回归模型和函数模型
2. 线性回归模型的参数估计

12.1 函数模型和回归模型

12.1.1 函数模型

$$y = f(x \mid \theta)$$

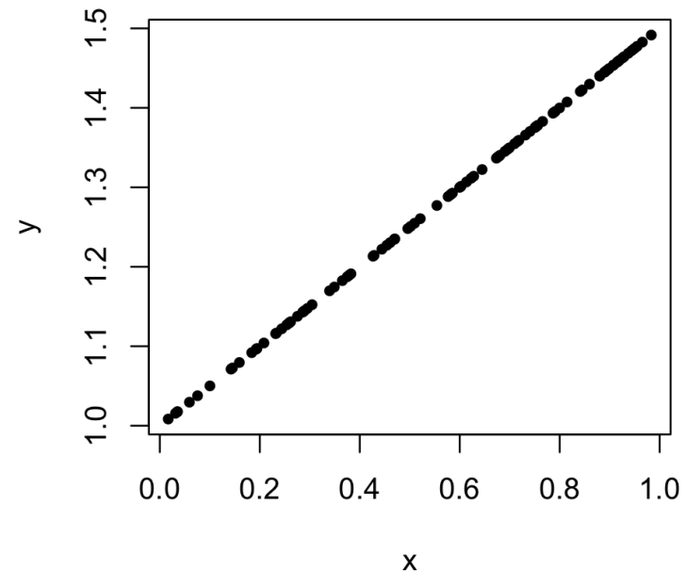
函数模型是简述因变量 y 和自变量 x 之间关系的一种模型，其中 θ 是模型参数，当确定模型参数后，这种模型的 x 能够唯一确定因变量 y。

12.1.1 函数模型

$$y = f(x \mid \theta)$$

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。比如

$$y = 0.5x + 1$$

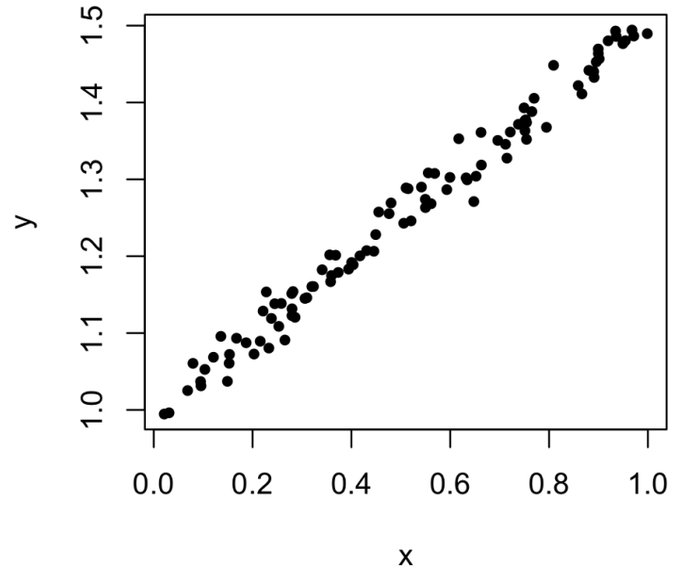


12.1.2 回归模型

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

回归模型是描述**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 $f\left(x \mid \theta\right)$ 所近似。

$$Y = 0.5x + 1 + \varepsilon$$



12.1.3 模型图像绘制

当我们拥有数据点 (x_i, y_i) 后，我们可以绘制这些数据点的散点图，可根据散点图判断模型的类型：

```r
a <- 1; b <- 0.5
x <- runif(100) # 生成[0,1]内服从均匀分布的随机数
y <- a + b*x    # 生成函数模型的数据点
Y <- a + b*x + rnorm(100, 0, 0.02) # 生成回归模型的数据点
plot(x, y, type="p", pch=20)
plot(x,Y, type="p", pch=20)
```

12.2 线性回归模型的参数估计

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

- 函数 `lm()`

```r
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
```

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

1. 我们所执行的拟合数据的代码

```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
   Min     1Q Median     3Q    Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

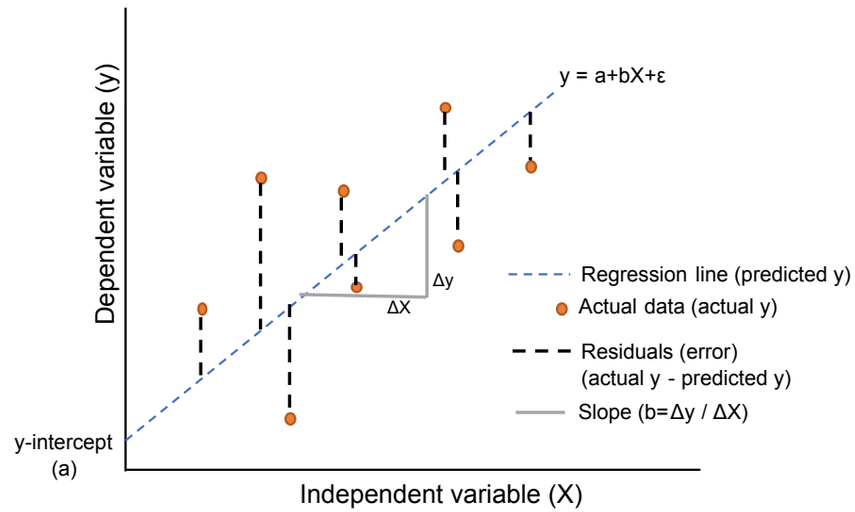
```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
    Min      1Q  Median      3Q     Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

1. 我们所执行的拟合数据的代码
2. 残差的五数概括



12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
   Min     1Q Median     3Q    Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

1. 我们所执行的拟合数据的代码
2. 残差的五数概括
3. 模型拟合结果：`Estimate` 这一列即模型参数 b, a 的估计值；后面三列用于检验原假设 $H_0: a = 0$ 和 $H_0: b = 0$ 是否成立。

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
   Min     1Q Median     3Q    Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

1. 我们所执行的拟合数据的代码
2. 残差的五数概括
3. 模型拟合结果
4. ε 的标准差估计

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
   Min     1Q Median     3Q    Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

1. 我们所执行的拟合数据的代码
2. 残差的五数概括
3. 模型拟合结果
4. ε 的标准差估计
5. 对于 R^2 的估计，其描述响应变量的变异在多大程度上能由自变量所解释。

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们拟用如下的线性模型来进行拟合

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

```
Call:
lm(formula = w ~ h, data = tmpData)

Residuals:
   Min     1Q Median     3Q    Max
-8.887 -2.056  0.007  2.164 10.143

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.747907   1.691185  -51.88   <2e-16 ***
h             0.887884   0.009532   93.15   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.038 on 998 degrees of freedom
Multiple R-squared:  0.8968,    Adjusted R-squared:  0.8967
F-statistic:  8677 on 1 and 998 DF,  p-value: < 2.2e-16
```

1. 我们所执行的拟合数据的代码
2. 残差的五数概括
3. 模型拟合结果
4. ε 的标准差估计
5. 对于 R^2 的估计
6. 对于整个模型的检验：

H_0：响应变量和解释变量无关系

12.2.1 线性回归模型的参数估计

例 我们获得了 1000 个人的身高 h 及其体重 W，我们将之前的线性模型改成一个更为复杂的模型

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

我们可以使用如下的R语言代码进行拟合：

```r
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
```

12.2.1 线性回归模型的参数估计

我们可以对比两个模型的残差平方和：

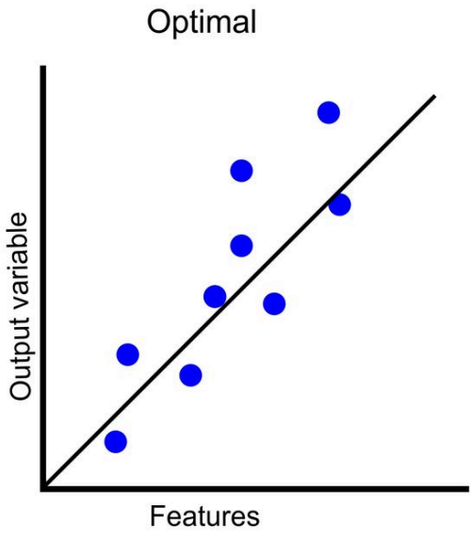
$$\mathrm{RSS} = \sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2$$

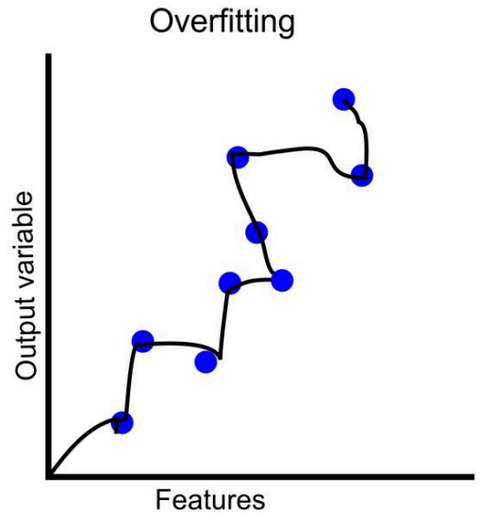
我们可以使用如下的R语言代码分别计算两个模型的残差平方和：

```r
sum(tmpLm$residuals ** 2)
sum(tmpLm2$residuals ** 2)
```

12.2.1 线性回归模型的参数估计

我们可以对比两个模型的残差平方和，通过计算，我们可以发现第二个模型的残差平方和小于第一个模型，但是它真的更好吗?





12.2.2 AIC准则

为了得到一个相对合适的拟合效果，前人提出了AIC准则，即挑选使得

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

最小的模型，其中

- k 为模型参数的个数
- n 为样本量
- $Q(\hat{\theta})$ 为残差平方和

我们可以用R语言计算两个模型的AIC

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