第九章 一元线性回归

9.1 回归函数的简介

回归函数是统计学里面很重要的一种函数。我们什么时候用一元线性回归呢？通常我们一个指标是受很多因素的影响的，但是如果该指标的主要影响因素只有一个的时候，我们就可以用到一元线性回归来分析。

我们一元线性回归模型的标准表达式是：

$$y=ax+b$$

x 是自变量，y 是因变量。其中的 a,b 是回归模型的参数。

下面的小节，我们将介绍回归模型的估计技术。

9.2 最小二乘法原理

我们先来介绍一下最小二乘法。最小二乘法是一种数学优化方法，其目的是在函数与一系列观测值（数据）之间建立一个关系式。最小二乘法的原理是通过寻找一条直线（或曲线），使得所有数据点到这条直线（或曲线）的距离之和最小。这里的距离是指数据点到直线（或曲线）的垂直距离。

在使用最小二乘法时，需要先选定一个数学模型，通常是线性模型。然后，利用数据对模型中的参数进行估计，使得模型能够较好地拟合数据。拟合模型时，通常会采用误差的平方和作为损失函数，然后利用最小化损失函数的方法求解参数的最优解，从而得到最终的拟合结果。

最小二乘法在科学和工程领域中广泛应用，例如用于拟合数据、曲线拟合、回归分析、信号处理、计量经济学等领域。

我们平常的一元线性回归方程，通常可以表示为

$$y_i = \hat{w}_0 + \hat{w}_1 x_i + e_i$$

我们在统计学中，最小二乘法的表达式通常为

$$\min \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - \hat{w}_0 - \hat{w}_1 x_i)^2$$

上述方程中的 $\widehat{w_0}$ 和 $\widehat{w_1}$ 的求解表达式为

$$\hat{w}_0 = \bar{y} - \hat{w}_1 \bar{x}, \quad \hat{w}_1 = \frac{\sum x'_i y'_i}{\sum x'^2_i}$$

9.3 一元线性回归在统计学中的应用

这一小节我们会介绍统计学中一元线性回归的一些案例。我们以树的高度以及树的年龄为例子。我们总共调查了 24 棵树为样本，下面图 1 为这 24 棵树的数据

树龄/年 x_i	树高/m y_{ij}	x_i^2	$x_i y_{ij}$	树龄/年 x_i	树高/m y_{ij}	x_i^2	$x_i y_{ij}$
2	5.6	4	11.2	5	7.1	25	35.5
2	4.8	4	9.6	5	7.3	25	36.5
2	5.3	4	10.6	5	6.9	25	34.5
2	5.7	4	11.4	5	6.9	25	34.5
3	6.2	9	18.6	6	7.2	36	43.2
3	5.9	9	17.7	6	7.5	36	45.0
3	6.4	9	19.2	6	7.8	36	46.8
3	6.1	9	18.3	6	7.8	36	46.8
4	6.2	16	24.8	7	8.9	49	62.3
4	6.7	16	26.8	7	9.2	49	64.4
4	6.4	16	25.6	7	8.5	49	59.5
4	6.7	16	26.8	7	8.7	49	60.9

图 1

上面的表格计算我们可以用 R 语言来进行一元线性回归的计算，这样就给我们节省了很多计算时间。我们计算用的代码以及结果入下图：

```
> plants<-data.frame(age=rep(2:7,rep(4,6)),height=c(5.6,4.8,5.3,5.7,6.2,5.9,6.4,6.1,6.2,6.7,6.4,6.7,7.1,7.3,6.9,6.9,7.2,7.5,7.8,7.8,8.9,9.2,8.5,8.7))
> View(plants)
```

```
> plants.lm<-lm(height~age,data = plants)
> summary(plants.lm)
Call:
lm(formula = height ~ age, data = plants)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.65976 -0.22476 -0.00833  0.21524  0.70595 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  4.05405    0.19378   20.92 5.19e-16 ***
age          0.63429    0.04026   15.76 1.82e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
 1

Residual standard error: 0.3368 on 22 degrees of freedom
Multiple R-squared:  0.9186,	Adjusted R-squared:  0.9149 
F-statistic: 248.2 on 1 and 22 DF,  p-value: 1.821e-13
```

我们用 rstudio 来进行一元线性回归的模型计算节省了大量时间。

然后我们再进行图像输出就可以得出下图，如图 2

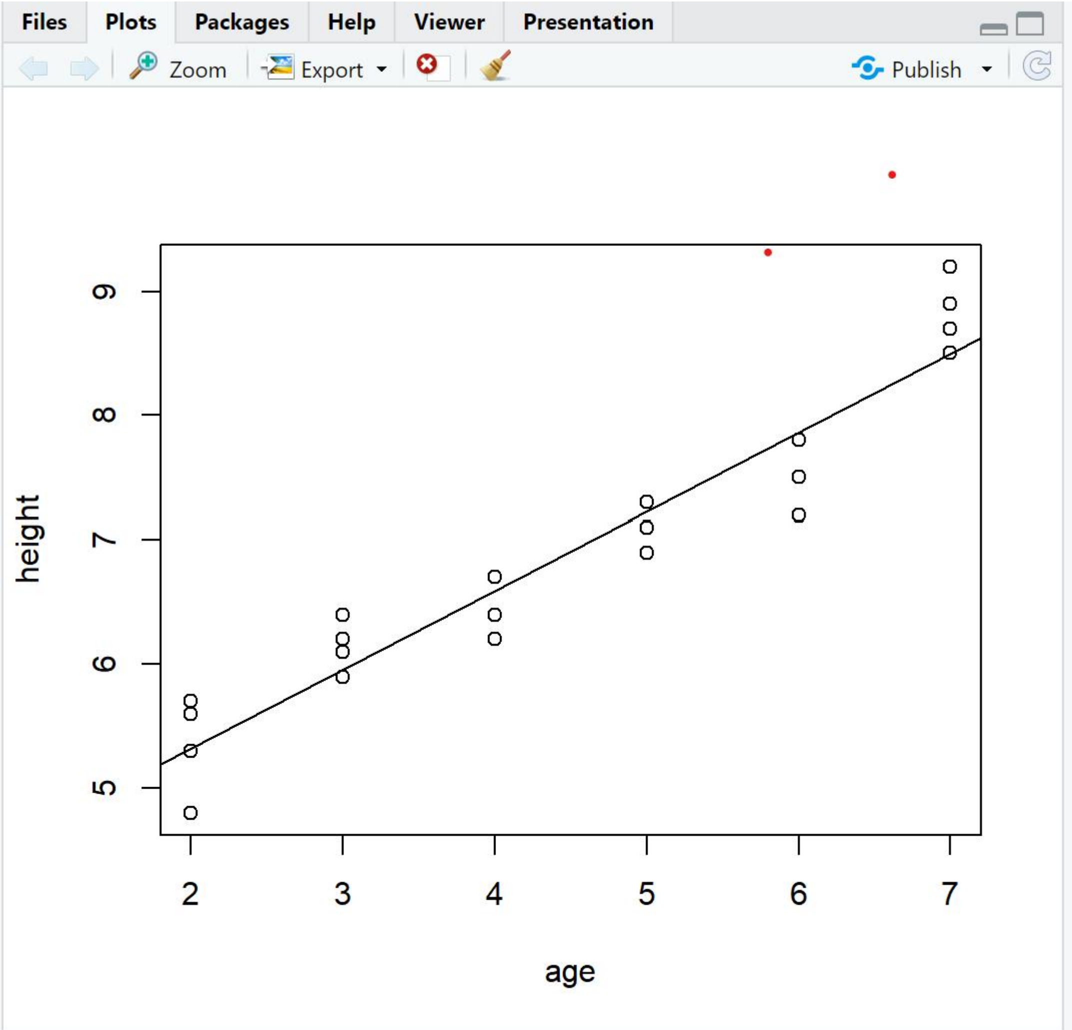

Files
Plots
Packages
Help
Viewer
Presentation
Zoom
Export
Publish
height
9
8
7
6
5
2
3
4
5
6
7
age

图 2

9.4 总体方差无偏估计的方法

首先我们引入一个新概念：总体方差。总体方差是指在统计学中，用来衡量一组数据的离散程度或者分散程度的一种方法。它是各个数据与其平均值偏差平方和的平均值。总体方差表示了数据之间的差异性或者波动程度，数值越大，代表数据之间的差异性越大，反之则越小。总体方差通常用符号 σ^2 表示，其中 σ 是总体标准差。

总体方差的无偏估计量是样本方差。下面我们来证明总体方差的无偏估计量

证明过程如下：

设总体的方差为 σ^2，样本的大小为 n，样本的观测值为 $X_1, X_2, ..., X_n$ ，样本均值为 x ，则样本方差 S^2 定义为：

$S^2 = \Sigma(X_i - x)^2 / (n-1)$

我们需要证明 $E(S^2) = \sigma^2$。

根据期望的线性性质和方差公式有：

$$
\begin{aligned}
E(S^2) &= E[\Sigma(X_i - x)^2 / (n-1)] \\
&= (1 / (n-1)) E[\Sigma(X_i^2 - 2X_i x + x^2)] \\
&= (1 / (n-1)) [\Sigma E(X_i^2) - 2E(x)\Sigma X_i + nx^2] \\
&= (1 / (n-1)) [n\sigma^2 - 2nx^2 + nx^2] \\
&= ((n-1)/ (n-1)) \sigma^2 \\
&= \sigma^2
\end{aligned}
$$

故样本方差 S^2 是总体方差 σ^2 的无偏估计量。

9.5 估计参数的概率分布

估计参数的概率分布通常采用贝叶斯方法。如果数据集为D，模型的参数为θ，则根据贝叶斯定理，参数的后验概率分布可以表示为：

$$P(\theta|D) = P(D|\theta)P(\theta) / P(D)$$

其中，P(D|θ)是给定参数θ下数据集D的似然函数，表示为经验概率密度函数；P(θ)是参数的先验概率分布，表示为先验概率密度函数，它体现了观测之前的先验知识；P(D)是数据集D的边际概率分布，是一个标准化因子，它使得后验概率概率分布满足归一性。

对于参数的估计，通常采用后验分布的期望或者中位数作为参数的最佳估计。还可以用后验分布的方差、置信区间等指标来对参数的不确定性进行描述。

我们在代码中可以先用summary语句来获取我们数据中的统计量。然后我们再介绍从confint函数，这个函数我们通常在R语言里面用来计算出线性回归模型的置信区间。

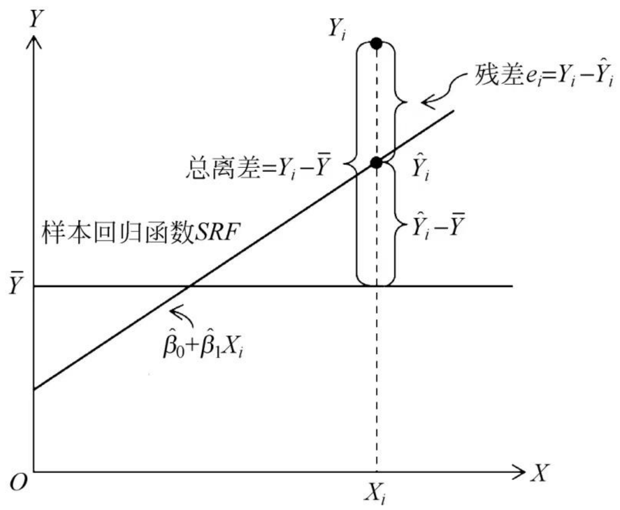
标准格式为confint(我们要添加的表格.lm)。

9.6 拟合优度的检验

拟合优度检验是统计学中经常用来检验模型拟合程度的方法。它可以用来验证模型是否能够很好地拟合现有数据，并且向未来数据提供准确的预测。

在进行拟合优度检验时，我们通常会采用残差平方和（RSS）和总平方和（TSS）之比来计算拟合优度，也就是R^2值。R^2值的范围在0到1之间，结果越接近1，表示模型的拟合优度越好，反之则拟合程度越差。

我们来看一张图



上图展示的是残差各个变量的关系。

在此模型里，s 值（残差）我们可以得出

$$s=\hat{\sigma}=\sqrt{\frac{\sum(y_i-\hat{y}_i)^2}{n-2}}$$

进而再进行多种演算我们可以得到

$$R_{\text{adj}}^2=R^2-\frac{p-1}{n-p}(1-R^2)$$

9.7 单个参数的检验

单个参数的检验是指在统计推断中，对于一个特定的参数进行的假设检验，例如平均值、方差、比例等。

对于一个单个参数进行检验的假设检验，一般使用 t 分布、卡方分布或者 F 分布来进行检验。其中 t 分布主要用于样本均值相对于总体均值的检验，卡方分布主要用于方差或标准差的检验，F 分布则用于两个方差或标准差的比较。

具体的检验步骤大致如下：

1. 确定假设检验的原假设和备择假设。
2. 确定显著性水平，即确定用于比较的临界值。
3. 根据样本数据计算统计量的值。
4. 根据显著性水平和自由度推算得到比较的临界值。

5. 比较统计量和临界值的大小，进行假设检验的决策。

若统计量的值落在拒绝域内，则拒绝原假设，否则接受原假设。