第5章 一元线性回归

CHAPTER 5

线性回归是统计分析中最常被用到的一种技术。在其他的领域，例如机器学习理论和计量经济研究中，回归分析也是不可或缺的重要组成部分。本章将要介绍的一元线性回归是最简单的一种回归分析方法，其中所讨论的诸多基本概念在后续更为复杂的回归分析中也将被常常用到。

5.1 回归分析的性质

回归一词最早由英国科学家弗朗西斯·高尔顿(Francis Galton)提出，他还是著名生物学家、进化论奠基人查尔斯·达尔文(Charles Darwin)的表弟。高尔顿深受进化论思想的影响，并把该思想引入到人类研究，从遗传的角度解释个体差异形成的原因。高尔顿发现，虽然存在一个趋势——父母高，儿女也高；父母矮，儿女也矮。但给定父母的身高，儿女辈的平均身高却趋向于或者“回归”到全体人口的平均身高。换句话说，即使父母双方都异常高或者异常矮，儿女的身高还是会趋向于人口总体的平均身高。这也就是所谓的普遍回归规律。高尔顿的这一结论被他的朋友，数学家、数理统计学的创立者卡尔·皮尔逊(Karl Pearson)所证实。皮尔逊收集了一些家庭的1000多名成员的身高记录，发现对于一个父亲高的群体，儿辈的平均身高低于他们父辈的身高；而对于一个父亲矮的群体，儿辈的平均身高则高于其父辈的身高。这样就把高的和矮的儿辈一同“回归”到所有男子的平均身高，用高尔顿的话说，这是“回归到中等”。

回归分析是被用来研究一个被解释变量(Explained Variable)与一个或多个解释变量(Explanatory Variable)之间关系的统计技术。被解释变量有时也被称为因变量(Dependent Variable)，与之相对应地，解释变量也被称为自变量(Independent Variable)。回归分析的意义在于通过重复抽样获得的解释变量的已知或设定值来估计或者预测被解释变量的总体均值。

在高尔顿的普遍回归规律研究中，他的主要兴趣在于发现为什么人口的身高分布存在有一种稳定性。现在关心的是，在给定父辈身高的条件下，找出儿辈平均身高的变化规律。也就是一旦知道了父辈的身高，怎样预测儿辈的平均身高。图5-1展示了对应于设定的父亲身高，儿子在一个假想人口总体中的身高分布情况。不难发现，对于任一给定的父亲身高，我们都能从图中确定出儿子身高的一个分布范围，同时随着父亲身高的增加，儿子的平

均身高也会增加。为了更加清晰地表示这种关系，在散点图上勾画了一条描述这些数据点分布规律的直线，用来表明被解释变量与解释变量之间关系，即儿子的平均身高与父亲身高之间的关系。这条直线就是所谓的回归线，后面还会对此进行详细讨论。

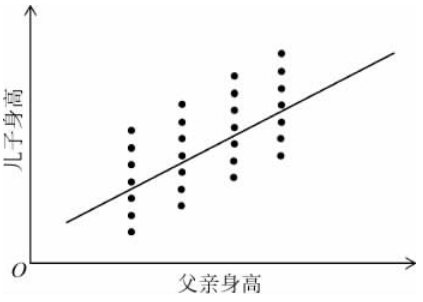

图 5-1　父亲身高与儿子身高的关系

在回归分析中，变量之间的关系与物理学公式中所表现的那种确定性依赖关系不同。回归分析中因变量与自变量之间所呈现出来的是一种统计性依赖关系。在变量之间的统计性依赖关系中，主要研究的是随机变量，也就是有着概率分布的变量。但是函数或确定性依赖关系中所要处理的变量并非是随机的，而是一一对应的关系。例如，粮食产量对气温、降雨和施肥的依赖关系是统计性质的。这个性质的意义在于：这些解释变量固然重要，但并不能据此准确地预测粮食的产量。首先是因为对这些变量的测量有误差，其次是还有很多影响收成的因素，很难一一列举。事实上，无论考虑多少个解释变量都不可能完全解释粮食产量这个因变量，毕竟粮食作物的生长过程是受到许许多多随机因素影响的。

与回归分析有密切关联的另外一种技术是相关分析，但两者在概念上仍然具有很大差别。相关分析是用来测度变量之间线性关联程度的一种分析方法。例如，常常会研究吸烟与肺癌发病率、金融发展与经济增长等之间的关联程度。而在回归分析中，对变量之间的这种关系并不感兴趣，回归分析更多的是通过解释变量的设定值来估计或预测因变量的平均值。

回归与相关在对变量进行分析时是存在很大分歧的。在回归分析中，对因变量和自变量的处理方法上存在着不对称性。此时，因变量被当作统计的、随机的，也就是存在着一个概率分布，而解释变量则被看成是（在重复抽样中）取有规定值的一个变量。因此在图 5-1 中，假定父亲的身高变量是在一定范围内分布的，而儿子的身高却反映在重复抽样后的一个由回归线给出的稳定值。但在相关分析中，将对称地对待任何变量，即因变量和自变量之间不加区别。例如，同样是分析父亲身高与儿子身高之间的相关性，那么这时我们所关注的将不再是由回归线给出的那个稳定值，儿子的身高变量也是在一定范围内分布的。大部分的相关性理论都建立在变量的随机性假设上，而回归理论往往假设解释变量是固定的或非随机的。

虽然回归分析研究是一个变量对另外一个或几个变量的依赖关系，但它并不意味着因果关系。莫里斯·肯达尔（Maurice Kendall）和艾伦·斯图亚蒂（Alan Stuart）曾经指出："一个统计关系式，不管多强也不管多么有启发性，都永远不能确立因果关系的联系；对因果关系的理念必须来自统计学以外，最终来自这种或那种理论。"比如前面谈到的粮食产量的例子中，将粮食产量作为降雨等因素的因变量没有任何统计上的理由，而是出于非统计上的原因。而且常识还告诉我们不能将这种关系倒转，即我们不可能通过改变粮食产量的做法来控制降雨。再比如，古人将月食归因于"天狗吃月"，所以每当发生月食时，人们就会敲锣打鼓意图吓走所谓的天狗。而且这种方法屡试不爽，只要人们敲锣打鼓一会儿，被吃掉的月亮就会恢复原样。显然，敲锣打鼓与月食结束之间有一种统计上的关系。但现代科技告

诉我们月食仅仅是一种自然现象，它与敲锣打鼓之间并没有因果联系，事实上即使人们不敲锣打鼓，被“吃掉”的月亮也会恢复原状。总之，统计关系本身不可能意味着任何因果关系。要谈及因果关系必须进行先验的或理论上的思考。

5.2　回归的基本概念

本节将从构建最简单的回归模型开始，结合具体例子向读者介绍与回归分析相关的一些基本概念。随着学习的深入，我们渐渐会意识到，更为一般的多变量之间的回归分析，在许多方面都是最简情形的逻辑推广。

5.2.1　总体的回归函数

经济学中的需求法则认为，当影响需求的其他变量保持不变时，商品的价格和需求量之间呈反向变动的关系，即价格越低，需求量越多；价格越高，需求量越少。据此，假设总体回归直线是线性的，便可以用下面的模型来描述需求法则

$$E(y \mid x_i) = w_0 + w_1 x_i$$

这是直线的数学表达式，它给出了与具体的 x 值相对应的(或条件的)y 的均值，即 y 的条件期望或条件均值。下标 i 代表第 i 个子总体，读作“在 x 取特定值 x_i 时，y 的期望值”。该式也称为非随机的总体回归方程。

这里需要指出，$E(y|x_i)$是 x_i 的函数，这意味着 y 依赖于 x，也称为 y 对 x 的回归。回归可以简单地定义为在给定 x 值的条件下 y 值分布的均值，即总体回归直线经过 y 的条件期望值，而上式就是总体回归函数的数学形式。其中，w_0 和 w_1 为参数，也称为回归系数。w_0 又称为截距，w_1 又称为斜率。斜率度量了 x 每变动一个单位，y 的均值的变化率。

回归分析就是条件回归分析，即在给定自变量的条件下，分析因变量的行为。所以，通常可以省略“条件”二字，表达式 $E(y|x_i)$也简写成 $E(y)$。

5.2.2　随机干扰的意义

现通过一个例子来说明随机干扰项的意义。表 5-1 给出了 21 种车型燃油消耗(单位：L/100km)和车重(单位：kg)。下面在 R 中使用下列命令读入数据文件，并绘制散点图，还可以用一条回归线拟合这些散点。

```
> cars <- read.csv("c:/racv.csv")
> plot(lp100km ~ mass.kg, data = cars,
 + xlab = "Mass (kg)", ylab = "Fuel consumption (l/100km)")
> abline(lm(lp100km ~ mass.kg, data = cars))
```

从表 5-1 中不难看出，车的油耗与车重呈正向关系，即车辆越重，油耗越高。如果用数学公式来表述这种关系，很自然地会想到采用直线方程来将这种依赖关系表示成下式

$$y_i = E(y) + e_i = w_0 + w_1 x_i + u_i$$

其中，u_i 表示误差项。上式也称为随机总体回归方程。

表 5-1 车型及相关数据

Make	L/100km	mass/kg
Alpha Romeo	9.5	1242
Audi A3	8.8	1160
BA Falcon Futura	12.9	1692
Chrysler PT Cruiser Classic	9.8	1412
Commodore VY Acclaim	12.3	1558
Falcon AU II Futura	11.4	1545
Holden Barina	7.3	1062
Hyundai Getz	6.9	980
Hyundai LaVita	8.9	1248
Kia Rio	7.3	1064
Mazda 2	7.9	1068
Mazda Premacy	10.2	1308
Mini Cooper	8.3	1050
Mitsubishi Magna Advance	10.9	1491
Mitsubishi Verada AWD	12.4	1643
Peugeot 307	9.1	1219
Suzuki Liana	8.3	1140
Toyota Avalon CSX	10.8	1520
Toyota Camry Ateva V6	11.5	1505
Toyota Corolla Ascent	7.9	1103
Toyota Corolla Conquest	7.8	1081

易见，某一款车型的燃油消耗量等于两部分之和：第一部分是由相应重量决定的燃油消耗期望 $E(y)=w_0+w_1x_i$，也就是在重量取 x_i 时，回归直线上相对应的点，这一部分称为系统的或者非随机的部分；第二部分 u_i 称为非系统的或随机的部分，在本例中由除了车重以外的其他因素所决定。

误差项 u_i 是一个随机变量，因此，其取值无法先验地知晓，通常用概率分布来描述它。随机误差项可能代表了人类行为中一些内在的随机性。即使模型中已经包含了所有的决定燃油消耗的有关变量，燃油消耗的内在随机性也会发生变化，这是做任何努力都无法解释的。即使人类行为是理性的，也不可能是完全可以预测的。所以在回归方程中引入 u_i 是希望可以反映人类行为中的这一部分内在随机性。

此外，随机误差项可以代表测量误差。在收集、处理统计数据时，由于仪器的精度、操作人员的读取或登记误差，总是会导致有些变量的观测值并不精准地等于实际值。所以误差项 u_i 也代表了测量误差。

随机误差项也可能代表了模型中并未包括变量的影响。有时在建立统计模型时，并非事无巨细、无所不包的模型就是最好的模型。恰恰相反，有时只要能说明问题，建立的模型可能越简单越好。即使知道其他变量可能对因变量有影响，我们也倾向于将这些次要因素归入随机误差项 u_i 中。

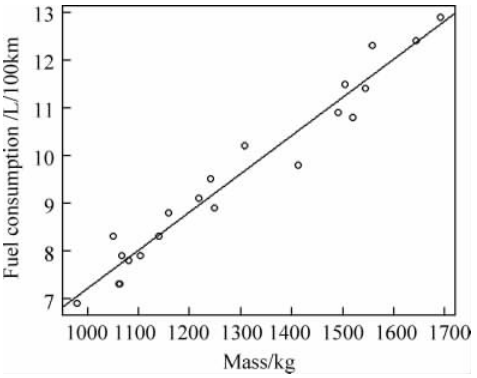

图 5-2　油耗与车重的关系

5.2.3　样本的回归函数

如何求得总体回归函数中的参数 w_0 和 w_1 呢？显然在实际中，很难获知整个总体的全部数据。更多的时候，我们仅有来自总体的一部分样本。于是任务就变成了根据样本提供的信息来估计总体回归函数。下面来看一个类别数据的例子。

一名园艺师想研究某种树木的树龄与树高之间的关系，于是他随机选定了 24 株树龄在 2～7 年的树苗，每个树龄选择 4 棵，并记录每棵树苗的高度，具体数据如表 5-2 所示。表中同时给出了每个树龄对应的平均树高，例如对于树龄为 2 的 4 棵树苗，它们的平均树高是 5.35。但在这个树龄下，并没有哪棵树苗的树高恰好等于 5.35。那么我们如何解释在某一个树龄下，具体某一棵树苗的树高呢？不难看出每个树龄对应的一棵树苗的高度等于平均树高加上或减去某一个数量，用数学公式表达即为

$$y_{ij} = w_0 + w_1 x_i + u_{ij}$$

某一个树龄 i 下，第 j 棵树苗的高度可以看作两个部分的和：第一部分为该树龄下所有树苗的平均树高，即 $w_0 + w_1 x_i$，反映在图形上，就是在此树龄水平下，回归直线上相对应的点；另一部分是随机项 u_{ij}。

表 5-2　树高与树龄

树龄/年	树高/m				平均树高/m
2	5.6	4.8	5.3	5.7	5.350
3	6.2	5.9	6.4	6.1	6.150
4	6.2	6.7	6.4	6.7	6.500
5	7.1	7.3	6.9	6.9	7.050
6	7.2	7.5	7.8	7.8	7.575
7	8.9	9.2	8.5	8.7	8.825

在上述例子中，并无法获知所有树苗的高度数据，而仅仅是从每个树龄中抽取了 4 棵树苗作为样本。而且类别数据也可以向非类别数据转换，我们也会在后面演示 R 中处理这类问题的方法。

样本回归函数可以用数学公式表示为

$$\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$$

其中，$\hat{y}_i$ 是总体条件均值 $E(y|x_i)$ 的估计量；$\hat{w}_0$ 和 $\hat{w}_1$ 分别表示 w_0 和 w_1 的估计量。并不是所有样本数据都能准确地落在各自的样本回归线上，因此，与建立随机总体回归函数一样，我们需要建立随机的样本回归函数。即

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

式中，e_i 表示 u_i 的估计量。通常把 e_i 称为残差(Residual)。从概念上讲，它与 u_i 类似，样本回归函数生成 e_i 的原因与总体回归函数中生成 u_i 的原因是相同的。

回归分析的主要目的是根据样本回归函数

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

来估计总体回归函数

$$y_i = w_0 + w_1 x_i + u_i$$

样本回归函数是总体回归函数的近似。那么能否找到一种方法，使得这种近似尽可能地接近真实值？换言之，一般情况下很难获得整个总体的数据，那么如何建立样本回归函数，使得$\hat{w}_0$ 和$\hat{w}_1$ 尽可能接近 w_0 和 w_1 呢？我们将在下一小节介绍相关技术。

5.3 回归模型的估计

本小节介绍一元线性回归模型的估计技术，并结合之前给出的树龄与树高关系的例子，演示在 R 中进行线性回归分析的方法。

5.3.1 普通最小二乘法原理

在回归分析中，最小二乘法是求解样本回归函数时最常被用到的方法。本小节就来介绍它的基本原理。一元线性总体回归方程为

$$y_i = w_0 + w_1 x_i + u_i$$

由于总体回归方程不能进行参数估计，因此只能对样本回归函数

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

进行估计。因此有

$$e_i = y_i - \hat{y}_i = y_i - \hat{w}_0 - \hat{w}_1 x_i$$

从上式可以看出，残差 e_i 是 y_i 的真实值与估计值之差。估计总体回归函数的最优方法是，选择 w_0、w_1 的估计值$\hat{w}_0$、$\hat{w}_1$，使得残差 e_i 尽可能小。最小二乘法的原则是选择合适的参数 $\hat{w}_0$、$\hat{w}_1$，使得全部观察值的残差平方和为最小。

最小二乘法用数学公式可以表述为

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

总而言之，最小二乘原理就是所选择的样本回归函数使得所有 y 的估计值与真实值差的平方和为最小。这种确定参数$\hat{w}_0$ 和$\hat{w}_1$ 的方法就叫做最小二乘法。

对于二次函数 $y=ax^2+b$ 来说，当 $a>0$ 时，函数图形的开口朝上，所以必定存在极小值。根据这一性质，因为 $\sum e_i^2$ 是$\hat{w}_0$ 和$\hat{w}_1$ 的二次函数，并且是非负的，所以 $\sum e_i^2$ 的极小值

总是存在的。根据微积分中的极值原理，当$\sum e_i^2$取得极小值时，$\sum e_i^2$对$\hat{w}_0$和$\hat{w}_1$的一阶偏导数为零，即

$$\frac{\partial \sum e_i^2}{\partial \hat{w}_0}=0,\quad \frac{\partial \sum e_i^2}{\partial \hat{w}_1}=0$$

由于

$$\sum e_i^2=\sum (y_i-\hat{w}_0-\hat{w}_1x_i)^2=\sum [(y_i-\hat{w}_1x_i)^2+\hat{w}_0^2-2\hat{w}_0(y_i-\hat{w}_1x_i)]$$

则得

$$\frac{\partial \sum e_i^2}{\partial \hat{w}_0}=-2\sum (y_i-\hat{w}_0-\hat{w}_1x_i)=0$$

$$\frac{\partial \sum e_i^2}{\partial \hat{w}_1}=-2\sum (y_i-\hat{w}_0-\hat{w}_1x_i)x_i=0$$

即

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

$$\sum x_iy_i=\hat{w}_0\sum x_i+\hat{w}_1\sum x_i^2$$

以上两式构成了以$\hat{w}_0$和$\hat{w}_1$为未知数的方程组，通常叫做正规方程组，或简称正规方程。解正规方程，得到

$$\hat{w}_0=\frac{\sum x_i^2\sum y_i-\sum x_i\sum x_iy_i}{n\sum x_i^2-(\sum x_i)^2}$$

$$\hat{w}_1=\frac{n\sum x_iy_i-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2}$$

等式左边的各项数值都可以由样本观察值计算得到。由此便可求出w_0、w_1的估计值$\hat{w}_0$、$\hat{w}_1$。

若设

$$\bar{x}=\frac{1}{n}\sum x_i,\quad \bar{y}=\frac{1}{n}\sum y_i$$

则可以将$\hat{w}_0$的表达式整理为

$$\hat{w}_0=\bar{y}-\hat{w}_1\bar{x}$$

由此便得到了总体截距w_0的估计值。其中，$\hat{w}_1$的表达式如下

$$\hat{w}_1=\frac{\sum x_iy_i-n\bar{x}\bar{y}}{\sum x_i^2-n\bar{x}^2}$$

这也就是总体斜率w_1的估计值。

为了方便起见，在实际应用中，经常采用离差的形式表示$\hat{w}_0$和$\hat{w}_1$。为此设

$$x'_i=x_i-\bar{x},\quad y'_i=y_i-\bar{y}$$

因为

$$\sum x_i'y_i' = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i y_i - \bar{x} y_i - x_i \bar{y} + \bar{x}\,\bar{y})$$
$$= \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\,\bar{y} = \sum x_i y_i - n\bar{x}\,\bar{y}$$
$$\sum x_i'^2 = \sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$$

所以$\hat{w}_0$、$\hat{w}_1$的表达式可以写成

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

5.3.2 一元线性回归的应用

上一小节中已经给出了最小二乘法的基本原理，而且还给出了计算斜率的几种不同方法。现在就以树高与树龄关系的数据为例来实际计算回归函数的估计结果。

正如前面说过的那样，类别数据可以转化成非类别数据，进而完成一元线性回归分析。其方法就是通过重复类别项从而将原来以二维数据表示的因变量转化为一维数据的形式。例如，在 R 中可以采用下列方法组织树高与树龄关系的数据。

```
> 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))
```

上述代码将会得到如表 5-3 所示的数据组织形式。根据上一小节所得出的计算公式，我们还需计算相应的 x_i^2 和 $x_i y_{ij}$，这些数据也一并在表中列出。

表 5-3 树龄与树高数据

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

基于表 5-3 中的数据进而可以算得

$$\bar{x} = 4.5, \quad \bar{y} = 6.908$$

$$n\bar{x}^2 = 486, \quad n\bar{x}\,\bar{y} = 746.1$$

$$\sum x_i^2 = 556, \quad \sum x_i y_i = 790.5$$

进而可以算得模型中估计的截距和斜率如下

$$\hat{w}_1 = \left(\sum x_i y_i - n\bar{x}\,\bar{y}\right)/\left(\sum x_i^2 - n\bar{x}^2\right) \approx 0.634\ 29$$

$$\hat{w}_0 = \bar{y} - \hat{w}_1\,\bar{x} \approx 4.054\ 05$$

由此便得到最终的估计模型为

$$\hat{y}_i = 4.054\ 05 + 0.634\ 29x_i$$

或

$$y_i = 4.054\ 05 + 0.634\ 29x_i + e_i$$

当然，在 R 中并不需要这样繁杂的计算过程，仅需几条简单的命令就可以完成数据的线性回归分析。示例代码如下。

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

由上述代码产生的模型估计如下，其中截距的估计值由 Intercept 项中的 Estimate 条目给出，斜率的估计值由 age 项中的 Estimate 条目给出，具体数值已经用方框标出。这些数据与我们人工算得的结果是一致的。输出结果中的其他数据将在后续的篇幅中加以讨论。

```
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
```

模型的拟合结果由图 5-3 给出，代码如下。

```
> plot(height ~ age, data = plants)
> abline(plants.lm)
```

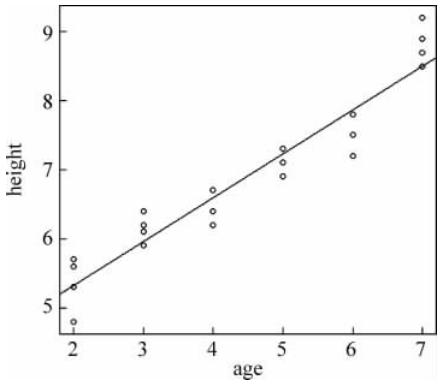

图 5-3　线性回归拟合结果

5.3.3　经典模型的基本假定

为了对回归估计进行有效的解释，就必须对随机干扰项 u_i 和解释变量 X_i 进行科学的假定，这些假定称为线性回归模型的基本假定。主要包括以下几个方面。

1. 零均值假定

由于随机扰动因素的存在，y_i 将在其期望附近上下波动，如果模型设定正确，y_i 相对于其期望的正偏差和负偏差都会有，因此随机项 u_i 可正可负，而且发生的概率大致相同。平均来看，这些随机扰动项有相互抵消的趋势。

2. 同方差假定

对于每个 x_i，随机干扰项 u_i 的方差等于一个常数 σ^2，即解释变量取不同值时，u_i 相对于各自均值的分散程度是相同的。同时也不难推证因变量 y_i 与 u_i 具有相同的方差。因此，该假定表明，因变量 y_i 可能取值的分散程度也是相同的。

前两个假设可以用公式 $u_i \sim N(0,\sigma^2)$ 来表述，通常我们都认为随机扰动（噪声）符合一个均值为 0，方差为 σ^2 的正态分布。

3. 相互独立性

随机扰动项彼此之间都是相互独立的。如果干扰的因素是全随机的，相互独立的，那么变量 y_i 的序列值之间也是互不相关的。

4. 因变量与自变量之间满足线性关系

这是建立线性回归模型所必需的。如果因变量与自变量之间的关系是杂乱无章、全无规律可言的，那么谈论建立线性回归模型就显然是毫无意义的。

R 中提供了 4 种基本的统计图形，用于对线性回归模型的假设基础进行检验。下面就用车重与燃油消耗的例子来说明这几种图形的意义。在 R 中输入下列代码，则可绘制出如图 5-4 所示的 4 张统计图形。

```
> cars.lm <- lm(lp100km ~ mass.kg, data = cars)
> par(mfrow = c(2, 2))
> plot(cars.lm)
```

图 5-4(a)是一幅残差对拟合值的散点图。图中的 x 轴是拟合值，也就是当 i 取不同值

时，有相应的$\hat{y}_i$值。y轴表示的是残差值，即e_i值。该图用于检验回归模型是否合理，是否有异方差性以及是否存在异常值。其中实线表示的附加线是采用局部加权回归散点修匀法(LOcally WEighted Scatterplot Smoothing，LOWESS)绘制的。如果残差的分布大致围绕着x轴，或红色附加线基本贴近x轴，则模型基本是无偏的；另外，如果残差的分布范围不随预测值的改变而大幅变化，则可以认为同方差假设成立。所以图形显示其模型基本上没有什么问题。

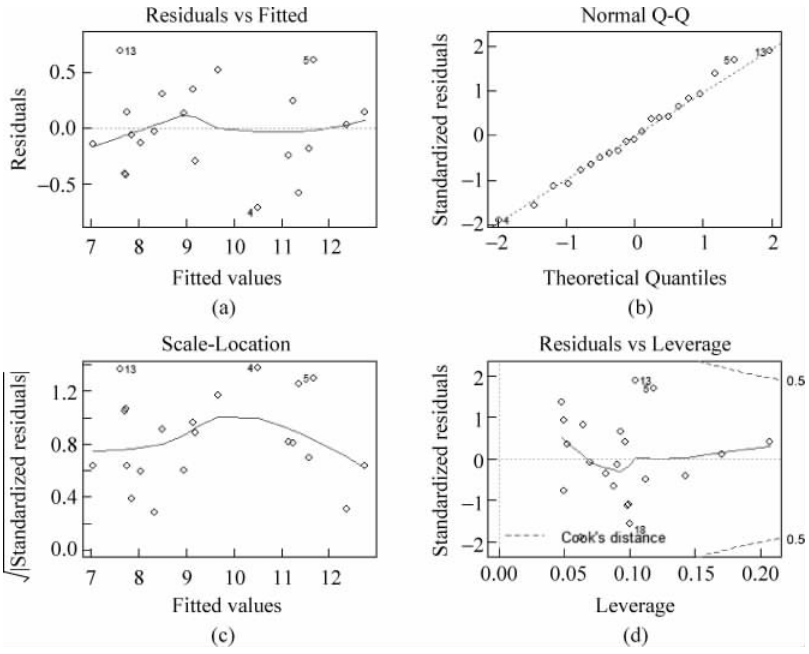

图 5-4　线性回归模型的诊断信息

图 5-4(b)展示了一幅标准化残差的 QQ 图，即将每个残差都除以残差标准差，然后再将结果与正态分布做比较。本书前面也已经对 QQ 图进行过较为详细的介绍，理想的结果是 QQ 图中的散点排列成一条直线，当然适度的偏离也是可以接受的。毕竟我们的采样点有限，根据中央极限定理，我们可以认为当采样点的数量足够大时，其结果会更加逼近正态分布。注意到在应用线性回归分析时，随机干扰项u_i应当满足正态分布这个假定，而残差相当于是对u_i的估计。如果图中散点的分布较大地偏离了直线，表明残差的分布是非正态的或者不满足同方差性，那么随机干扰的正态性自然也是不满足的。在我们给出的例子中，残差的正态性得到了较好的满足。

图 5-4(c)作用大致与第一幅图相同。图中的x轴是拟合值，y轴表示的是相应的标准化残差值绝对值的平方根。如果标准化残差的平方根大于 1.5，则说明该样本点位于 95%置信区间之外。中间的实线偏离于水平直线的程度较大，则意味着异方差性。尽管图中的实线表示的附加线不是一条完全水平的直线，但这种小的偏离主要是因为样本点的数量较小，所以图形显示我们的模型基本上没有什么问题。

图 5-4(d)是标准化残差对杠杆值的散点图，其作用是检查样本点中是否有异常值。如果删除样本点中的某一条数据，由此造成的回归系数变化过大，就表明这条数据对回归系数

的计算产生了明显的影响，这条数据就是异常值。需要好好考虑是否在模型中使用这条数据。设有帽子矩阵 $\boldsymbol{H}$，该矩阵的诸对角线元素记为 h_{ii}，这就是杠杆值(Leverage)。杠杆值用于评估第 i 个观测值离其余 $n-1$ 个观测值的距离有多远。对一元回归来说，其杠杆值为

$$h_{ii}=\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum(x_i-\bar{x})^2}$$

此外，图中还添加了 LOWESS 曲线和库克距离(Cook's Distance)曲线。库克距离用于诊断各种回归分析中是否存在异常数据。库克距离太大的样本点可能是模型的强影响点或异常值点，值得进一步检验。一个通常的判断准则是当库克距离大于 1 时就需要引起注意，图中显示所有点的库克距离都在 0.5 以内，所以没有异常点。

在本小节最后，尝试在 R 中自行绘制图 5-4(d)。这个过程有助于读者更好地理解杠杆值的意义。表 5-4 给出了操作步骤计算所得的中间结果。这些计算步骤需要用到的三个值，即斜率 0.008 024、截距 −0.817 768 和的残差标准差 0.3891，这些值都可以从线性回归的输出结果中直接得到。

表 5-4　中间结果数据

$(x_i-\bar{x})^2$	杠杆值	$\hat{y}_i$	e_i	标准化残差
2308.574	0.049 903	9.148 040	0.351 960	0.904 549
16 912.38	0.064 347	8.490 072	0.309 928	0.796 525
161 565.7	0.207 427	12.758 84	0.141 160	0.362 786
14 872.38	0.062 323	10.512 12	−0.712 120	−1.830 170
71 798.48	0.118 636	11.683 62	0.616 376	1.584 107
65 000.72	0.111 913	11.579 31	−0.179 310	−0.460 840
52 005.72	0.099 059	7.703 720	−0.403 720	−1.037 570
96 129.53	0.142 703	7.045 752	−0.145 750	−0.374 590
1768.002	0.049 368	9.196 184	−0.296 180	−0.761 200
51 097.53	0.098 161	7.719 768	−0.419 770	−1.078 820
49 305.15	0.096 388	7.751 864	0.148 136	0.380 714
322.2880	0.047 938	9.677 624	0.522 376	1.342 524
57 622.86	0.104 615	7.607 432	0.692 568	1.779 923
40 381.86	0.087 562	11.146 01	−0.246 020	−0.632 270
124 575.4	0.170 839	12.365 66	0.034 336	0.088 245
5047.764	0.052 612	8.963 488	0.136 512	0.350 840
22 514.29	0.069 888	8.329 592	−0.029 590	−0.076 050
52 878.10	0.099 922	11.378 71	−0.578 710	−1.487 310
46 204.53	0.093 321	11.258 35	0.241 648	0.621 043
34 986.81	0.082 225	8.032 704	−0.132 700	−0.341 050
43 700.91	0.090 844	7.856 176	−0.056 176	−0.144 370

下面给出绘制图形的 R 代码。

```
> plot(Std_Residuals ~ Leverage, xlab = "Leverage",
 + ylab = "Standardized residuals",
 + xlim = c(0,0.21), ylim = c( - 2,2), main = "Residuals vs Leverage")
> abline(v = 0.0, h = 0.0, lty = 3, col = "gray60")
> par(new = TRUE)
> lines(lowess(Std_Residuals~Leverage ), col = 'red')
```

执行上述代码，结果如图 5-5 所示，易见与 R 自动生成的效果一致。

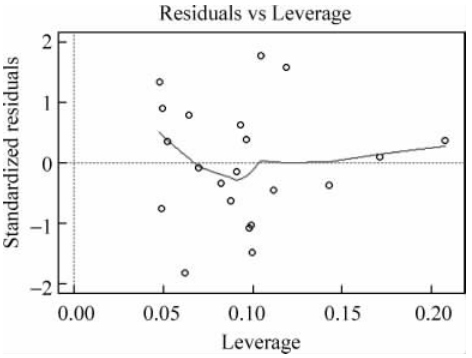

图 5-5　标准化残差对杠杆值的散点图

5.3.4　总体方差的无偏估计

前面谈到回归模型的基本假定中有这样一条：随机扰动(噪声)符合一个均值为 0，方差为 σ^2 的正态分布，即 $u_i \sim N(0,\sigma^2)$ 来表述。随机扰动 u_i 的方差 σ^2 又称为总体方差。由于总体方差 σ^2 未知，而且随机扰动项 u_i 也不可度量，所以只能从 u_i 的估计量——残差 e_i 出发，对总体方差 σ^2 进行估计。可以证明总体方差 σ^2 的无偏估计量为

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

证明：因为

$$\bar{y} = \frac{1}{n}\sum y_i$$

即 $\bar{y}$ 是有限个 y_i 的线性组合，所以当 $y_i = w_0 + w_1 x_i + u_i$，同样有

$$\bar{y} = w_0 + w_1 \bar{x} + \bar{u}$$

所以可得

$$\begin{aligned} y_i' &= y_i - \bar{y} = w_0 + w_1 x_i + u_i - (w_0 + w_1 \bar{x} + \bar{u}) \\ &= w_1 (x_i - \bar{x}) + (u_i - \bar{u}) = w_1 x_i + (u_i - \bar{u}) \end{aligned}$$

又因为

$$\left.\begin{aligned} e_i &= y_i - \hat{y}_i = y_i - \hat{w}_0 - \hat{w}_1 x_i = y_i' + \bar{y} - \hat{w}_0 - \hat{w}_1 (x_i' + \bar{x}) \\ \hat{w}_0 &= \bar{y} - \hat{w}_1 \bar{x} \end{aligned}\right\} \Rightarrow e_i = y_i' - \hat{w}_1 x_i'$$

所以有

$$e_i = w_1 x_i' + (u_i - \bar{u}) - \hat{w}_1 x_i' = (u_i - \bar{u}) - (\hat{\beta}_1 - \beta_1) x_i'$$

进而有

$$\begin{aligned} \sum e_i^2 &= \sum [(u_i - \bar{u}) - (\hat{w}_1 - w_1) x_i']^2 \\ &= (\hat{w}_1 - w_1)^2 \sum x_i'^2 + \sum (u_i - \bar{u})^2 - 2(\hat{w}_1 - w_1) \sum x_i' (u_i - \bar{u}) \end{aligned}$$

对上式两边同时取期望，则有

$$E\left(\sum e_i^2\right) = E\left[(\hat{w}_1 - w_1)^2 \sum x_i'^2\right] + E\left[\sum (u_i - \bar{u})^2\right]$$

$$-2E\left[(\hat{w}_1-w_1)\sum x'_i(u_i-\bar{u})\right]$$

然后对上式右端各项分别进行整理，可得

$$E\left[\sum(u_i-\bar{u})^2\right]=E\left[\sum(u_i^2-2u_i\bar{u}+\bar{u}^2)\right]=E\left[n\bar{u}^2+\sum u_i^2-2\bar{u}\sum u_i\right]$$

$$=E\left[\sum u_i^2-\frac{1}{n}\left(\sum u_i\right)^2\right]=\sum E(u_i^2)-\frac{1}{n}E\left(\sum u_i\right)^2$$

$$=\sum E(u_i^2)-\frac{1}{n}\left(\sum u_i^2+2\sum_{i\neq j}u_iu_j\right)$$

$$=n\sigma^2-\frac{1}{n}n\sigma^2-0=(n-1)\sigma^2$$

其中用到了 u_i 互不相关以及 $u_i\sim N(0,\sigma^2)$ 这两条性质。

一个变量与其均值的离差之总和恒为零，该结论可以简证如下

$$\bar{x}=\frac{1}{n}\sum x_i\Rightarrow n\bar{x}=\sum x_i\Rightarrow\sum\bar{x}=\sum x_i\Rightarrow\sum(x_i-\bar{x})=0$$

又因为 $\bar{y}$ 是一个常数，所以有

$$\sum x'_iy'_i=\sum x'_i(y_i-\bar{y})=\sum x'_iy_i-\bar{y}\sum x'_i$$

$$=\sum x'_iy_i-\bar{y}\sum(x_i-\bar{x})=\sum x'_iy_i$$

进而得到

$$\hat{w}_1=\frac{\sum x'_iy_i}{\sum x'^2_i}=\frac{\sum x'_iy_i}{\sum x'^2_i}=\sum k_iy_i$$

其中

$$k_i=\frac{x'_i}{\sum x'^2_i}$$

这其实说明 $\hat{w}_1$ 是 y 的一个线性函数；它是 y_i 的一个加权平均，以 k_i 为权数，从而它是一个线性估计量。同理，$\hat{w}_0$ 也是一个线性估计了。易证 k_i 满足下列性质

$$\sum k_i=\sum\left[\frac{x'_i}{\sum x'^2_i}\right]=\frac{1}{\sum x'^2_i}\sum x'_i=0$$

$$\sum k_i^2=\sum\left[\frac{x'_i}{\sum x'^2_i}\right]^2=\frac{\sum x'^2_i}{\left(\sum x'^2_i\right)^2}=\frac{1}{\sum x'^2_i}$$

$$\sum k_ix'_i=\sum k_ix_i=1$$

于是有

$$\hat{w}_1=\sum k_iy_i=\sum k_i(w_0+w_1x_i+u_i)$$

$$=w_0\sum k_i+w_1\sum k_ix_i+\sum k_iu_i=w_1+\sum k_iu_i$$

即

$$\hat{w}_1-w_1=\sum k_iu_i$$

以此为基础可以继续前面的整理过程，其中再次用到了 u_i 的互不相关性

$$E\left[(\hat{w}_1-w_1)\sum x'_i(u_i-\bar{u})\right]=E\left[\sum k_iu_i\sum x'_i(u_i-\bar{u})\right]$$

$$= E\Big[\sum k_i u_i \sum (x'_i u_i - x'_i \bar{u})\Big]$$
$$= E\Big[\sum k_i u_i \sum x'_i u_i - \bar{u}\sum k_i u_i \sum x'_i\Big]$$
$$= E\Big[\sum k_i u_i \sum x'_i u_i\Big] = E\Big[\sum k_i x'_i u_i^2\Big] = \sigma^2$$

此外还有

$$E\Big[(\hat{w}_1 - w_1)^2 \sum x_i'^2\Big] = E\Big[\Big(\sum k_i u_i\Big)^2 \sum x_i'^2\Big]$$
$$= E\Big[\sum \left(\frac{x'_i}{\sum x_i'^2} u_i\right)^2 \sum x_i'^2\Big] = E\Big[\sum (x'_i u_i)^2 / \sum x_i'^2\Big] = \sigma^2$$

综上可得

$$E\Big(\sum e_i^2\Big) = (n-1)\sigma^2 + \sigma^2 - 2\sigma^2 = (n-2)\sigma^2$$

原结论得证，可知 $\hat{\sigma}^2$ 是 σ^2 的无偏估计量。

5.3.5　估计参数的概率分布

中央极限定理表明，对于独立同分布的随机变量，随着变量个数的无限增加，其和的分布近似服从正态分布。随机项 u_i 代表了在回归模型中没有单列出来的其他所有影响因素。在众多的影响因素中，每种因素对 y_i 的影响可能都很微弱，如果用 u_i 来表示所有这些随机影响因素之和，则根据中央极限定理，就可以假定随机误差项服从正态分布，即 $u_i \sim N(0, \sigma^2)$。

因为 $\hat{w}_0$ 和 $\hat{w}_1$ 是 y_i 的线性函数，所以 $\hat{w}_0$ 和 $\hat{w}_1$ 的分布取决于 y_i。而 y_i 与随机干扰项 u_i 具有相同类型的分布，所以为了讨论 $\hat{w}_0$ 和 $\hat{w}_1$ 的概率分布，就必须对 u_i 的分布做出假定。这个假定十分重要，如果没有这一假定，$\hat{w}_0$ 和 $\hat{w}_1$ 的概率分布就无法求出，再讨论两者的显著性检验也就无的放矢了。

根据随机项 u_i 的正态分布假定可知，y_i 服从正态分布，根据正态分布变量的性质，即正态变量的线性函数仍服从正态分布，其概率密度函数由其均值和方差唯一决定。于是可得

$$\hat{w}_0 \sim N\left(w_0, \sigma^2 \frac{\sum x_i^2}{n\sum x_i'^2}\right)$$
$$\hat{w}_1 \sim N\left(w_1, \frac{\sigma^2}{\sum x_i'^2}\right)$$

并且 $\hat{w}_0$ 和 $\hat{w}_1$ 的标准差分布为

$$se(\hat{w}_0) = \sqrt{\sigma^2 \frac{\sum x_i^2}{n\sum x_i'^2}}$$
$$se(\hat{w}_1) = \sqrt{\frac{\sigma^2}{\sum x_i'^2}}$$

以 $\hat{w}_1$ 的分布为例，如图 5-6 所示，$\hat{w}_1$ 是 w_1 的无偏估计量，$\hat{w}_1$ 的分布中心是 w_1。易见，标准差可以用来衡量估计值接近于其真实值的程度，进而判定估计量的可靠性。

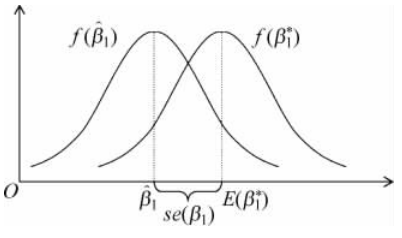

图 5-6　估计量的分布及其偏移

此前，已经证明 $\hat{\sigma}^2$ 是 σ^2 的无偏估计量，那么由此可知 $\hat{w}_0$ 和 $\hat{w}_1$ 的方差及标准差的估计量分别为

$$\mathrm{var}(\hat{w}_0)=\hat{\sigma}^2\frac{\sum x_i^2}{n\sum x_i'^2},\quad se(\hat{w}_0)=\hat{\sigma}\sqrt{\frac{\sum x_i^2}{n\sum x_i'^2}}$$

$$\mathrm{var}(\hat{w}_1)=\frac{\hat{\sigma}^2}{\sum x_i'^2},\quad se(\hat{w}_1)=\frac{\hat{\sigma}}{\sqrt{\sum x_i'^2}}$$

例如，在车重与油耗的例子中，一元线性回归的分析结果如下。其中，截距的估计值 $\hat{\beta}_0$ 的标准差为 0.506 422，斜率的估计值 $\hat{\beta}_1$ 的标准差为 0.000 387，这两个值已经用方框标出。

```
> summary(cars.lm)

Call:
lm(formula = lp100km ~ mass.kg, data = cars)

Residuals:
      Min        1Q    Median        3Q       Max
-0.71186 -0.24574 -0.02938  0.24193  0.69276

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.817768   0.506422  -1.615    0.123
mass.kg      0.008024   0.000387  20.733 1.65e-14 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3891 on 19 degrees of freedom
Multiple R-squared: 0.9577,    Adjusted R-squared: 0.9554
F-statistic: 429.9 on 1 and 19 DF, p-value: 1.653e-14
```

标准差可以被用来计算参数的置信区间。例如在本题中，w_0 的 95%的置信区间为

$$\begin{aligned}&-0.8178\pm c_{0.975}(t_{19})\times 0.5064\\=&-0.8178\pm 2.093\times 0.5064\\=&(-1.878,0.242)\end{aligned}$$

同理可以计算 w_1 的 95%的置信区间为

$$\begin{aligned}&0.008\,024\pm c_{0.975}(t_{19})\times 0.000\,387\\=&0.008\,024\pm 2.093\times 0.000\,387\\=&(0.0072,0.0088)\end{aligned}$$

其中，因为残差的自由度为 21−2=19，所以数值 2.093 是自由度为 19 的 t 分布值。当然在 R 中可以通过如下代码来完成上述计算过程。

```
> confint(cars.lm)
                    2.5 %     97.5 %
(Intercept) -1.877722151 0.24218677
mass.kg      0.007213806 0.00883382
```

5.4　正态条件下的模型检验

以样本观察值为基础，用最小二乘法求得样本回归直线，从而对总体回归直线进行拟合。但是拟合的程度怎样，必须要进行一系列的统计检验，从而对模型的优劣做出合理的评估，本节就介绍与模型评估检验有关的内容。

5.4.1　拟合优度的检验

由样本观察值(x_i, y_i)得出的样本回归直线为$\hat{y}_i = \hat{w}_0 + \hat{w}_1 x_i$，$y$ 的第 i 个观察值 y_i 与样本平均值$\bar{y}$的离差称为 y_i 的总离差，记为 $y_i' = y_i - \bar{y}$，不难看出总离差可以分成两部分，即

$$y_i' = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

其中一部分$\hat{y}_i' = \hat{y}_i - \bar{y}$是通过样本回归直线计算的拟合值与观察值的平均值之差，它是由回归直线（即解释变量）所解释的部分。另一部分 $e_i = y_i - \hat{y}_i$ 是观察值与回归值之差，即残差。残差是回归直线所不能解释的部分，它是由随机因素、被忽略掉的因素、观察误差等综合影响而产生的。各变量之间的关系如图 5-7 所示。

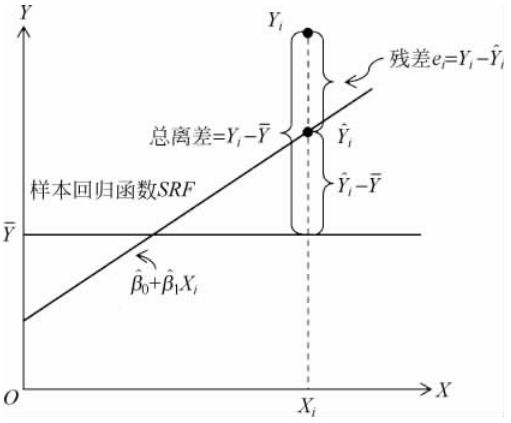

图 5-7　总离差分解

由回归直线所解释的部分$\hat{y}_i' = \hat{y}_i - \bar{y}$的绝对值越大，则残差的绝对值就越小，回归直线与样本点(x_i, y_i)的拟合就越好。

因为

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$$

如果用加总 y 的全部离差来表示显然是不行的，因为

$$\sum (y_i - \bar{y}) = \sum y_i - \sum \bar{y} = n\bar{y} - n\bar{y} = 0$$

所以考虑利用加总全部离差的平方和来反映总离差，即

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= \sum [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2 \\ &= \sum (y_i - y_i)^2 + \sum (\hat{y}_i - \bar{y})^2 + 2\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\end{aligned}$$

其中

$$\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

这是因为

$$\begin{aligned}\sum (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum e_i(\hat{w}_0 + \hat{w}_1 x_i - \bar{y}) \\ &= (\hat{w}_0 - \bar{y})\sum e_i + \hat{w}_1 \sum x_i e_i \\ &= (\hat{w}_0 - \bar{y})\sum e_i + \hat{w}_1 \sum x_i (y_i - \hat{w}_0 - \hat{w}_1 x_i)\end{aligned}$$

注意最小二乘法对于 e_i 有零均值假定，所以对其求和结果仍为零。而上述式子中最后一项为零，则是由最小二乘法推导过程中极值存在条件（令偏导数等于零）所保证的。

于是可得

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

或者写成

$$\sum y_i'^2 = \sum e_i^2 + \sum \hat{y}_i'^2$$

其中，

$$\sum y_i'^2 = \sum (y_i - \bar{y})^2$$

称为总离差平方和（Total Sum of Squares），用 SS_{total} 表示

$$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$$

称为残差平方和（Residual Sum of Squares），用 SS_{residual} 表示

$$\sum \hat{y}_i'^2 = \sum (\hat{y}_i - \bar{y})^2$$

称为回归平方和（Regression Sum of Squares），用 $SS_{\text{regression}}$ 表示。

总离差平方和可以分解成残差平方和与回归平方和两部分。总离差分解公式还可以写成

$$SS_{\text{total}} = SS_{\text{residual}} + SS_{\text{regression}}$$

这一公式也是方差分析 ANOVA 的原理基础，这一点在后续的章节中我们还会详细介绍。

在总离差平方和中，如果回归平方和比例越大，残差平方和所占比例就越小，表示回归直线与样本点拟合得越好；反之，就表示拟合得不好。把回归平方和与总离差平方和之比定义为样本判定系数，记为

$$R^2 = SS_{\text{regression}} / SS_{\text{total}}$$

判断系数 R^2 是一个回归直线与样本观察值拟合优度的数量指标，R^2 越大则拟合优度就越好；相反，R^2 越小，则拟合优度就越差。

注意 R 中指示判定系数的标签是“Multiple R-squared”，例如，在前面给出的树高与树龄的例子中，$R^2=0.9186$（$=91.86\%$），这表明模型的拟合程度较好。此外，R 的输出中还给出了所谓的调整判定系数，调整判定系数是对 R^2 的修正，指示标签为“Adjusted R-squared”。例如，在树高与树龄的例子中调整判定系数大小为 0.9149。

在具体解释调整判定系数的意义之前，还需先考查一下进行线性回归分析时，R 中输出的另外一个值——残差标准误差（Residual Standard Error）。在树高与树龄的例子中，R 给出的结果数值是 0.3368。所谓残差的标准误差其实就是残差的标准差（Residual Standard Deviation）。前面已经证明过，在一元线性回归中，总体方差 σ^2 的无偏估计量为

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

所以残差的标准差为

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

如果将这一结论加以推广(即不仅限于一元线性回归),则有

$$s = \hat{\sigma} = \sqrt{\frac{SS_{\text{residual}}}{n - \text{被估计之参数的数量}}}$$

因为在一元线性回归中,被估计的参数只有 β_0 和 β_1 两个,所以此时被估计之参数的数量就是 2。而在树高与树龄的例子中,研究单元的数量 $n=24$,因此在 R 中的输出结果上有一句“on 22 degrees of freedom”。

调整判定系数的定义为

$$1 - R^2_{\text{adj}} = s^2 / s_y^2$$

根据前面给出的公式可知

$$s^2 = \frac{SS_{\text{residual}}}{n-p}$$

其中,p 是模型中参数的数量。以及

$$s_y^2 = \frac{SS_{\text{total}}}{n-1}$$

一般认为调整判定系数会比判定系数更好地反映回归直线与样本点的拟合优度。那么其理据何在呢?注意残差 e_i 是扰动项 u_i 的估计值,因为 u_i 的标准差 σ 无法计算,所以借助 e_i 对其进行估计,而且也可以证明其无偏估计的表达式需要借助自由度来进行修正。另一方面,本书前面也曾经证明过当用样本来估计总体时,方差的无偏估计需要通过除以 $n-1$ 来进行修正。所以采用上述公式来计算会得到更加准确的结果。

经过简单的代数变换,可得出 R^2_{adj} 的另外一种算式

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

对于树高与树龄的例子有

$$R^2_{\text{adj}} = 0.9186 - \frac{2-1}{24-2}(1 - 0.9186) \approx 0.9149$$

这与 R 中输出的结果相同。通常情况下,R^2_{adj} 的值都会比 R^2 的值略小,且两者的差异一般都不大。

5.4.2　整体性假定检验

如果随机变量 X 服从均值为 μ、方差为 σ^2 的正态分布,即 $X \sim N(\mu, \sigma^2)$,则随机变量 $Z=(X-\mu)/\sigma$ 是标准正态分布,即 $Z \sim N(0,1)$。统计理论表明,标准正态变量的平方服从自由度为 1 的 χ^2 分布,用符号表示为

$$Z^2 \sim \chi_1^2$$

其中,χ^2 的下标表示自由度为 1。与均值、方差是正态分布的参数一样,自由度是 χ^2 分布的

参数。在统计学中自由度有各种不同的含义，此处定义的自由度是平方和中独立观察值的个数。

总离差平方和 SS_{total} 的自由为 $n-1$，因变量共有 n 个观察值，由于这 n 个观察值受 $\sum y'_i = \sum (y_i - \bar{y}) = 0$ 的约束，当 $n-1$ 个观察值确定以后，最后一个观察值就不能自由取值了，因此 SS_{total} 的自由为 $n-1$。

回归平方和 $SS_{\text{regression}}$ 的自由度是由自变量对因变量的影响决定的，因此它的自由度取决于解释变量的个数。在一元线性回归模型中，只有一个解释变量，所以 $SS_{\text{regression}}$ 的自由度为 1。在多元回归模型中，如果解释变量的个数为 k 个，则其中 $SS_{\text{regression}}$ 的自由度为 k。因为 $SS_{\text{regression}}$ 的自由度与 SS_{residual} 的自由度之和等于 SS_{total} 的自由度，所以 SS_{residual} 的自由度为 $n-2$。

平方和除以相应的自由度称为均方差。因此 $SS_{\text{regression}}$ 的均方差为

$$\begin{aligned}\frac{\sum \hat{y}'^2_i}{1} &= \sum (y_i - \bar{y})^2 = \sum (\hat{w}_0 + \hat{w}_1 x_i - \bar{y})^2 \\ &= \sum [\hat{w}_0 + \hat{w}_1 (\bar{x} + x'_i) - \bar{y}]^2 = \sum [\bar{y} - \hat{w}_1 \bar{X} + \hat{w}_1 (\bar{x} + x'_i) - \bar{y}]^2 \\ &= \sum (\hat{w}_1 x'_i)^2 = \hat{w}_1^2 \sum x'^2_i\end{aligned}$$

而且还有 SS_{residual} 的均方差为 $\left(\sum e_i^2\right)/(n-2)$。可以证明，在多元线性回归的条件下（即回归方程中有 k 个解释变量 x_i，$i=1,2,\cdots,k$），有

$$\sum \hat{y}'^2_i \sim \chi^2_k$$

$$\sum e_i^2 \sim \chi^2_{(n-k-1)}$$

根据基本的统计学知识可知，如果 Z_1 和 Z_2 分别是自由度为 k_1 和 k_2 的分布变量，则其均方差之比服从自由度为 k_1 和 k_2 的 F 分布，即

$$F = \frac{Z_1/k_1}{Z_2/k_2} \sim F(k_1, k_2)$$

那么

$$F = \frac{\left(\sum \hat{y}'^2_i\right)/k}{\left(\sum e_i^2\right)/(n-k-1)} \sim F(k, n-k-1)$$

下面就利用 F 统计量对总体线性的显著性进行检验。首先，提出关于 k 个总体参数的假设

$$H_0: w_1 = w_2 = \cdots = w_k = 0$$

$$H_0: w_i \text{ 不全为 } 0, \quad i = 1,2,\cdots,k$$

进而根据样本观察值计算并列出方差分析数据如表 5-5 所示。

表 5-5　方差分析表

方差来源	平方和	自由度	均方差
SS_{residual}	$\sum \hat{y}'^2_i$	k	$\left(\sum \hat{y}'^2_i\right)/k$
$SS_{\text{regression}}$	$\sum e_i^2$	$n-k-1$	$\left(\sum e_i^2\right)/(n-k-1)$
SS_{total}	$\sum y'^2_i$		

然后在 H_0 成立的前提下计算 F 统计量

$$F=\frac{(\sum y_i'^2)/k}{(\sum e_i^2)/(n-k-1)}$$

对于给定的显著水平 α，查询 F 分布表得到临界值 $F_\alpha(1,n-k-1)$，如果 $F>F_\alpha(1,n-k-1)$，则拒绝原假设，说明犯第一类错误的概率非常之小。也可以通过与这个 F 统计量对应的 P 值来判断，说明如果原假设成立，得到此 F 统计量的概率很小即为 P 值。这个结果说明我们的回归模型中的解释变量对因变量是有影响的，即回归总体是显著线性的。相反，若 $F<F_\alpha(1,n-k-1)$，则接受原假设，即回归总体不存在线性关系，或者说解释变量对因变量没有显著的影响关系。

例如，对于树龄与树高的例子，给定 $\alpha=0.05$，可以查表或者在 R 中输入下列语句得到 $F_{0.05}(1,22)$ 的值。

```
> qf(0.05, 1, 22, lower.tail = FALSE)
[1] 4.30095
```

其中，参数 lower.tail 是一个逻辑值，模型情况下它的值为 FALSE，此时给定服从某分布的随机变量 X，求得的概率是 $P[X\leqslant x]$，如果要求 $P[X>x]$，要么用 $1-P[X\leqslant x]$，要么就令 lower.tail 的值为 TRUE。

经过简单计算易知 $\sum y_i'^2=28.162\,666\,3$，$\sum e_i^2=2.496\,047\,632$。由此便可算得 $F=248.223\,892\,3$。当然，R 中给出的线性回归分析结果也包含了这个结果。因为 $F>F_{0.05}(1,22)$，所以有理由拒绝原假设 H_0，即证明回归总体是显著线性的。也可以通过与这个 F 统计量对应的 P 值来判断，此时可以在 R 中使用下面的代码得到相应的 P 值。

```
> pf(248.2238923, 1, 22, lower.tail = FALSE)
[1] 1.821097e-13
```

可见，P 值远远小于 0.05，因此有足够的把握拒绝原假设。

本小节所介绍的其实就是方差分析(ANOVA)的基本步骤。在本书的后续章节中，还将对方差分析做专门介绍。一元线性回归模型中对模型进行整体性检验只用后面介绍的 t 检验即可。但在多元线性回归模型中，F 检验是检验统计假设的非常有用和有效的方法。

5.4.3　单个参数的检验

前面介绍了利用 R^2 来估计回归直线的拟合优度，但是 R^2 却不能告诉我们估计的回归系数在统计上是否显著，即是否显著地不为零。实际上确实有些回归系数是显著的，而有些又是不显著的。下面就来介绍具体的判断方法。

本章前面曾经给出了 $\hat{w}_0$ 和 $\hat{w}_1$ 的概率分布，即

$$\hat{w}_0\sim N\left(w_0,\sigma^2\frac{\sum x_i^2}{n\sum x_i'^2}\right)$$

$$\hat{w}_1\sim N\left(w_1,\frac{\sigma^2}{\sum x_i'^2}\right)$$

但在实际分析时，由于 σ^2 未知，只能用无偏估计量 $\hat{\sigma}^2$ 来代替。此时，一元线性回归的最小二乘估计量 $\hat{w}_0$ 和 $\hat{w}_1$ 的标准正态变量服从自由度为 $n-2$ 的 t 分布，即

$$t = \frac{\hat{w}_0 - w_0}{se(\hat{w}_0)} \sim t(n-2)$$

$$t = \frac{\hat{w}_1 - w_1}{se(\hat{w}_1)} \sim t(n-2)$$

下面以 w_1 为例，演示利用 t 统计量对单个参数进行检验的具体步骤。首先对回归结果提出如下假设

$$H_0: w_1 = 0$$

$$H_1: w_1 \neq 0$$

即在原假设条件下，解释变量对因变量没有影响。在备择假设条件下，解释变量对因变量有（正的或者负的）影响，因此备择假设是双边假设。

以原假设 H_0 构造 t 统计量并由样本观察值计算其结果，则

$$t = \frac{w_1}{se(\hat{w}_1)}$$

其中

$$se(\hat{w}_1) = \frac{\hat{\sigma}}{\sqrt{\sum x_i'^2}} = \sqrt{\frac{\sum e_i^2}{(n-2)\sum x_i'^2}}$$

可以通过给定的显著性水平 α，检验自由度为 $n-2$ 的 t 分布表，得临界值 $t_{\frac{\alpha}{2}}(n-2)$。如果 $|t| > t_{\frac{\alpha}{2}}(n-2)$，则拒绝 H_0，此时接受备择假设犯错的概率很小，即说明 w_1 所对应的变量 x 对 y 有影响。

相反，若 $|t| \leqslant t_{\frac{\alpha}{2}}(n-2)$，则无法拒绝 H_0，即 w_1 与 0 的差异不显著，说明 w_1 所对应的变量 x 对 y 没有影响，变量之间的线性关系不显著。对参数的显著性检验，还可以通过 P 值来判断，如果相应的 P 值很小，则可以拒绝原假设，即参数显著不为零。

例如，在树龄与树高的例子中，很容易算得

$$\sum x_i'^2 = 70$$

于是可得到 $se(\hat{w}_1) = 0.3368/\sqrt{70} = 0.040\ 26$，进而有 $t = 0.634\ 29/0.040\ 26 = 15.754\ 84$。相应的 P 值可以在 R 中用下列代码算得。

```
> 2 * (1 - pt(15.75484,22))
[1] 1.820766e - 13
```

经过计算所得之 t 值为 15.754 84，其 P 值几乎为 0。P 值越低，拒绝原假设的理由就越充分。现在来看，我们已经有足够的把握拒绝原假设，可见变量之间具有显著的线性关系。

5.5 一元线性回归模型预测

预测是回归分析的一个重要应用。这种所谓的预测通常包含两个方面，对于给定的点，一方面要估计它的取值，另一方面还应对可能取值的波动范围进行预测。

5.5.1　点预测

对于给定的 $x=x_0$，利用样本回归方程可以求出相应的样本拟合值$\hat{y}_0$，以此作为因变量个别值 y_0 或其均值 $E(y_0)$的估计值，这就是所谓的点预测。比如树龄与树高的例子，如果购买了一棵树苗，并且想知道该树的树龄达到 4 年时，其树高预计为多少。此时你希望求得的值，其实是树龄为 4 的该种树木的平均树高或者是期望树高。

已知含随机扰动项的总体回归方程为

$$y_i = E(y_i)+u_i = w_0 + w_1 x_i + u_i$$

当 $x=x_0$ 时，y 的个别值为

$$y_0 = w_0 + w_1 x_0 + u_0$$

其总体均值为

$$E(y_0) = w_0 + w_1 x_0$$

样本回归方程在 $x=x_0$ 时的拟合值为

$$\hat{y}_0 = \hat{w}_0 + \hat{w}_1 x_0$$

对上式两边取期望，得

$$E(\hat{y}_0) = E(\hat{w}_0 + \hat{w}_1 x_0) = w_0 + w_1 x_0 = E(y_0)$$

这表示在 $x=x_0$ 时，由样本回归方程计算的$\hat{y}_0$ 是个别值 y_0 和总体均值 $E(y_0)$的无偏估计，所以$\hat{y}_0$ 可以作为 y_0 和 $E(y_0)$的预测值。

5.5.2　区间预测

对于任一给定样本，估计值$\hat{y}_0$ 只能作为 y_0 和 $E(y_0)$的无偏估计量，不一定能够恰好等于 y_0 和 $E(y_0)$。也就是说，两者之间存在误差，这个误差就是预测误差。由这个误差开始，期望得到 y_0 和 $E(y_0)$的可能取值范围，这就是区间预测。

定义误差 $\delta_0=\hat{y}_0-E(y)$，由于$\hat{y}_0$ 服从正态分布，所以 δ_0 是服从正态分布的随机变量。而且可以得到 δ_0 的数学期望与方差如下

$$\begin{aligned}
E(\delta_0) &= E[\hat{y}_0 - E(y)] = 0\\
\mathrm{var}(\delta_0) &= E[\hat{y}_0 - E(y)]^2 = E[\hat{w}_0 + \hat{w}_1 x_0 - (w_0 + w_1 x_0)]^2\\
&= E[(\hat{w}_0 - w_0)^2 + 2(\hat{w}_0 - w_0)(\hat{w}_1 - w_1) + (\hat{w}_1 - w_1)^2 x_0^2]\\
&= \mathrm{var}(\hat{w}_0) + 2x_0\,\mathrm{cov}(\hat{w}_0, \hat{w}_1) + \mathrm{var}(\hat{w}_1)x_0^2
\end{aligned}$$

其中，$\hat{w}_0$ 和$\hat{w}_1$ 的协方差为

$$\begin{aligned}
\mathrm{cov}(\hat{w}_0, \hat{w}_1) &= E[(\hat{w}_0 - w_0)(\hat{w}_1 - w_1)]\\
&= E[(\bar{y} - \hat{w}_1\,\bar{x} - w_0)(\hat{w}_1 - w_1)]\\
&= E[(w_0 + w_1\,\bar{x} + \bar{u} - \hat{w}_1\,\bar{x} - w_0)(\hat{w}_1 - w_1)]\\
&= E\{[-(\hat{w}_1 - w_1)\bar{x} + \bar{u}](\hat{w}_1 - w_1)\}\\
&= \bar{x}E(\hat{w}_1 - w_1)^2 + E(\bar{u}\,\hat{w}_1)
\end{aligned}$$

因为

$$E(\hat{w}_1 - w_1)^2 = \mathrm{var}(\hat{w}_1) = \frac{\sigma^2}{\sum x_i'^2}$$

$$E(\bar{u}\,\hat{w}_1) = \frac{1}{n}E\left(\sum u_i \sum \frac{x_i'}{\sum x_i'^2} y_i\right)$$

$$= \frac{1}{n}\left(\sum_{i=j} x_i' \Big/ \sum x_i'^2\right)E(u_i y_i) + \frac{1}{n}\left(\sum_{i\neq j} x_i' \Big/ \sum x_i'^2\right)E(u_i y_i)$$

$$= \frac{\sigma^2 \sum x_i'}{\sum x_i'^2}E(u_i y_i) = 0$$

所以

$$\mathrm{cov}(\hat{w}_0, \hat{w}_1) = -\frac{\bar{x}\sigma^2}{\sum x_i'^2}$$

于是可得

$$\mathrm{var}(\delta_0) = \frac{\sigma^2 \sum x_i'^2}{n \sum x_i'^2} - \frac{2\sigma^2 x_0\,\bar{x}}{\sum x_i'^2} + \frac{\sigma^2 x_0^2}{\sum x_i'^2}$$

$$= \frac{\sigma^2}{\sum x_i'^2}\left(\frac{\sum x_i'^2 - n\bar{x}}{n} + \bar{x}^2 - 2x_0\,\bar{x} + x_0^2\right)$$

$$= \frac{\sigma^2}{\sum x_i'^2}\left[\frac{\sum x_i'^2}{n} + (x_0 - \bar{x})^2\right] = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}\right]$$

由 δ_0 的数学期望与方差可知

$$\delta_0 \sim N\left\{0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}\right]\right\}$$

将 δ_0 标准化，则有

$$\frac{\delta_0}{\sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}}} \sim N(0,1)$$

由于 σ 未知，所以用 $\hat{\sigma}$ 来代替，根据抽样分布理论及误差 δ_0 的定义，有

$$\frac{\hat{y}_0 - E(y_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}}} \sim t(n-2)$$

那么 $E(y_0)$ 的预测区间为

$$\hat{y}_0 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}} \leqslant E(y_0) \leqslant \hat{y}_0 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum x_i'^2}}$$

其中 α 为显著水平。

在 R 中可以使用下面的代码来获得总体均值 $E(y_0)$ 的预测区间。

```
> predict(plants.lm,
 + newdata = data.frame(age = 4),
```

```
+ interval = "confidence")
       fit      lwr      upr
1 6.59119 6.442614 6.739767
```

在此基础上，还可以对总体个别值 y_0 的可能区间进行预测。设误差 $e_0=y_0-\hat{y}_0$，由于 $\hat{y}_0$ 服从正态分布，所以 e_0 也服从正态分布。而且可以得到 e_0 的数学期望与方差如下。

$$E(e_0)=E(y_0-\hat{y}_0)=0$$

$$\mathrm{var}(e_0)=\mathrm{var}(y_0-\hat{y}_0)$$

由于 $\hat{y}_0$ 与 y_0 相互独立，并且

$$\mathrm{var}(y_0)=\mathrm{var}(w_0+w_1x_0+u_0)=\mathrm{var}(u_0)$$

$$\mathrm{var}(\hat{y}_0)=E[\hat{y}_0-E(y_0)]^2=\mathrm{var}(\delta_0)$$

所以

$$\mathrm{var}(e_0)=\mathrm{var}(y_0)+\mathrm{var}(\hat{y}_0)=\mathrm{var}(u_0)+\mathrm{var}(\delta_0)$$

$$=\sigma^2+\sigma^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}\right]=\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}\right]$$

由 e_0 的数学期望与方差可知

$$e_0\sim N\left\{0,\sigma^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}\right]\right\}$$

将 e_0 标准化，则有

$$\frac{e_0}{\sigma\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}}}\sim N(0,1)$$

由于 σ 未知，所以用 $\hat{\sigma}$ 来代替，根据抽样分布理论及误差 e_0 的定义，有

$$\frac{y_0-\hat{y}_0}{\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}}}\sim t(n-2)$$

那么 y_0 的预测区间为

$$\hat{y}_0-t_{\frac{\alpha}{2}}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}}\leqslant y_0\leqslant\hat{y}_0+t_{\frac{\alpha}{2}}\cdot\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum x_i'^2}}$$

在 R 中可以使用下面的代码来获得总体个别值 y_0 的预测区间。

```
> predict(plants.lm, newdata = data.frame(age = 4), interval = "prediction")
      fit      lwr      upr
1 6.59119 5.877015 7.305366
```

可见在执行 predict 函数时，通过选择参数“confidence”或“prediction”即可实现对 y_0 或者 y_0 期望及其置信区间(或称置信带)的估计。而且 y_0 期望的置信区间要比 y_0 的置信区间更窄。