第 5 章

Chapter 5

线性回归——使用 R

学习成果

通过本章的学习，您将能够：

- 解释回归分析，通常用于根据预测变量预测一个结果（目标或响应）变量的值；
- 创建一个简单线性回归模型；
- 使用残差与拟合图（residual vs. fitted plot）、正态分位数图（normal Q-Q plot）、位置尺度图（scale location plot）和残差与杠杆图（residuals vs. leverage plot）验证模型。

5.1 概述

回归分析是一种估计变量之间关系的统计过程。当回归分析关注的是一个因变量（也称目标或响应变量）和一个或多个自变量（也称预测变量）之间的关系时，会包括许多建模和分析变量的技术。简单线性回归用于确定一个因变量和单个自变量之间的线性关系的程度。通常，回归分析用于以下三种目的：对目标变量的预测（预告）；对 x 和 y 之间关系的建模；对假设的测试。

5.2 模型拟合

R 语言中的模型是数据点序列的表示，具有看起来嘈杂的点云。模型拟合是指选择最适合的模型描述一组数据，R 有不同类型的模型，下面列出这些模型及其对应的命令。

- 线性模型（Linear Model，LM）：lm()是 R 中的一个线性模型函数，用于构建一个简单回归模型。
- 广义线性模型（Generalised Linear Model，GLM）：通过给出线性预测器的一种符号描述和误差分布的描述进行定义。
- 混合效应线性模型（Linear Model for Mixed Effect，LME）。
- 非线性最小二乘（Non-linear Least Square，NLS）：确定一个非线性模型的参数的非线性（加权）最小二乘估计。

- 广义加法模型(Generalised Additive Model,GAM)：GAM 只是一种统计模型,通常,响应变量和预测变量之间的线性关系会被几个用于数据建模和捕获非线性特征的非线性平滑函数代替。

每种模型都有特定的函数,数据点的分布基于描述模型的函数。

5.3　线性回归

R 中的线性回归由两个主要变量组成,它们通过两个变量的指数(幂)为 1 的方程相互关联。在数学上,当绘制图形时,线性关系表示一条直线。在存在非线性关系的情况下,变量的指数不等于 1,它会在图上创建一条曲线。

一般的线性回归方程为

$$y = ax + b$$

其中,y 是一个响应变量,x 是一个预测变量,a 和 b 是常数,称为系数。

5.3.1　R 中的 lm()函数

```
lm(formula, data, subset, weights, na.action,
    method ="qr", model =TRUE, x =FALSE, y =FALSE, qr =TRUE,
    singular.ok =TRUE, contrasts =NULL, offset, …)
```

其中,

- formula 表示 x 和 y 之间的关系。
- data 包含模型中的变量。
- subset 是一个可选的向量,它指定了用于模型拟合的观测值的一个子集。
- weights 是一个可选的向量,它指定了模型拟合过程中的权重,它是一个数字向量或 NULL。
- na.action 是一个可选的函数,它指定了对于包含 NA 的数据如何反应的动作。
- method 是用于拟合的方法。
- model,x,y,qr：如果对应参数为 TRUE,则返回模型矩阵、模型框和 QR 分解。
- singular.ok：如果该参数为 FALSE,则奇异拟合(singular fit)会出错。
- contrasts：一个可选的列表偏移量,用于指定包含在线性预测器中的先前已知的成分。

线性回归中的 lm()函数的简单语法为 lm(formula, data),其中删除了可选参数。

确定一个学生数据集(student data set)的预测变量和响应变量之间的关系模型。预测向量存储着学生在学习上投入的时间,响应向量存储着学生的分数。

1. 检查数据集中的数据

考虑表 5.1 所给出的数据集 D:\student.csv,显示了学生在学习上投入的时间(NoOfHours)和分数(Freshmen_Score)。

表 5.1 student 数据集中的数据

小 时 数	成 绩	小 时 数	成 绩
2	55	4.5	82
2.5	62	5	75
3	65	5.5	83
3.5	70	6	85
4	77	6.5	88

2. 将数据集中的数据读取到数据框中

使用 read.table()函数以表格形式读取文件 D:\student.csv,并从中创建一个数据框 HS,使用与文件中的行和字段变量相对应的大小写。

```
>HS <-read.table("D:/student.csv", sep=",",header=TRUE)
>HS
   NoOfHours     Freshmen_Score
1        2.0                 55
2        2.5                 62
3        3.0                 65
4        3.5                 70
5        4.0                 77
6        4.5                 82
7        5.0                 75
8        5.5                 83
9        6.0                 85
10       6.5                 88
```

3. 查看数据框中数据的结果汇总

使用 summary()函数生成结果汇总。这里,对所有数字变量计算出最小值、第一四分位数、中值、均值、第三四分位数和最大值。

```
>summary(HS)
   NoOfHours    Freshmen_Score
Min.   : 2.000   Min.   : 55.00
1st Qu.: 3.125   1st Qu.: 66.25
Median : 4.250   Median : 76.00
Mean   : 4.250   Mean   : 74.20
3rd Qu.: 5.375   3rd Qu.: 82.75
Max.   : 6.500   Max.   : 88.00
```

4. 查看数据框的内部结构

显示 R 对象 HS 的内部结构,它表明有 2 个变量 NoOfHours 和 Freshmen_Score 的 10

组观测值。

```
>str(HS)
'data.frame'     : 10 obs. of 2 variables:
$NoOfHours       : num 2 2.5 3 3.5 4 4.5 5 5.5 6 6.5
$Freshmen_Score  : int 55 62 65 70 77 82 75 83 85 88
```

5. 绘制 R 对象

绘制 R 对象 HS，x 轴使用 HS$NoOfHours，y 轴使用 HS$Freshmen_Score，如图 5.1 所示。

```
>plot (HS$NoOfHours, HS$Freshmen_Score)
```

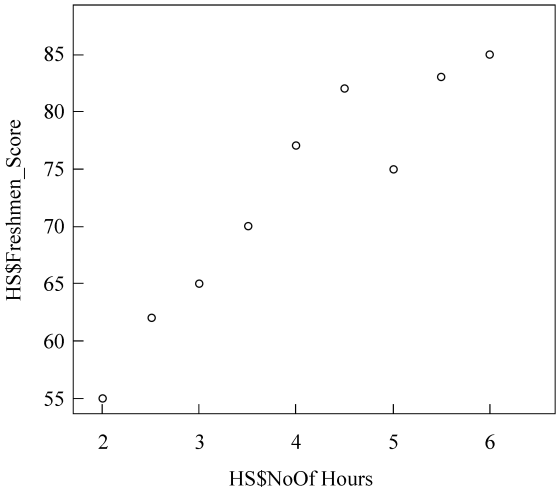

图 5.1　预测变量与响应变量的散点图

在图 5.2 所示的均值处(Freshmen_Score 的均值为 74.20)绘制一条水平线横穿该图。

```
>abline (h=mean (HS$Freshmen_Score))
```

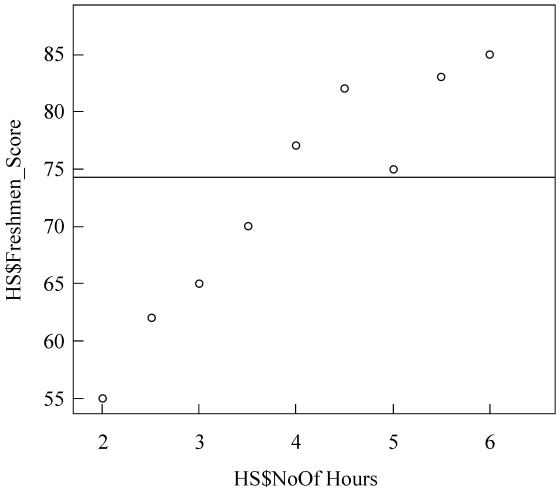

图 5.2　在均值处绘制一条直线的预测变量与响应变量的散点图

当使用均值预测 Freshmen_Score 的分数时，在某些情况下可以观察到实际（观测）值和预测值之间的显著差异。

例如，第一个学生的 Freshmen_Score 为 55，如果使用均值预测分数，则会预测它是 74.20。在这里，观测值小于期望值。对于第 10 个学生，观测值为 88，观测值大于预测分数 74.20。

这表明要想预测期望的分数，则还应该考虑其他因素。

6. 相关系数

$$r=\frac{n\left(\sum xy\right)-\left(\sum x\right)\left(\sum y\right)}{\sqrt{\left[n\sum x^2-\left(\sum x\right)^2\mid n\sum y^2-\left(\sum y\right)^2\right]}}$$

对于学生数据集，其计算为

$$\begin{aligned}r&=10\times 3295.5-42.5\times 742/\text{square root of}([10\times 201.25-1806.25]\times\\&\quad[10\times 56130-550564])\\&=32955-31545/1488.05\\&=1410/1488.05\\&=0.947548805\\&=0.95\end{aligned}$$

7. 使用 R 中的 cor()函数确定线性关联的度和方向

线性关联的度（degree）和方向（direction）可以使用相关性确定。学习用时数和 GPA 分数之间关联的 Pearson 相关系数显示如下：

```
>cor(HS$NoOfHours, HS$Freshmen_Score)
[1] 0.9542675
```

这里的相关性值表明学习的用时数和学生的分数之间有一个强关联关系。

但是，对于相关性有以下一些需要注意的问题。

① 对于非线性关系，相关性不是一个度量关联的合适方法。如果要确定两个变量是否线性相关，则应该使用散点图。

② 异常值会影响 Pearson 相关性，箱线图可用于辨别异常值，异常值对 Spearman 相关性的影响是很小的。因此，如果没有合适的理由将异常值从分析中删除，则 Spearman 相关性是首选。

③ 相关性的取值接近于 0 表明变量不是线性关联的，但是这些变量可能依然是相关的，因此建议绘制这些数据。

④ 相关性并不意味着因果关系，即基于相关性的值并不能判断一个变量可以引起另一个变量。

⑤ 相关性分析仅仅有助于确定关联度。

因为相关性分析在确定因果关系时可能是不合适的，因此需要使用回归技术度量变量之间关系的本质。

使用一个数学模型 $y=f(X)$ 表示回归模型，其中 y 是因变量，X 是预测变量（x_1，x_2…

x_n)的集合。

一般而言，$f(X)$是线性或非线性的形式。

① 线性形式：$f(X)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\varepsilon$。

② 非线性形式：$f(X)=\beta_0+\beta_1x_1^{p1}+\beta_2x_2^{p2}+\cdots+\beta_nx_n^{pn}+\varepsilon$。

一些常用的线性形式如下。

① 简单线性形式：有一个预测变量和一个因变量，即 $f(X)=\beta_0+\beta_1x_1+\varepsilon$。

② 多重线性形式：有多个预测变量和一个因变量，即

$$f(X)=\beta_0+\beta_1x_1+\beta_2x_2+\cdots+\beta_nx_n+\varepsilon$$

一些常用的非线性形式如下。

① 多项式形式：$f(X)=\beta_0+\beta_1x_1^{p1}+\beta_2x_2^{p2}+\cdots+\beta_nx_n^{pn}+\varepsilon$。

② 二次型：$f(X)=\beta_0+\beta_1x_1^2+\beta_2x_1+\varepsilon$。

③ Logistic 形式：

$$f(X)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\cdots\beta_nx_n)}}+\varepsilon$$

其中，β_0，β_1，$\beta_2\cdots\beta_n$ 是回归系数，ε 是预测中的误差。回归系数和预测中的误差都是实数。

当一个回归模型是线性形式时，这样的回归称为线性回归。类似地，当一个回归模型是非线性形式时，这样的回归称为非线性回归。

由于学生投入的学习时间与学生的分数之间的散点图表明二者存在线性关联，因此建立一个线性回归模型以量化这种关系的本质。

注意：本章主要讨论线性回归。

在这个例子中，将使用学生投入的学习时间预测他/她的学习成绩。因此，Freshmen_Score 可被看作因变量，而 NoOfHours 可被看作预测变量。这是一个简单线性回归的例子，因为其只有一个预测变量和一个因变量。

因此，用来预测学生成绩的回归模型可以被表示为

$$\text{Freshmen\_Score}=\beta_0+(\beta_1\times\text{NoOfHours})+\varepsilon$$

8. 使用 lm()函数创建线性模型

下面计算系数：

(a) Intercept　　(b)HS$NoOfHours

```
>x<-HS$NoOfHours
>y<-HS$Freshmen_Score
>n<-nrow(HS)
>xmean<-mean(HS$NoOfHours)
>ymean<-mean(HS$Freshmen_Score)
>xiyi<-x * y
>numerator<-sum(xiyi)-n * xmean * ymean
>denominator<-sum(x^2)-n * (xmean^2)
>b1<-numerator / denominator
>b0<-ymean-b1 * xmean
>b1
```

```
[1] 6.884848
>b0
[1] 44.93939
```

使用 lm( )构建模型。在这里,HS $ Freshmen_Score 是响应变量或目标变量,HS $ NoOfHours 是预测变量。图 5.3 是对模型的可视化表示。

```
>model_HS <- lm(HS$Freshmen_Score ~ HS$NoOfHours)
>model_HS
Call:
lm(formula = HS$Freshmen_Score ~ HS$NoOfHours)
Coefficients:
(Intercept)  HS$NoOfHours
     44.939         6.885
>summary(model_HS)
Call:
lm(formula = HS$Freshmen_Score ~ HS$NoOfHours)
Residuals:
    Min       1Q   Median       3Q      Max
-4.3636  -1.5803  -0.3727  0.7712  6.0788
Coefficients
              Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept)    44.9394      3.4210   13.136  1.07e-06 ***
HS$NoOfHours     6.8848      0.7626    9.028  1.81e-05 ***
---
Signif. codes: 0 '***" 0.001 '**' 0.01 '* ' '0.05' '.' 0.1 '' 1
Residual standard error: 3.463 on 8 degrees of freedom
Multiple R-squared: 0.9106, Adjusted R-squared: 0.8995
F-statistic: 81.51 on 1 and 8 DF, p-value: 1.811e-05
>plot(HS$NoOfHours, HS$Freshmen_Score, co1="blue", main =
"Linear    Regression",
+abline(lm(HS$Freshmen_Score ~ HS$NoOfHours)), cex = 1.3, pch = 16, xlab = "No of
hours of study",
+ylab = "Student Score")
```

9. 对输出的说明

输出中的第一项是 formula(lm(formula = HS $ Freshmen_Score ~ HS $ NoOfHours)),R 使用它拟合数据。lm( )是 R 中的一个线性模型函数,用于构建一个简单回归模型。HS $ NoOfHours 是预测变量,HS $ Freshmen_Score 是目标/响应变量。

模型输出中的下一项描述了 residuals。实际观测值(例子中的 HS $ Freshmen_Score)和模型预测的响应值之间的差称为 residuals。模型输出的 residuals 部分将其汇总成 5 点,即最小值、1Q(第一四分位数)、中值、3Q(第三四分位数)和最大值。在评估模型对数据的拟合程度时,应该在均值为 0 上寻找这些点之间的对称分布值,如表 5.2 所示。

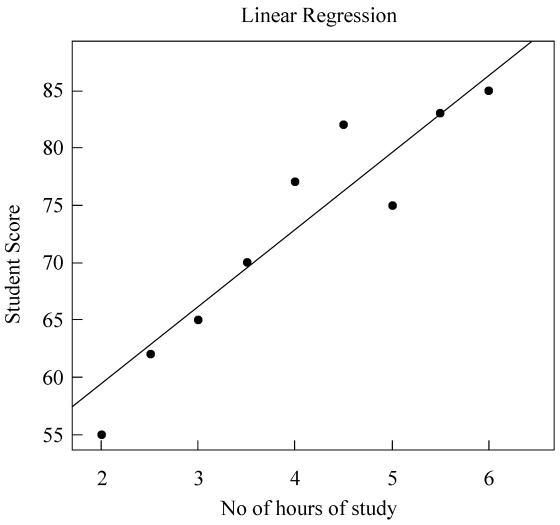

图 5.3　线性回归图

表 5.2　模型输出

小时数	成绩	预测值	残差值 （实际值－估计值）
2	55	58.70909	－3.70909
2.5	62	62.15152	－0.15152
3	65	65.59394	－0.59394
3.5	70	69.03636	0.96364
4	77	72.47879	4.52121
4.5	82	75.92121	6.07879（maximum value）
5	75	79.36364	－4.36364（minimum value）
5.5	83	82.80606	0.19394
6	85	86.24848	－1.24848
6.5	88	89.69091	－1.69091

要计算 5 个汇总点，需要将数据以升序排列。

（－4.36364，－3.70909，－1.69091，－1.24848，－0.59394，－0.15152，0.19394，0.96364，4.52121，6.07879）

最小值：－0.436364。

1Q：在位置 3.25 处。

获取 3.5 位置处的值＝（－1.69090－1.24848）/2＝－1.46969。

获取 3.25 位置处的值＝（－1.69090－1.46969）/2＝－1.580295。

中值：（－0.59394 －0.15152）/2＝－0.37273（中值在位置 5.5 处）。

3Q：在位置 7.75 处。

获取位置 7.5 处的值＝(0.19394＋0.96364)/2＝0.57879。

获取位置 7.75 处的值＝(0.57879＋0.96364)/2＝0.771215。

最大值：6.07879。

下一部分描述了模型的系数(coefficient)。从理论上来讲，在简单线性回归中，系数是两个未知的常数，表示线性模型中的截距(intercept)和斜率(slope)。

10. 系数：估计值

估计值(estimate)包含两行。第一行是截距，所有预测变量 X＝0 时是响应变量 Y 的均值。请注意，只有当模型中的每个 X 的实际值为 0 时均值才有用。系数的第二行是斜率或者 HS_NoOfHours 对 Freshmen_Score 的影响。模型中的斜率证明了 NoOfHours 每增加一小时，Freshmen_Score 就会提高 6.8848 分。

11. 系数：标准误差

标准误差(standard error)度量了从响应变量的实际平均值估计的系数变化的平均量。理想的情况是：相对于其系数，这应该是一个较小的数。

12. 系数：t-value

系数 t-value 是指系数估计值距离 0 有多少标准差(standard deviation)的度量，它应该是远离 0 的，因为可以声明 HS_NoOfHours 和 Freshmen_Score 之间的关系是存在的。t-value 是系数除以标准误差((44.9394/3.4210)＝13.1363)。一般来说，t-value 也用来计算 p-value。

13. 系数：Pr(＞t)

在模型输出中发现的字母缩写 pr(＞t)是观测到的任何值等于或大于 t 的概率。偶然情况下的一个小的 p-value 表明不太可能观察到预测变量(HS_NoOfHours)和响应变量(Freshmen_Score)之间的关系。典型的情况是：5%及以下的 p-value 是一个很好的临界点。

注意：signif.Codes 与每个估计值相关联。

3 颗星(或星号)表示一个非常显著的 p-value。

由***标明的系数是 p-value＜0.001 的系数。

由**标明的系数是 p-value＜0.01 的系数。

14. 残差标准误差(Residual Standard Error)

残差标准误差是对线性回归拟合质量的度量。从理论上而言，每个线性模型都假设包含一个误差项 E，它可以防止从预测变量中完美地预测出响应变量。下面计算均方根误差(Root Mean Squared Error，RMSE)，它是均方残差的平方根。

考虑如表 5.3 所示的学生数据集。

表 5.3　学生数据集

小时数	成绩	预测值	残差值（实际值－估计值）	残差平方和
2	55	58.70909	－3.70909	13.75734863
2.5	62	62.15152	－0.15152	0.02295831
3	65	65.59394	－0.59394	0.352764724
3.5	70	69.03636	0.96364	0.92860205
4	77	72.47879	4.52121	20.44133986
4.5	82	75.92121	6.07879	36.95168786
5	75	79.36364	－4.36364	19.04135405
5.5	83	82.80606	0.19394	0.037612724
6	85	86.24848	－1.24848	1.55870231
6.5	88	89.69091	－1.69091	2.859176628

注意：以下部分将会展示利用 predict()和 resid()函数计算预测值和残差值的方法。

残差标准误差＝(残差的平方之和/模型自由度)的平方根。

- 残差的平方之和＝95.95154715。
- 模型的自由度＝8。

自由度是由数据集的行数－列数或变量数指定的，student 数据集中有 10 行和 2 列 HS$NoOfHours 和 HS$Freshmen_Score，即 10－2＝8。

可以使用 df.residual()函数在 R 中计算自由度。

```
>df.residual (model_HS)
[1] 8
```

残差标准误差＝(95.95154715/8)的平方根。

残差标准误差＝(11.99394339)的平方根。

残差标准误差＝3.463227309。

15. 拟合优度(Multiple R-squared)和修正的拟合优度(Adjusted R-squared)

拟合优度 R-squared(R2R2)统计提供了模型和实际数据拟合程度的一种度量，它的表示形式为方差比例(proportion of variance)。R2R2 是一种用于预测变量(HS_NoOfHours)和响应/目标变量(Freshmen_Score)之间的线性关系的度量，它总是介于 0 和 1 之间(也就是说，一个接近于 0 的数字代表一种不能很好地解释响应变量中方差的回归，而一个接近 1 的数字则解释了响应变量中观测到的方差)。

拟合优度又称确定系数，它给出了有多少数据点落在回归方程形成的直线内的观点。该系数越大，当绘制数据点和直线时，直线会穿过更高比例的数据点。如果系数是 0.80，那么 80%的点应该落在回归线内。值 1 和 0 表明回归线分别表示全部数据或没有数据。一个较大的系数表明对观测值更好的拟合。