第六章 线性回归

章节引入

回归分析（regression analysis）是统计分析中最重要的思想之一。大部分的统计分析问题，都可以被视为一个回归问题。首先，回归分析建模的目标是因变量Y。这里的因变量，一般可以随着一些因素的改变而变化。在实际应用中，因变量刻画的是业务的核心诉求，是科学研究的关键问题。这里我们举几个在业务问题中常用到的因变量的例子：

【例 1】在求职市场上，求职者最关心的莫过于岗位薪资。若能有针对性的提高相应技能，升职加薪不再是梦。因此，这里的因变量 Y 是岗位薪资，它的取值是连续的。

【例 2】截至 2016 年 5 月 25 日的北京住宅年内交易数据显示，北京市已经全面进入二手房时代。二手房定价是二手房交易过程中重要的环节之一。若能根据住房的特征，更准确地估计价格，住房业主将会获得更准确的市场定位。因此，这里的因变量 Y 是房价，它的取值也是连续的。

确定了统计建模目标后，还需要知道有哪些可能的因素会影响因变量。在回归分析中，把这些起到解释作用的变量称为自变量 X（或解释性变量）。例如，在例 1 中，想要预测岗位薪资，则可能有影响的自变量包括岗位的地区、公司的类别与规模、应聘者的工作经验等；在例 2 中，想要预测住房的单位面积房价，则可能有影响的自变量包括住房所在城区、卧室数量、房屋面积、楼层、是否为学区房等。

对于因变量与自变量之间的回归关系，我们特别关注以下问题：

（1）哪些自变量对 Y 具有显著的解释能力？如何选出这些自变量？

（2）这些有用的自变量与 Y 的相关关系是正是负？

（3）每个自变量在回归关系中的权重大小是多少？

明确了这些问题，也就对回归分析有了深入的理解。按照因变量形式的不同，回归模型又被划分成不同的类型。针对连续型因变量，可以主要通过线性回归模型建模。

案例引入

背景介绍

随着“互联网+”和大数据时代的到来，越来越多的数据科学公司如雨后春笋般涌现。传统行业也面临着“互联网+”时代下的创新转型，对于数据分析及相关领域有大量人才需求，各行各业与数据分析相关的招聘岗位越来越多。

在数据分析相关岗位的招聘中，合理定位岗位薪资，找出与岗位薪资挂钩的特殊技能尤其关键。在市场层面上，可以了解数据分析人才市场现状，合理化市场资源配置；在公司层面上，可以为公司招聘提供借鉴，为数据分析人才定制薪资提供参考；对于应聘者而言，能够更科学地进行职业测评，实现准确地自我定位；对于高校来说，能够明确学生的培养方向，优化应用统计及数据分析方向人才培养方案。

数据介绍

本章使用数据分析岗位招聘薪酬数据集。该数据收集自各大招聘网站发布的数据分析岗位招聘的相关信息，共包含了 6682 条岗位招聘数据。数据集的每一列分别对应：岗位薪资、是否要求掌握 R、SPSS、Excel、Python、MATLAB、Java、SQL、SAS、Stata、EViews、Spark、Hadoop、公司类别、公司规模、学历要求、工作经验、公司地点。详细的数据说明表如表 6-1 所示。

表 6-1：数据分析岗位招聘薪酬数据集变量说明表

变量类型	变量名称		详细说明	取值范围
因变量	薪资		单位：元 / 月	1500-5000 元
自变量	软件要求	R	共 2 个水平	0（不要求），1（要求）
		SPSS	共 2 个水平	
		Excel	共 2 个水平	
		Python	共 2 个水平	
		MATLAB	共 2 个水平	
		Java	共 2 个水平	
		SQL	共 2 个水平	
		SAS	共 2 个水平	

	Stata	共 2 个水平	
	EViews	共 2 个水平	
	Spark	共 2 个水平	
	Hadoop	共 2 个水平	
	公司类别	共 6 个水平	合资、外资、民营公司等
	公司规模	共 6 个水平	少于 50 人、50-500 人等
	学历要求	共 7 个水平	无、中专、高中、大专等
	工作经验	单位：年	0-10 年
	地区	共 2 个水平	北上深、非北上深

描述分析

在回归分析中，我们关注的问题是：哪些因素会影响薪资水平？在目前竞争激励的数据分析人才招聘市场上，这是求职者们首先关注的问题。在上一章方差分析中，我们发现，掌握 R 语言的岗位薪资显著高于不要求掌握 R 语言的岗位。那么，是否还有其他因素（例如：学历水平、其他编程语言等）与岗位薪资有关？它们与薪资水平的相关程度如何？这正是接下来回归分析将要探讨的问题。在正式进行回归分析之前，先对数据进行粗略的描述分析，简要观察各个自变量和因变量之间是否有一定的关系。

（1）因变量：岗位薪资

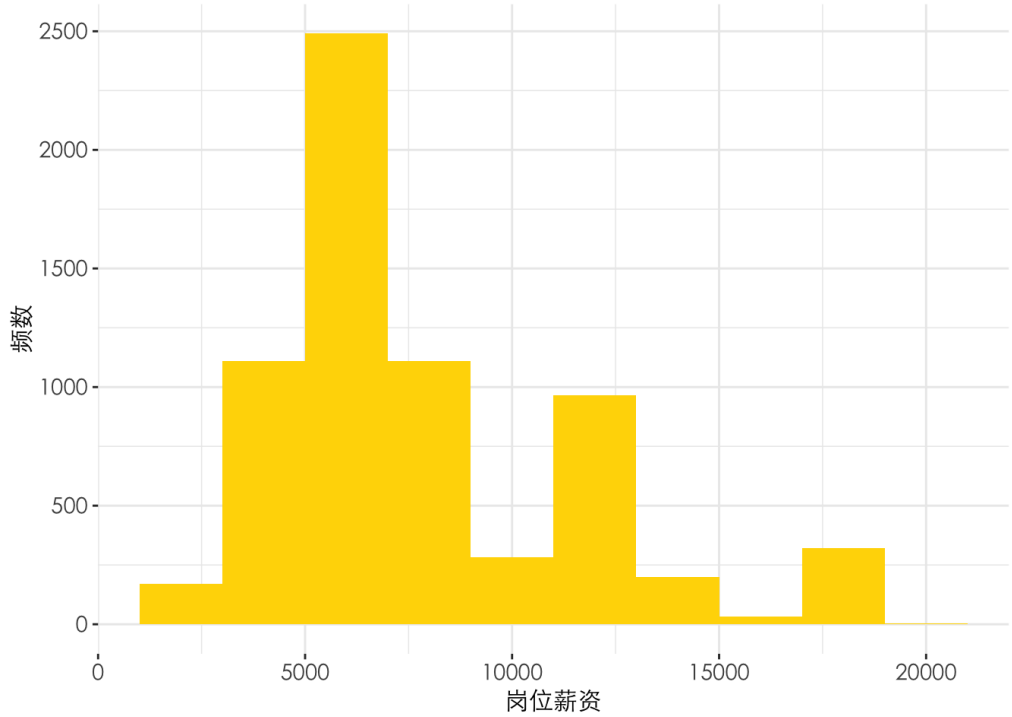


图 6-1：岗位薪资频数直方图

该数据中的因变量薪资呈右偏分布，如图 6-1 所示。最高的月薪为 19999.5 元/月，对应的岗位是一个规模为 1000-5000 人的国企，这个岗位要求申请人有 2 年的工作经验。从整体情况来看，有 75%的岗位月薪低于 10000 元。

（2）自变量：学历要求

一般来说，学历要求也是影响薪资水平的一个关键因素。从图 6-2 可以看到，在五种学历要求中，研究生学历的薪资中位数最高，达到了 8999.75 元，对应的对数薪资为 9.10（图 6-2）；其次为本科学历的岗位，达到了 8999.50 元；薪资水平最低的为中专学历，月薪仅有 5249.50 元。水平最高的学历岗位比最低的月薪高出了 3750.25 元，但是仅根据描述分析，我们仍然不能说明学历对薪资有统计意义上的显著影响。

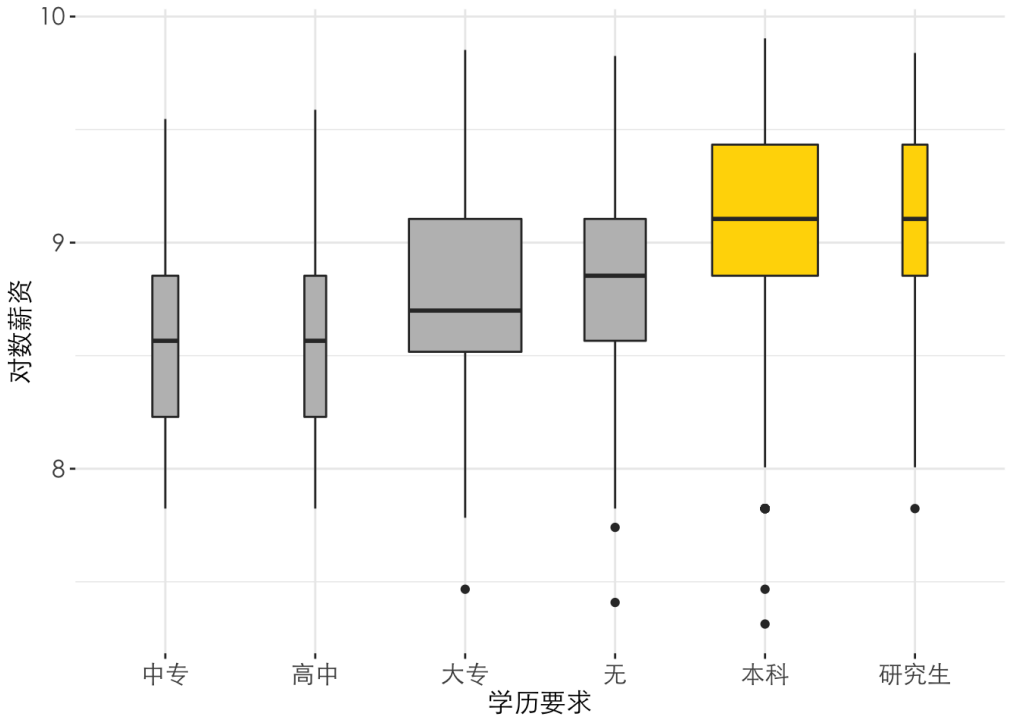


图 6-2：对数薪资与学历要求箱线图

（3）自变量：软件要求

除了 R 语言以外，掌握其他编程语言是否对薪资有影响呢？图 6-3 以 Python 和 SPSS 为例进行描述分析。从数目上看，要求掌握 Python 的岗位占比为 4.4%，要求掌握 SPSS 的岗位占比约为 8.0%。从箱线图中可以看出，要求掌握这两种软件的薪资中位数均高于不要求掌握该软件的岗位。

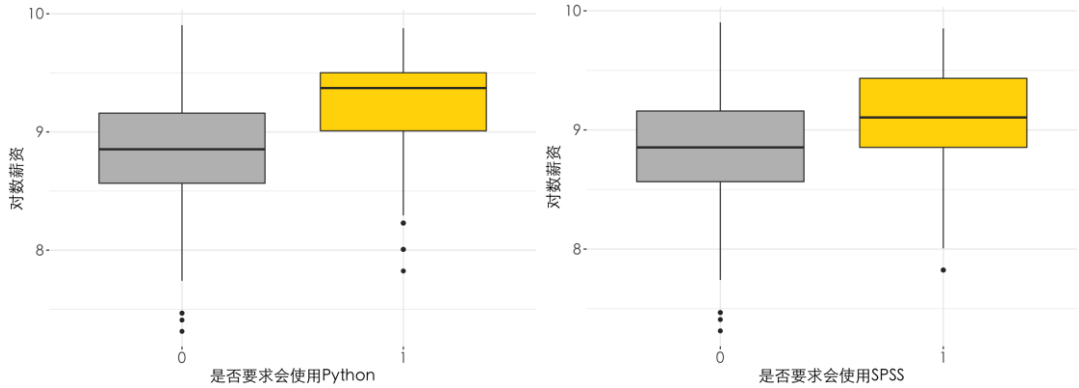


图 6-3：对数薪资与软件要求箱线图

本章难点

（1）了解回归分析的基本思想，掌握线性回归模型的形式，对模型有清晰的理

解；

（2）掌握回归模型参数的估计方法，能在实际案例中对参数进行合理的解读；掌握回归系数与回归方程的假设检验、回归模型的诊断等问题；

（3）了解模型变量选择的基本思想；能够使用 R 语言实现线性回归模型的建立、诊断、解读与预测。

6.1 模型形式

如何研究因变量和自变量之间的关系？如果不做任何的模型假设，这是一个很难完成的任务。简单而言，统计模型就是解读因变量和自变量相关关系的函数表达式。假设我们有p个自变量$X_1, X_2, \cdots, X_p$，考虑模型 1：

$$\text{模型 1：} Y = f(X) = f(X_1, \cdots, X_p)$$

其中$f(\cdot)$是一个提前设定的函数形式。但是，以上模型往往不能很好地解释数据。这是因为模型 1 刻画了一种确定的函数关系：给定自变量取值，因变量是唯一确定的。但事实上，这是一个非常不合理的性质。例如，对于岗位描述一致（即：具有相同的 X）的两个职位，它们的薪水也可能存在差异。这是因为数据中包含的自变量信息不足以完全刻画岗位薪资的水平。为刻画这种不确定性，可以引入一个噪音项ε。一般认为噪音项与自变量互相独立，完全随机，但也能够对因变量产生影响。因此，可以将模型 1 稍加修正，使其包含噪音项ε：

$$\text{模型 2：} Y = f(X, \varepsilon)$$

在引入了噪音项之后，模型 2 对现实数据具有更好的刻画能力。

应该考虑什么样的模型形式呢？在所有的函数形式中，线性函数具有易于解释、稳定性高的特点。如果使用线性函数，模型 2 可以具体表达成一个自变量及噪音项线性组合的形式，这就是线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

其中，$\beta_0, \beta_1, \ldots, \beta_p$是 p+1 个未知参数，$\beta_0$为常数项，$\beta_1, \ldots, \beta_p$称为回归系数。Y 为因变量，$X_1, X_2, \ldots, X_p$为$p$个可观测的预测性变量，称为自变量，$\varepsilon$为随机误差项。可以看到，从模型形式上，自变量与随机误差项都对 Y 产生影响。同时，线性回归模型具有很好的可解释性，是在实际应用中得以广泛应用的统计模型之一。

6.2 模型理解

给定线性回归模型形式，本节介绍如何解读线性回归模型。设Y表示岗位薪资，X_1表示工作经验，X_2表示掌握 R 语言技能（0-1 变量），线性回归模型写为：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

首先，由于随机误差项ε是不被自变量X所包含的，因此它是一个无法控制的因素，也就没有精确预测的可能。一般而言，假设$E(\varepsilon) = 0$且ε与自变量 X 互相独立。因此，以下等式成立：

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

对于一个回归模型，我们需要关注的是未知参数$\beta_0, \beta_1, \beta_2$的解读。

6.2.1 回归系数的理解

回归系数即$\beta = (\beta_0, \beta_1, \beta_2)^\top$，具体分为常数项$\beta_0$和变量系数$\beta_1, \beta_2$。

（1）β_0 的解读

β_0又称为截距项。首先注意到：

$$E(Y|X_1 = 0, X_2 = 0) = \beta_0$$

这说明β_0是所有自变量都取值为 0 时，因变量的期望值。在岗位招聘数据中，自变量都取 0 意味着一个工作经验要求为 0 年，且不需要掌握 R 语言的岗位平均薪资为β_0，反映了不考虑自变量的情况下对因变量的平均预期。

（2）β_1, β_2的解读

接着，我们来讨论β_1, β_2，由于它们的性质是一样的，这里主要介绍β_1的理解。按照模型的形式，$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$，这是在给定自变量$X$的情况下，因变量$Y$的期望值。如果保持$X_2$不变，而对$X_1$增加一个单位，即$X_1$变成$X_1 + 1$，则会产生一个新的因变量值，记为$Y^\star$，并且$E(Y^\star) = \beta_0 + \beta_1(X_1 + 1) + \beta_2 X_2$。因此，这个变化对因变量期望值的改变量是$\Delta = E(Y^\star) - E(Y) = \beta_1$。由此可见，$\beta_1$反映的是在其他解释变量不变的情况下，因变量期望值对X_1的敏感度。对于β_2的理解也是类似的。

若$\beta_1 = 0$，那么自变量X_1不再参与到因变量Y的生成机制中。此时，自变量

X_1对 Y 的期望值没有任何影响，即X_1工作经验这个因素对于我们关心的薪资完全不重要。若$\beta_1 > 0$，那么在控制其他自变量的情况下，薪资与工作经验是正相关的；若$\beta_1 < 0$，在控制其他自变量的情况下，薪资和工作经验是负相关的。

需要注意的是，我们在理解任何一个回归系数β_i时，一定要在控制其他自变量不变的前提下进行解读。如果仅仅说X_i取值增加一个单位，Y的期望值就会增加β_i是不正确的。因为X_i变化时，其他自变量也可能在变化，此时，因变量Y的期望值变化量就不再是β_i。

6.2.2 定性变量转换及回归系数理解

在实际数据中，变量信息往往是复杂的，自变量既包括定量变量，也包括定性变量。为学习定性变量在回归分析中的处理方式及解读，在前半部分讨论的基础上，再引入一个新的自变量：公司类别X_3（多水平定性变量）。其中，工作经验对应回归系数在上一节中已经详细解读，下面重点讨论定性变量X_2（是否掌握 R 语言）和X_3（公司类别）的处理和解读。

1. X_2为二分类变量，对应的取值 1 表示掌握 R 语言，0 表示不掌握 R 语言。此时线性回归模型系数可以进一步解释为，自变量取分类“1”时，因变量的值平均比自变量取分类“0”时高多少。例如：自变量“掌握 R 语言”对应的系数为β_2，说明要求掌握 R 语言的岗位，薪资平均比不要求掌握 R 语言的岗位高出β_2。

2. X_3为多水平定性变量。对于水平定性变量在回归分析前需要进行预处理。一般可以将定性变量的各水平转换为一组哑变量。哑变量是一种特殊的变量，只有 0 和 1 两个取值。一个多水平的定性变量可以用一组哑变量来表示。例如，这里可以将公司类别（取值为 6 个水平）转换为 5 个哑变量，分别为：是否为合资、是否为外资、是否为上市公司、是否为民营公司、是否为创业公司。如果上述 5 个哑变量取值全部为 0，那么公司类型就是国企（记为基准组）。一般来说，若分类变量取值为k个互斥的水平，则可以将这个分类变量转换为$k-1$个哑变量，分别对应其中$k-1$个取值，未转换为哑变量的水平则定义为基准组。因此，经过转换之后的线性回归表达式可以写为：

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{31} X_{31} + \beta_{32} X_{32} + \cdots + \beta_{35} X_{35} + \varepsilon$$

其中，$X_{31},\cdots,X_{35}$为X_3转换的哑变量。

在 R 中可以直接将分类变量设置为因子（factor）类型，这样在回归分析时程序将自动完成哑变量的转换，而不需要手动转换为哑变量，通过设置因子水平（level）可以自定义基准组。

```
# 转换为 factor 型变量，地区以河北为基准，公司类别以国企为基准，公司规模以少于 50 人为基准，学历以无为基准

jobinfo$公司类别 <- factor(jobinfo$公司类别, levels = c("国企","合资","外资","上市公司","民营公司","创业公司"))

jobinfo$公司规模 <- factor(jobinfo$公司规模, levels = c("少于 50 人","50-500 人","500-1000 人","1000-5000 人","5000-10000 人","10000 人以上"))

jobinfo$学历要求 <- factor(jobinfo$学历要求, levels = c("无","中专","高中","大专","本科","研究生"))
```

线性回归模型为自变量X_3每一个取值分别拟合一个系数。自变量公司类别共有 6 个取值，分别为国企、民营公司、上市公司、合资、外资和创业公司。如果取基准组为“国企”，则其他 5 个哑变量的回归系数表示取该分类水平时，因变量的值平均比**基准组（国企）**高多少。比如“民营公司”对应的系数为β_{31}，说明民营公司的平均薪资比国企的平均薪资高β_{31}。

6.2.3 交互项的解读

当自变量X_i变化时，其他自变量也可能在变化。例如一般要求掌握 R 语言编程的公司，可能同时也要求有一定的工作经验，此时如果只在控制工作经验不变的情况下解读掌握 R 语言对薪资的影响，得出的结论可能不够合理。因此引入交互项这一概念，用来描述两个变量的交互关系。引入交互项的线性回归模型可以写作：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

称其中的X_1X_2项为变量“工作经验”与“是否掌握 R 语言”的交互项。此时有：

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$

$$= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \varepsilon$$

由以上模型表达式可以看出，对于不要求掌握 R 语言的岗位（$X_2 = 0$），工作经验每增加一年，薪资平均增加β_1；对于要求掌握 R 语言的岗位（$X_2 = 1$），工作

经验每增加一年，薪资平均增加$\beta_1+\beta_2$。这说明是否掌握 R 语言，改变了因变量薪资对自变量工作经验的敏感度。

特别要提醒的是，在实际的数据分析中，如果加入过多交互项，会导致模型复杂，稳定性差。因此，交互项不宜加入过多，是否加入交互项需要结合数据的描述分析与案例背景进行讨论。

6.2.4 σ^2的理解

最后来讨论$Var(\varepsilon)=\sigma^2$的含义。根据模型形式和概率论知识，可以得到$\sigma_Y^2=Var(\beta_1X_1+\beta_2X_2)+\sigma^2$，这说明因变量$Y$的波动程度由自变量和随机误差项的波动程度共同组成，且$\sigma_Y^2\geq\sigma^2$，因此，可以通过比较σ_Y^2和σ^2的相对大小来判断随机误差项在该模型中的作用。当$\sigma^2/\sigma_Y^2\approx1$时，所有的自变量几乎对因变量的波动性毫无解释作用，此时的模型是一个不够理想的情况；相反，当$\sigma^2/\sigma_Y^2\approx0$时，随机误差项对因变量的影响非常小，自变量可以非常充分地解释因变量的波动，此时模型是比较理想的。

6.3 基本假定

对于一个实际问题，例如案例引入部分介绍的预测岗位薪资的问题，可以收集到n个样本观测数据。将这n个岗位的信息记为$(x_{i1},x_{i2},\dots,x_{ip};y_i),i=1,2,\dots,n$，则线性回归模型可以表示为：

$$\begin{cases}y_1=\beta_0+\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_px_{1p}+\varepsilon_1\\ y_2=\beta_0+\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_px_{2p}+\varepsilon_2\\ \cdots\cdots\\ y_n=\beta_0+\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_px_{np}+\varepsilon_n\end{cases}\tag{6.1}$$

这个模型写成矩阵的形式为：

$$y=X\beta+\varepsilon$$

其中，

$$y=\begin{bmatrix}y_1\\y_2\\\vdots\\y_n\end{bmatrix}\quad X=\begin{bmatrix}1&x_{11}&x_{12}&\cdots&x_{1p}\\1&x_{21}&x_{22}&\cdots&x_{2p}\\\vdots&\vdots&\vdots&&\vdots\\1&x_{n1}&x_{n2}&\cdots&x_{np}\end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

矩阵X是$n \times (p+1)$维的矩阵，也称为资料矩阵或设计矩阵（design matrix）。

为了进行模型的参数估计，在线性回归中有如下基本假定：

1. 自变量矩阵满秩。

解读：当 X 矩阵不满秩时，存在 X 的某一列可以被其他列通过线性组合的形式表示。也就是说，这一列变量的信息可以被其他变量完全“替代”，因此，可以对这样的列进行剔除，使 X 满足满秩的条件。一般认为 X 矩阵是非随机的。

2. 随机误差项ε_i独立同分布。它们的均值是 0，方差相同，即：

$$E(\varepsilon_i) = 0, \qquad var(\varepsilon_i) = \sigma^2$$

解读：当$E(\varepsilon_i) \neq 0$时，可以令$\varepsilon_i = \ \varepsilon_i - \ E(\varepsilon_i)$，同时变化截距项 $\beta_0 = \beta_0 + E(\varepsilon_i)$。经过变化后仍可以满足扰动项期望为 0 的假设。同时，为了便于技术处理，我们假设噪音项独立，且方差一致。如果不满足这个假定，则可以通过修正模型形式（例如：通过时间序列建模）进行改善。

3. 随机误差项满足正态分布假设：$\varepsilon_i \sim N(0, \sigma^2)$。

解读：实际数据很难满足完全正态性的假设。一般这一假设可以放宽，例如要求扰动项服从对称形式的分布。如果观察到因变量呈现右偏分布，常用的操作是对因变量取对数，使得其在取对数后近似呈现对称分布。

由上述假定和多元正态分布的性质可知，随机向量y服从n维正态分布，回归模型（6.1）满足：

$$E(y) = X\beta$$

$$Var(y) = \sigma^2 I_n$$

因此，

$$y \sim N(X\beta, \sigma^2 I_n)$$

6.4 回归参数的估计

既然回归分析是围绕各个自变量的系数展开的，那么如何确定这些未知参数的值呢？一般来说，我们可以通过观测到的有限数据，对未知参数给出合理的估

计。这里我们介绍最常用的两种方法：普通最小二乘估计和最大似然估计。

6.4.1 普通最小二乘估计

对于回归模型$y = X\beta + \varepsilon$，希望找到一个回归系数β，使得$X\beta$与 y 的某种距离足够小。这就是最小二乘法的想法。对于任意一个样本i，定义估计的因变量为$\hat{y} = X\beta$，则可以计算因变量与自变量之间欧氏距离的平方：

$$(y_i - \hat{y})^2 = \left(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p\right)^2$$

其中$y_i - \hat{y}$称为残差。如果把所有样本的残差平方相加，则可以得到**残差平方和**：

$$RSS(\beta) = \Sigma_{i=1}^{n}\left(y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p\right)^2$$

最小二乘法即寻找$\beta_0, \beta_1, \ldots, \beta_p$的估计值 $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$，使得**残差平方和**达到最小的估计方法，即$\hat{\beta} = \text{argmin}_{\beta_0,\beta_1,\ldots,\beta_p} RSS(\beta)$。$\hat{\beta}$具备许多优良的性质，例如，在满足模型设定的前提下，$\hat{\beta}$是β的一个无偏估计。同时，当收集到更多样本时，$\hat{\beta}$ 的估计也将更为精确。

最小二乘估计是求解线性回归问题中常用的估计方法，但不是唯一的方法。例如，在计算残差和时，可以不考虑欧氏距离的平方，而采用绝对值距离，即 $\Sigma_{i=1}^{n}|y_i - \beta_0 - \beta_1 x_1 - \cdots - \beta_p x_p|$ 。但是，相对于使用绝对值距离，最小二乘估计的计算性能更好。这是因为最小二乘估计可以显式求解。对最小二乘法的目标函数进行整理，可得：

$$RSS(\beta) = (y - X\beta)^{\mathsf{T}}(y - X\beta)$$

为求解目标函数的最小值，可以对残差平方和$RSS(\beta)$关于β求偏导，并令其等于 0：

$$\frac{\partial RSS(\beta)}{\partial \beta} = -2X^{\mathsf{T}}(y - X\beta) = 0$$

可以显式解得

$$\hat{\beta} = (X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y$$

由于 X 是满秩矩阵，则$X^{\mathsf{T}}X$为正定矩阵，回归拟合值为

$$\hat{y} = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}y = Hy$$

其中$H = X(X^{\mathsf{T}}X)^{-1}X^{\mathsf{T}}$称为帽子矩阵。

最后我们再讨论一下如何对噪音项的方差σ^2进行估计。尽管σ^2的估计不会引入到β的估计中，但是，它会对$\hat{\beta}$的统计推断产生影响。在模型的假设部分，我们提到了$E(\varepsilon)=0$，所以有$E(\varepsilon^2)=\sigma^2$。因此，随机误差项的方差$\sigma^2$可以通过如下方式估计：

$$n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 = n^{-1}\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1X_{i1}-\beta_2X_{i2}-\cdots-\beta_pX_{ip}\right)^2$$

由于$\beta=\left(\beta_0,\beta_1,\cdots,\beta_p\right)^{\mathrm{T}}$是未知参数，因此用最小二乘估计$\hat{\beta}=\left(\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_p\right)^{\mathrm{T}}$来代替：

$$n^{-1}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1X_{i1}-\hat{\beta}_2X_{i2}-\cdots-\hat{\beta}_pX_{ip}\right)^2$$

尽管以上估计量在大样本的前提下具有优良的性质，但它仍是一个有偏估计（即$E(\hat{\sigma}^2)\neq\sigma^2$）。为了获得$\sigma^2$的无偏估计量，可以对这个估计量修正如下：

$$\hat{\sigma}^2=\frac{1}{n-p-1}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1X_{i1}-X_{i2}-\cdots-\hat{\beta}_pX_{ip}\right)^2=\frac{RSS}{n-p-1}$$

可以证明，在正态假设$y\sim N(X\beta,I_n\sigma^2)$下，估计量 $\hat{\beta}$和 $\hat{\sigma}^2$有如下性质：

（1）$\hat{\beta}$服从正态分布 $N(\beta,(X^{\mathrm{T}}X)^{-1}\sigma^2)$

（2）$\frac{RSS}{\sigma^2}$服从分布$\chi^2(n-p-1)$

6.4.2 极大似然估计

除了最小二乘估计，也可以采用极大似然估计的方法，对回归模型的参数进行估计。由于ε服从正态分布，则y的概率分布为：

$$y\sim N(X\beta,\sigma^2I_n)$$

因此，可以写出似然函数及其对数形式：

$$L(\beta)=(2\pi)^{-n/2}(\sigma^2)^{-n/2}\exp\left(-\frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\beta)^{\mathrm{T}}(\mathrm{y}-\mathrm{X}\beta)\right)$$

$$\ln L(\beta)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln(\sigma^2)-\frac{1}{2\sigma^2}(\mathrm{y}-\mathrm{X}\beta)^{\mathrm{T}}(\mathrm{y}-\mathrm{X}\beta)$$

最小化对数似然函数等价于最小化$(y-X\beta)^{\mathrm{T}}(y-X\beta)$，因此极大似然估计的结果与普通最小二乘法对于回归系数估计的结果是相同的。

6.5 假设检验

使用线性回归方程拟合随机变量y与变量$X_1, \dots X_p$之间的关系后，是否可以确定它们之间一定具有线性关系呢？实际上，由于存在随机性，我们不能断定变量之间的关系，因此在求解出线性回归方程后，还需要对回归方程进行显著性检验。本节将介绍两种检验的方法，一个是回归系数显著性的 t 检验，另一个是回归方程显著性的 F 检验。

6.5.1 回归系数的 t 检验

当X_j对因变量实际没有显著影响时，$\beta_j = 0$，因此可以建立假设：

$$H_0: \ \beta_j = 0 \quad vs \quad H_1: \ \beta_j \neq 0, \quad j = 1,2, \dots, p$$

由于$\hat{\beta} \sim N(\beta, \sigma^2 (X^\top X)^{-1})$，记$(X^\top X)^{-1} = (c_{ij}) \quad i,j = 1,2, \dots, p$，于是有

$$E(\hat{\beta}_j) = \beta_j, \ \mathrm{var}(\hat{\beta}_j) = c_{jj}\sigma^2$$

$$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2), j = 0,1,2, \cdots, p$$

得到 t 统计量

$$T_j = \frac{\hat{\beta}_j}{\sqrt{c_{jj}\hat{\sigma}^2}}$$

联系在第五章介绍的内容，在H_0成立的条件下，以上 t 统计量服从自由度为 n-p-1 的 t 分布，即$T_j \sim t(n-p-1)$。因此，可以利用 t 统计量对回归系数的显著性进行检验。

6.5.2 回归方程的 F 检验

除了考虑某一个自变量的影响，还需要对回归方程整体的意义进行考察。对线性回归方程的显著性检验即判断自变量$X_1, \dots, X_p$整体是否对因变量有显著性影响。因此，原假设与备择假设为

$$H_0: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0 \quad vs \quad H_1: \exists j \ (1 \leq j \leq p) \ s.t. \quad \beta_j \neq 0$$

记回归平方和（解释平方和）为$ESS = \Sigma_i(\hat{y}_i - \bar{y})^2$，总平方和$TSS = \Sigma_i(y_i - \bar{y})^2$。可以证明，

$$\sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

即$TSS = ESS + RSS$。在H_0成立的情况下，构建满足$F(p, n-p-1)$的 F 统计量：

$$F = \frac{ESS/p}{RSS/(n-p-1)} \sim F(p, n-p-1)$$

因此，可利用 F 统计量对回归方程的总体显著性进行检验。

6.6 模型评价

为评价建立的线性回归模型在多大程度上反映了模型的拟合程度，我们采用样本决定系数R^2来评价回归模型的拟合效果。样本决定系数R^2定义如下：

$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = \frac{\mathrm{ESS}}{\mathrm{TSS}}$$

其中，

残差平方和：$RSS = \Sigma_{i=1}^{n}(y_i - \hat{y}_i)^2$

回归平方和：$ESS = \Sigma_{i=1}^{n}(\hat{y}_i - \bar{y})^2$

总离差平方和：$TSS = \Sigma_{i=1}^{n}(y_i - \bar{y})^2$

R^2表达式由分子分母组成，其中分子可以表示因变量Y中我们能够了解和把握的部分，即Y的信息中被X解释的部分；分母表示 Y 的总的波动程度，既包含能够被自变量 X 解释的部分（回归平方和），也包括不能够被自变量 X 解释的部分（残差平方和）。

但是，R^2有一个缺点，即鼓励过度拟合。简单地说，只要增加自变量X的个数，R^2值就会增大。这使得R^2大的模型往往变量个数也较多，因此模型复杂度较高。为避免得到过于复杂的模型，在模型评价时一般更加经常使用的是调整后的R^2：

$$\text{Adjusted } R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$$

当模型的复杂度提高，自变量的个数 p 越来越多时，此时RSS的下降会和p的增大产生对抗。如果RSS下降幅度巨大，说明新加入的自变量对解释因变量有很大的影响，此时调整后的R^2会上升，反之则会下降。

在实际分析中，我们应当考虑哪一种判决系数呢？只要样本量足够大，调整R^2中的$(n-1)/(n-p-1)$就会非常接近 1，此时，两种判决系数的作用是相似的。但是，如果涉及到模型的比较与选择，使用调整后的R^2更好，因为未调整的R^2更倾向于复杂的模型，从而不具备控制模型复杂度的能力。

6.7 回归诊断

建立回归模型后，还需要对模型进行回归诊断。回归诊断就像是给模型“看病”，如果模型存在重大问题，那么通过诊断就能够看出端倪。回归诊断主要着重四个方面：异方差、强影响点、多重共线性和正态性。

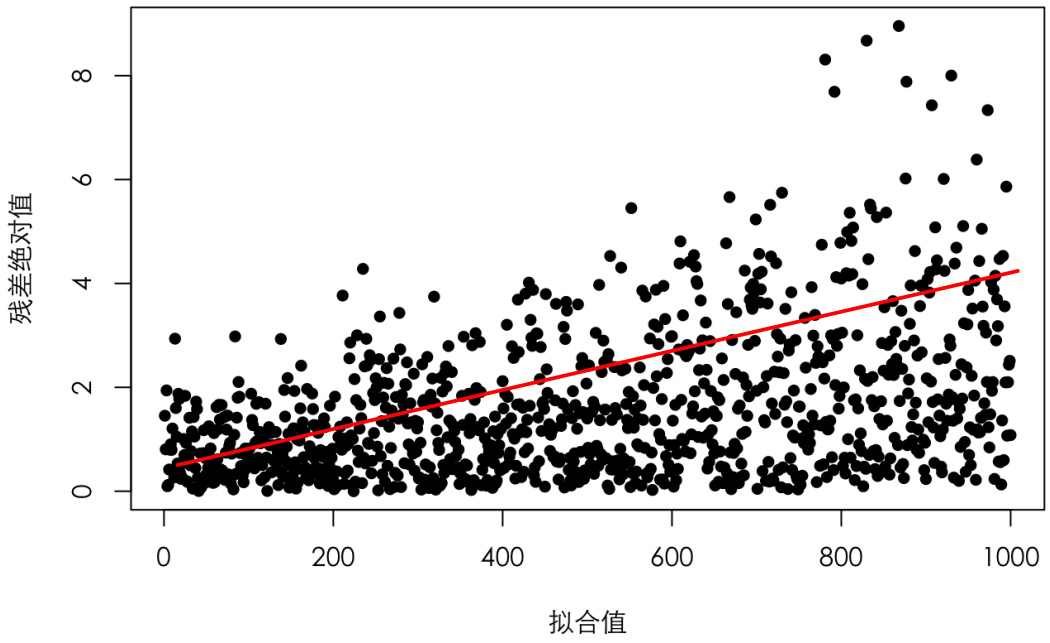
6.7.1 异方差

线性回归模型假设随机误差ε_i方差相同，如果不满足这一假设，则会产生异方差的问题。当模型存在异方差时，尽管此时仍然能够得到无偏的参数估计结果，但可能会存在以下问题：

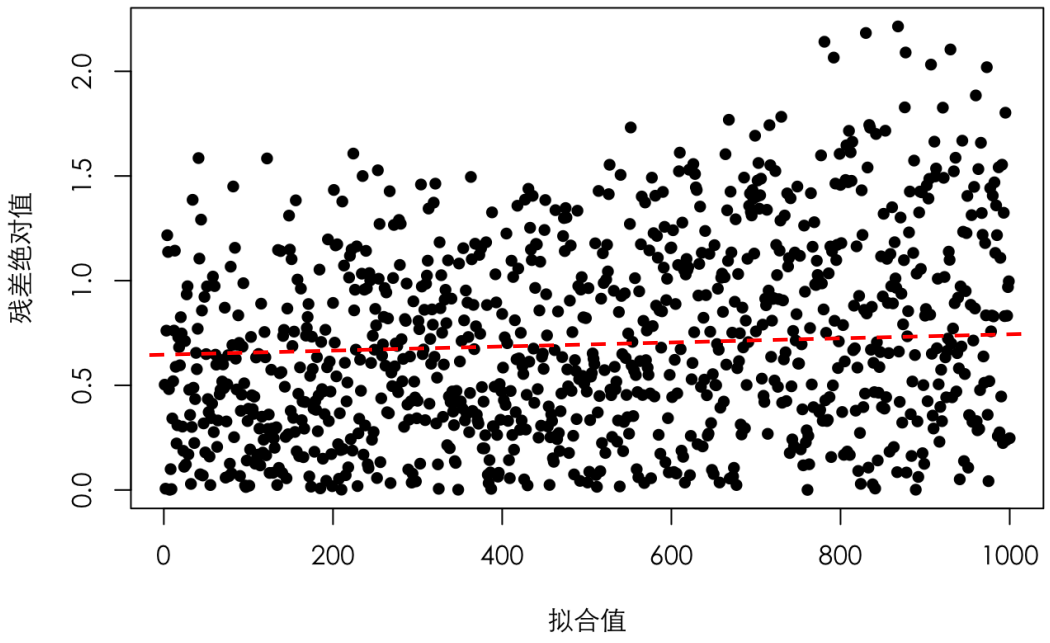
（1）参数的显著性检验失效；

（2）回归方程的应用效果不理想；

如何诊断异方差问题呢？一般通过绘制残差图的方法进行分析。具体而言，可以以回归拟合值$\hat{y}_i$为横轴，残差e_i为纵轴，绘制散点图。若散点图随着横轴的增大呈现出发散或汇聚的现象，则认为存在异方差。如图 6-4 (a)所示，可以观察到残差的波动随着$\hat{y}_i$值的增大而增大，说明存在异方差问题。可以尝试对 y 做对数变换（图 6-4 (b)），进行修正。



（a）



（b）

图 6-4：残差与拟合值散点图

6.7.2 强影响点

强影响点是指对回归模型估计结果有较大影响的点。如图 6-5 所示，点(x^*, y^*)可以认为是强影响点。强影响点的存在会使得回归线向自身靠拢（如图中虚线所

示），因此会对回归方程的估计结果产生较大影响。当存在强影响点时，可以剔除这些点后再进行回归分析。

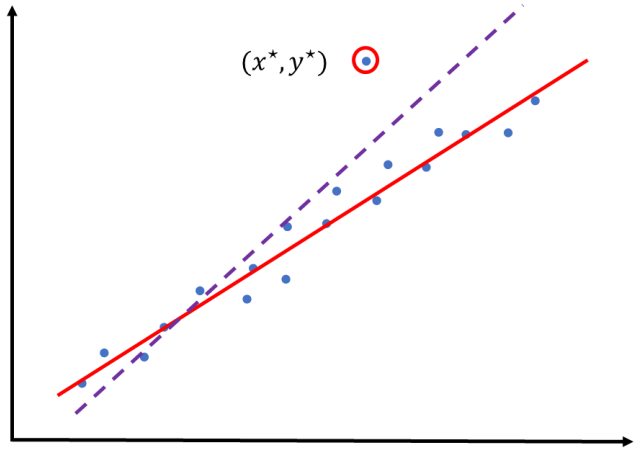


图 6-5：线性回归分析中的强影响点及其对回归直线的影响

如何判断一个样本是否是强影响点呢？著名统计学家 D. R. Cook 建议按照样本的影响程度对样本进行打分。如果一个样本是强影响点，那么它的得分就会较高。这个分值后来也被称作 Cook 距离（Cook's distance）。Cook 距离是如何计算的呢？想要评判一个样本的“影响力”，可以剔除该样本后对回归方程重新进行估计。如果剔除样本前后的回归估计差异不大，那么该样本不是强影响点；如果差异巨大，则该样本就具有强影响点的嫌疑。具体而言，记$\hat{\beta}$为我们基于所有样本获得的系数估计值，$\hat{y}_j$为使用$\hat{\beta}$得到的第j个样本的拟合值。对于一个给定的样本i，如果要判断这个样本的影响力，则剔除该样本后，重新估计回归系数，记为$\hat{\beta}_{(i)}$，此时得到的模型对第j个样本的拟合值为$\hat{y}_{j(i)}$，则 Cook 距离的计算方式如下：

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_{j(i)}\right)^2}{(p+1)s^2}$$

其中，$s^2 = e^{\top}e/(n-p)$，$e = y - \hat{y}$。如果有少数一两个样本的 Cook 距离特别大，则应该考虑将这样的样本剔除后，再重新拟合回归模型。一般认为，Cook 距离大于 1 的样本点具有强影响点的嫌疑。

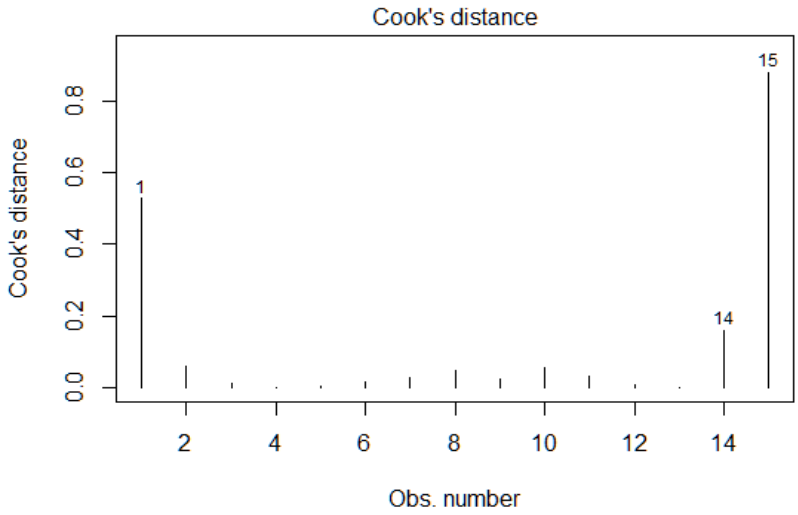


图 6-6：样本点的 Cook 距离

6.7.3 多重共线性

多元线性回归模型有一个基本假定，要求$rank(X) = p + 1$，即自变量之间是线性无关的，任何一个自变量不能被其他的自变量所“代替”。若存在不全为 0 的$p + 1$个数$c_0, c_1, \dots, c_p$，使得

$$c_0 + c_1 x_{i1} + \cdots + c_p x_{ip} \approx 0, \qquad i = 1,2, \dots, n$$

则称自变量$X_1, X_2, \dots, X_p$之间存在多重共线性（multi-collinearity）。多重共线性会对模型系数的估计产生影响。当存在完全共线性时，$c_0 + c_1 x_{i1} + \cdots + c_p x_{ip} = 0 \ (i = 1,2, \dots, n)$，$rank(X) < p + 1$，此时$|X^{\top} X| = 0$，方程组$X^{\top} X\beta = X^{\top} y$的解不唯一，无法求得线性回归系数的最小二乘估计。

在实际问题中，自变量间的相关度极强时，就会产生近似共线性的情形，即$c_0 + c_1 x_{i1} + \cdots + c_p x_{ip} \approx 0 \ (i = 1,2, \dots, n)$，此时$|X^{\top} X| \approx 0$。在回归结果中，常表现为回归方程整体显著，但单个系数不显著。此时难以判断自变量对因变量的影响程度，甚至还会出现回归系数符号与真实情况相反的问题。因此多重共线性的诊断是十分必要的。

目前诊断多重共线性的主要方法为方差膨胀因子法。自变量X_j的方差膨胀因子可以这样计算：将自变量X_j对其他自变量进行线性回归，计算回归方程的决定

系数R_j^2。决定系数R_j^2越高，则X_j被其他自变量解释的程度越高，共线性也就越大。为了分析方便，可以将R_j^2稍作变换，得到方差膨胀因子（variance inflation factor, VIF）：

$$\mathrm{VIF_j} = \frac{1}{1 - R_j^2}$$

经验表明，当$\mathrm{VIF_j} \geq 10$时，自变量X_j与其余自变量之间有严重的多重共线性，且这种多重共线性可能会过度地影响最小二乘估计值，因此建议剔除该自变量。

6.7.4 正态性

当自变量值固定时，因变量呈现正态分布，因此残差值也应当服从均值为 0 的正态分布。一般使用“正态 Q-Q 图”（normal Q-Q）来查看是否符合正态性假设。正态 Q-Q 图的横轴为正态分布对应的概率值，纵轴为标准化残差对应的概率值。若满足正态性假设，则图中的散点应当落在呈 45 度角的直线上（图 6-7）。

在实际问题中，正态性的分布特征很难完美满足。一般来说，当残差的分布存在较强的非对称性时，可以考虑对因变量进行变换（例如：对数变换），缓解残差分布的非正态性特征。

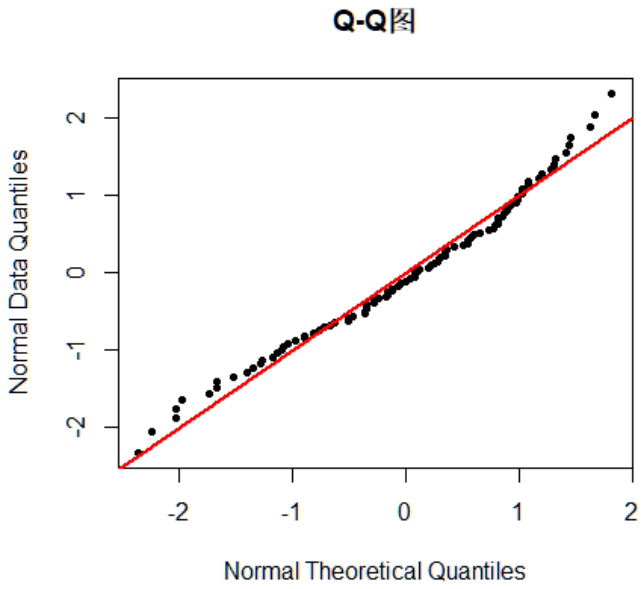


图 6-7：正态 Q-Q 图

6.8 变量选择

6.8.1 逐步回归法

当回归模型纳入的自变量较多时，往往模型复杂度过高，此时难以得到稳定性高、解释性强的回归结果，会影响模型的应用效果。回归模型的变量选择是指对自变量进行筛选。筛选目标是选出能够对因变量起到重要解释性的自变量，同时剔除解释性低的无关自变量。在变量选择的过程中如果遗漏了某些重要变量，回归方程的拟合及预测效果就会较差；如果考虑了过多的自变量，也会增加不必要的模型复杂度，降低模型可解释性。因此，选择合适的自变量在回归问题中十分关键。本节将介绍自变量选择的常用方法——逐步回归法（stepwise method）。

逐步回归法中，模型每次会添加或删除一个变量，直到达到某个判断的准则停止标准为止。逐步回归法主要包括三种变量选择模式：向前逐步回归、向后逐步回归和向前向后逐步回归。向前逐步回归（forward stepwise regression）每次添加一个自变量进入模型，直到添加变量不会使模型效果有所改进为止；向后逐步回归（backward stepwise regression）从包含所有自变量的模型开始，每次删除一个自变量，直到模型效果不再改进为止；向前向后逐步回归（通常简称逐步回归，stepwise regression）则结合了向前逐步回归和向后逐步回归的思想，每次添加一个变量后，对此时模型中的自变量进行重新评价，将对模型没有贡献的自变量剔除，因此可以实现自变量反复添加、删除，直到获得最优模型为止。

6.8.2 信息准则

信息准则是评价变量选择过程中模型拟合程度及复杂度的综合性指标，常用的信息准则有 AIC 准则与 BIC 准则。一般通过逐步回归法得到使 AIC 准则或 BIC 准则达到最优的模型。

AIC 准则（Akaike information criterion, AIC）是根据最大似然估计原理提出的一种较为一般的模型选择准则。设模型的似然函数为$L(\beta)$，β的维数为m，则 AIC 定义为：

$$AIC = -\frac{2}{n}L(\beta) + 2\frac{m}{n}$$

其中，n 表示样本量。由于似然函数越大的估计量越好，同时模型复杂程度越低（m 越小）的模型可解释性越好，因此选择**使得 AIC 达到最小的模型是最优模型**。

一般而言，当模型中变量较多，从而复杂度提高（m 增大）时，似然函数$L(\beta)$也会增大，从而使 AIC 变小。但是 m 过大时，似然函数增速减缓，导致 AIC 增大。因此使用 AIC 准则可以在提高模型拟合度（增大似然函数的值）的同时，降低模型复杂度，避免模型过拟合。

BIC 准则（Bayesian information criterion, BIC）与 AIC 准则的设计原理相似，但 BIC 准则对于模型复杂度的“惩罚”力度比 AIC 准则更强：

$$BIC = -2L(\beta) + \log(n) \times m$$

根据 AIC 和 BIC 的表达式可以看出，在数据量 n 较大时，BIC 准则对于模型复杂度的“惩罚”力度更强，因此在 BIC 准则下得到的最优模型一般更加简洁。

6.9 模型实现

6.9.1 R 语言中的基本函数

R 语言中实现线性回归可以使用 lm()函数。lm()函数中包含两个主要的参数：回归公式（formula）和数据集（data），格式为`myfit <- lm(formula, data)`。其中，formula 指要拟合的模型形式，data 是一个数据框，包含了用于拟合模型的数据，函数返回的结果对象（myfit）储存在一个列表中，包含了所拟合模型的大量信息。回归公式（formula）的形式如下：

$$Y \sim X_1 + X_2 + \cdots + X_p$$

`~`符号左边为因变量，右边为自变量，每个自变量之间用`+`符号分隔。下表中的符号可以用不同方式修改这一表达式。

符号	用途	示例
~	分隔符号，左边为因变量，右边为自变量	Y~X1+X2
+	分隔自变量	
:	表示自变量的交互项	Y~X1+X2+X1:X2
*	表示所有可能交互项的简洁方式	Y~X1*X2
^	表示交互项达到某个次数	Y~X1+X2+X2^2

.	表示包含除因变量之外的所有变量	Y~.
-	表示从等式中移除某个变量	Y~. – X1
-1	删除截距项	Y~. – 1

在完成模型拟合后，将下表中所示的函数应用于 lm()返回的对象，可以得到更多额外的模型信息。

函数	用途
summary()	展示拟合模型的详细结果
coefficients()	列出拟合模型的模型参数
confint()	提供模型参数的置信区间（默认 95%）
fitted()	列出拟合模型的预测值
residuals()	列出拟合模型的残差值
vcov()	列出模型参数的协方差矩阵
AIC()	输出 AIC 值
plot()	生成评价拟合模型的诊断图
predict()	用拟合模型对新的数据集预测因变量值

6.9.2 实例分析

使用数据分析岗位招聘薪酬数据集。首先以岗位薪资为因变量，其他的变量为自变量，在 R 中使用 lm()函数建立回归模型。

1. 建立线性回归模型

```
## 建立线性模型
lm.fit1 = lm(aveSalary ~ ., data = jobinfo)
## 查看回归结果
summary(lm.fit1)
```

```
##
## Call:
## lm(formula = aveSalary ~ ., data = jobinfo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10071.2  -2153.6   -676.5   1654.8  13515.5
```

```
##
## Coefficients:
##                     Estimate Std. Error t value           Pr(>|t|)
## (Intercept)          6768.910    254.284  26.619 < 0.0000000000000002 ***
## R1                    466.596    218.212   2.138            0.03253 *
## SPSS1                 423.658    207.874   2.038            0.04158 *
## Excel1               -970.032     97.652  -9.934 < 0.0000000000000002 ***
## Python1               718.776    234.923   3.060            0.00222 **
## MATLAB1               -75.241    302.845  -0.248            0.80380
## Java1                 568.885    258.771   2.198            0.02795 *
## SQL1                 1275.237    148.395   8.594 < 0.0000000000000002 ***
## SAS1                  332.544    227.232   1.463            0.14339
## Stata1              -1062.972    814.138  -1.306            0.19172
## EViews1               419.079    844.015   0.497            0.61954
## Spark1                 -3.709    436.855  -0.008            0.99323
## Hadoop1              1736.165    330.751   5.249   0.0000001575250389 ***
## 公司类别合资          723.833    236.596   3.059            0.00223 **
## 公司类别外资          293.873    234.330   1.254            0.20985
## 公司类别上市公司      597.552    264.734   2.257            0.02403 *
## 公司类别民营公司      378.531    211.465   1.790            0.07349 .
## 公司类别创业公司      893.262    402.988   2.217            0.02668 *
## 公司规模 50-500 人    282.346    111.355   2.536            0.01125 *
## 公司规模 500-1000 人   77.016    149.989   0.513            0.60763
## 公司规模 1000-5000 人 427.060    156.551   2.728            0.00639 **
## 公司规模 5000-10000 人 339.565   278.032   1.221            0.22201
## 公司规模 10000 人以上  329.871    229.577   1.437            0.15080
## 学历要求中专        -1432.439    272.913  -5.249   0.0000001579146062 ***
## 学历要求高中        -1626.832    318.650  -5.105   0.0000003392908562 ***
## 学历要求大专         -930.555    121.848  -7.637   0.0000000000000253 ***
## 学历要求本科          776.026    126.739   6.123   0.0000000009705010 ***
```

```
## 学历要求研究生        1725.095    286.849   6.014   0.0000000019065795 ***
## 工作经验               682.515     23.073  29.580 < 0.0000000000000002 ***
## 地区非北上深         -2515.269    102.062 -24.645 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3105 on 6652 degrees of freedom
## Multiple R-squared:  0.3199, Adjusted R-squared:  0.3169
## F-statistic: 107.9 on 29 and 6652 DF,  p-value: < 0.00000000000000022
```

2. 回归诊断和模型修正

拟合回归模型之后，还需要通过回归诊断判断模型是否合适。虽然 summary() 函数对模型做了整体的描述，但是它并没有包含回归模型诊断的信息。在 R 语言中直接使用 plot()函数，就可以展现模型诊断的结果，实现方式如下。

```
## 对线性模型进行回归诊断
# 将画布分为 2*2 的 4 块
par(mfrow=c(2,2))
plot(lm.fit1, which = c(1:4))
```

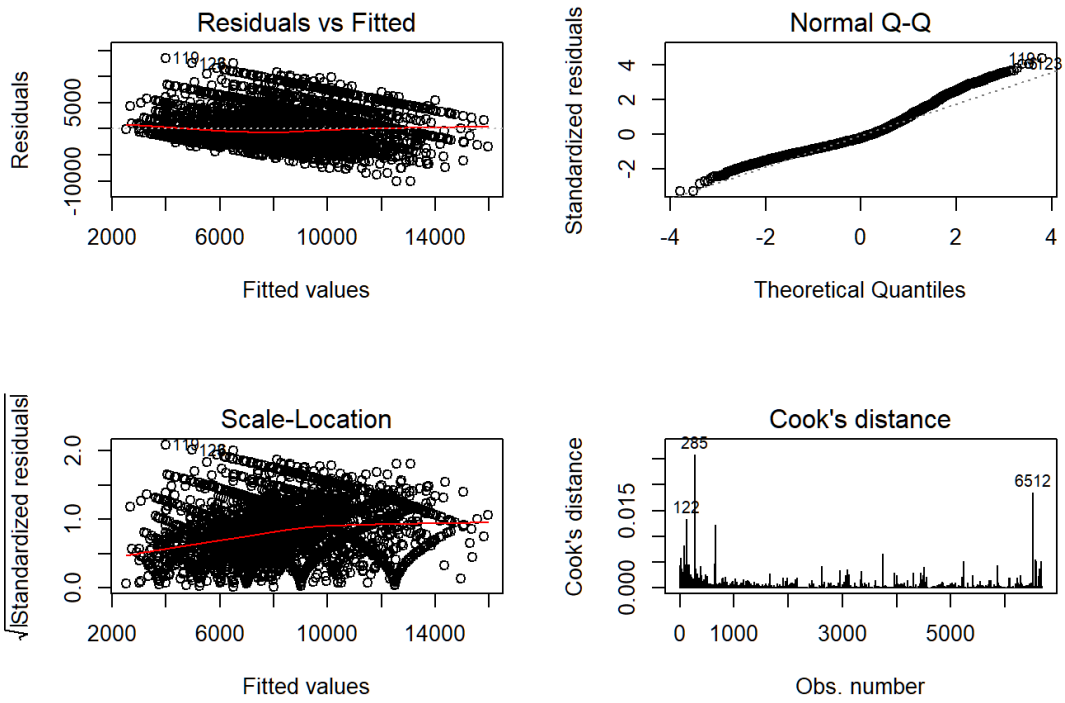


图 6-8：R 语言输出的回归诊断图像

左上图为残差与拟合值的散点图，若因变量和自变量呈现线性相关的关系，则残差值与拟合值没有任何的系统关联。图中所示残差并不随着拟合值的变化呈现规律性变化，因此基本满足线性假设；右上图为正态 Q-Q 图，当因变量服从正态分布时，图中的散点应该落在呈 45 度倾斜的直线上。根据图示，模型中的残差项在较大值的部分偏离直线，因此较大程度上偏离了正态分布；左下图为位置尺度图，若满足同方差假定，则水平线周围的点应呈现无规律随机分布，根据图中所示，随着拟合值增大，标准化残差呈现升高的规律，表明存在一定的异方差问题；右下图为样本点的库克距离，其最大值未超过 0.05，因此认为样本中不存在异常点。

针对以上模型诊断出现的问题，可将因变量进行对数变换，重新拟合回归模型。R 语言实现如下：

```
## 计算对数因变量
jobinfo$对数薪资 <- log(jobinfo$aveSalary)
# 建立对数线性模型，剔除平均薪资变量
lm.fit2 = lm(对数薪资 ~ .-aveSalary, data = jobinfo)
```

```
## 查看回归结果
summary(lm.fit2)
```

```
##
## Call:
## lm(formula = 对数薪资 ~ . - aveSalary, data = jobinfo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.59999 -0.26299 -0.02898  0.25620  1.43210
##
## Coefficients:
##                    Estimate Std. Error t value           Pr(>|t|)
## (Intercept)          8.753373   0.030511 286.892 < 0.0000000000000002 ***
## R1                  0.032774   0.026183   1.252            0.21071
## SPSS1               0.055665   0.024942   2.232            0.02566 *
## Excel1             -0.117507   0.011717 -10.029 < 0.0000000000000002 ***
## Python1             0.083433   0.028188   2.960            0.00309 **
## MATLAB1            -0.006056   0.036338  -0.167            0.86765
## Java1               0.069968   0.031050   2.253            0.02426 *
## SQL1                0.147355   0.017806   8.276 < 0.0000000000000002 ***
## SAS1                0.046839   0.027265   1.718            0.08586 .
## Stata1             -0.123253   0.097687  -1.262            0.20710
## EViews1             0.039811   0.101272   0.393            0.69425
## Spark1             -0.037998   0.052417  -0.725            0.46853
## Hadoop1             0.179699   0.039686   4.528 0.000006058448713169 ***
## 公司类别合资        0.087100   0.028389   3.068            0.00216 **
## 公司类别外资        0.025907   0.028117   0.921            0.35686
## 公司类别上市公司    0.074514   0.031765   2.346            0.01902 *
## 公司类别民营公司    0.041935   0.025373   1.653            0.09844 .
## 公司类别创业公司    0.103932   0.048354   2.149            0.03164 *
```

```
## 公司规模50-500人      0.034777   0.013361   2.603            0.00927 **
## 公司规模500-1000人    0.017690   0.017997   0.983            0.32567
## 公司规模1000-5000人   0.044287   0.018784   2.358            0.01842 *
## 公司规模5000-10000人  0.045901   0.033361   1.376            0.16890
## 公司规模10000人以上   0.048603   0.027546   1.764            0.07771 .
## 学历要求中专         -0.200786   0.032746  -6.132 0.000000000919981542 ***
## 学历要求高中         -0.223217   0.038234  -5.838 0.000000005527645379 ***
## 学历要求大专         -0.117711   0.014620  -8.051 0.000000000000000964 ***
## 学历要求本科          0.090812   0.015207   5.972 0.000000002469471753 ***
## 学历要求研究生        0.201236   0.034418   5.847 0.000000005250271705 ***
## 工作经验              0.084425   0.002769  30.495 < 0.0000000000000002 ***
## 地区非北上深         -0.374549   0.012246 -30.585 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3725 on 6652 degrees of freedom
## Multiple R-squared:  0.3476, Adjusted R-squared:  0.3448
## F-statistic: 122.2 on 29 and 6652 DF,  p-value: < 0.00000000000000022
```

拟合以对数线性回归模型之后，同样需要通过回归诊断判断模型是否满足基本假设。

```
# 将画布分为2*2的4块
par(mfrow=c(2,2))
plot(lm.fit2, which = c(1:4))
```

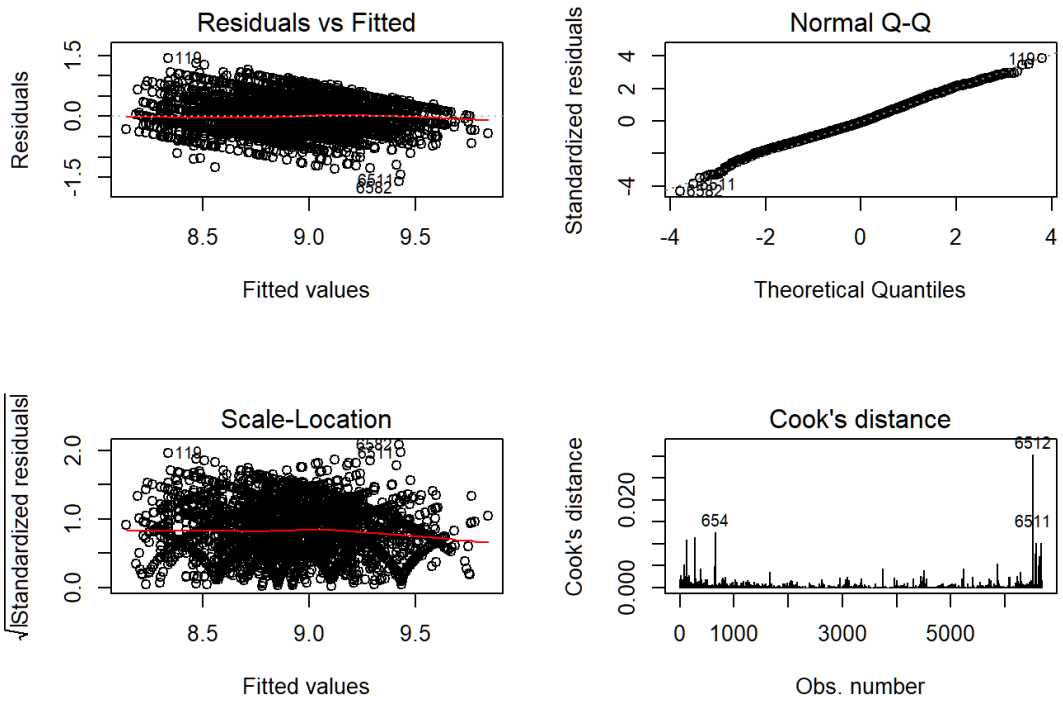


图 6-9：调整后的模型回归诊断图

由 6-9 图中的诊断结果所示，回归模型的异方差及非正态性都得到了改善。

接下来，使用 DAAG 包中的 vif()函数诊断模型的多重共线性，实现方式如下。

```
# 多重共线性诊断：计算 VIF 值
library(DAAG)
vif(lm.fit2)
```

```
##                R1             SPSS1            Excel1
##            1.9138            2.1948            1.1146
##           Python1           MATLAB1              Java1
##            1.6090            1.2310            1.4519
##              SQL1              SAS1            Stata1
##            1.3385            2.3449            1.0974
##           EViews1             Spark1            Hadoop1
##            1.0323            1.7191            1.8908
##      公司类别合资      公司类别外资    公司类别上市公司
```

```
##              3.5698              3.7260              2.3917
##     公司类别民营公司     公司类别创业公司     公司规模 50-500 人
##              6.6519              1.3644              2.1190
##   公司规模 500-1000 人 公司规模 1000-5000 人 公司规模 5000-10000 人
##              1.6593              1.6808              1.1451
##  公司规模 10000 人以上         学历要求中专         学历要求高中
##              1.2954              1.1550              1.1191
##         学历要求大专         学历要求本科       学历要求研究生
##              2.5233              2.6254              1.1862
##             工作经验         地区非北上深
##              1.0490              1.0905
```

根据结果可以看出，所有变量的 VIF 值都小于 10，因此认为自变量之间不存在较强的多重共线性。

3. 模型解读

R 输出的结果显示，模型的 F 检验$p\text{-}value < 0.05$，因此在 0.05 的显著性水平下，该模型整体是有意义的。每一个自变量的输出信息分别为：系数估计值、标准误、t 统计量值、t 检验的 p 值和对应的显著性水平。以变量“经验要求”为例，其系数估计值为 0.085，估计的标准误为 0.003，t 统计量为 30.567，若取显著性水平为 0.05，则“经验要求”与薪资有显著为正的相关性。

回归系数直接反映了因变量与自变量间的相关关系。线性回归模型的系数基本含义是，在控制其他自变量不变的条件下，某个自变量每变化 1 个单位，导致的（模型中的）因变量变化的平均值。在本案例中，使用对数线性回归模型，因此在解读上也稍有不同。利用偏微分的表达形式：$\partial\, log(y)\,/\partial x = \partial y/(y \times \partial x)$，可以看出，$x$的回归系数可以近似解读为：在控制其他自变量不变的条件下，某个自变量每变化 1 个单位，导致的薪资变化的比例。具体解读如下。

1）自变量为数值型变量时，系数解释为自变量每增加一个单位，因变量增加的比例。例如自变量“经验要求”的回归系数（0.084）可解释为：在控制其他因素不变的情况下，对数据分析人员的工作经验年限要求每多一年，相应岗位的薪资就平均高出 8.4%。

2）自变量为 0-1 定性变量时，线性回归模型系数可解释为自变量取分类“1”时，因变量相对于自变量取分类“0”时平均增加的比例。比如自变量“R”对应的系数为 0.033，说明要求掌握 R 的岗位，薪资平均比不要求掌握 R 的岗位高出 3.3%。这和本章开头的描述分析趋势是吻合的。

3）自变量为多分类自变量时，线性回归模型系数可以解释为自变量取该分类的某个水平时，因变量平均相对于基准水平增加的比例。比如自变量“公司规模”，其基准水平为“少于 50 人”，那么“5000-10000 人”对应的系数 0.046 可解读为：公司规模为 5000-10000 人的岗位薪资平均比公司少于 50 人的岗位高 4.6%。

但是，值得注意的是，由于我们在对数线性回归系数解读时做了近似处理，因此可能存在一些偏差。更为直接的方式是计算因变量增长率$exp(\beta_j)-1$，作为对自变量X_j系数的解读。

4. 模型选择

对于上一节建立的模型，使用 BIC 准则选择最优模型。在 R 中使用 step()函数作用于 lm()返回的模型。step()函数中，参数 direction 表示选择模型的方法，`forward`表示向前选择、`back`表示向后选择、`both`表示结合两个方向逐步选择；参数 k 决定逐步回归的准则，设置`k=2`使用 AIC 准则，设置`k=n`使用 BIC 准则，其中 n 表示样本量。实现方式如下所示。

```
## 使用BIC准则选择模型
n <- nrow(jobinfo)
lm.bic <- step(lm.fit2, direction = "both", k = log(n), trace = F)
summary(lm.bic)
```

```
##
## Call:
## lm(formula = 对数薪资 ~ SPSS + Excel + Python + SQL + Hadoop +
##     学历要求 + 工作经验 + 地区, data = jobinfo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -1.54248 -0.25992 -0.02571  0.25825  1.43138
##
## Coefficients:
##               Estimate Std. Error t value           Pr(>|t|)
## (Intercept)    8.827384   0.013680 645.276 < 0.0000000000000002 ***
## SPSS1          0.092435   0.018706   4.941  0.00000079418407451 ***
## Excel1        -0.117721   0.011646 -10.108 < 0.0000000000000002 ***
## Python1        0.110596   0.024510   4.512  0.00000652522467175 ***
## SQL1           0.160374   0.017389   9.223 < 0.0000000000000002 ***
## Hadoop1        0.191131   0.030947   6.176  0.00000000069579684 ***
## 学历要求中专   -0.202515   0.032791  -6.176  0.00000000069674498 ***
## 学历要求高中   -0.223880   0.038146  -5.869  0.00000000459536464 ***
## 学历要求大专   -0.116834   0.014630  -7.986  0.00000000000000163 ***
## 学历要求本科    0.094769   0.015130   6.263  0.00000000039996660 ***
## 学历要求研究生  0.201863   0.034147   5.912  0.00000000355550562 ***
## 工作经验        0.084348   0.002766  30.496 < 0.0000000000000002 ***
## 地区非北上深   -0.371999   0.011976 -31.061 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3732 on 6669 degrees of freedom
## Multiple R-squared:  0.3436, Adjusted R-squared:  0.3424
## F-statistic: 290.9 on 12 and 6669 DF,  p-value: < 0.00000000000000022
```

使用 BIC 准则进行逐步回归后，留下的自变量都对因变量有显著影响。自变量的系数估计值与模型选择之前略有区别，对于自变量系数的解读同理可得。

5. 模型预测

在选择出最优的回归模型之后，就可以对新收集到的样本数据进行预测。假设有一份北京市上市公司数据分析岗位的工作，该公司为 1500 人的中小型公司，这个工作要求申请人掌握 R、Python、SQL 和 Hadoop 的技能，并且有至少 3 年

的工作经验，最低学历为硕士。希望通过模型预测该岗位的薪资。

在 R 语言中，可以使用 predict()函数完成模型预测。但是，值得注意的是，直接做回归模型预测得到的仍是$\log(Y)$的预测值。如果需要得到Y的预测值则需要经过一些变换。

在因变量Y满足对数线性模型的基础上（即：$\log(Y) = X^{\top}\beta + \varepsilon$），可以推导得到：

$$E(Y) = exp(X^{\top}\beta + \sigma^2/2)$$

因此，预测的薪资计算方式为：

$$\hat{y} = exp(X^{\top}\hat{\beta} + \hat{\sigma}^2/2)$$

在 R 中实现如下：

```
## 新样本
testdata <- data.frame(R = 1, SPSS = 0, Excel = 0, Python = 1, MATLAB = 0, Java = 0, SQL = 1, SAS = 0, Stata = 0, EViews = 0, Spark = 0, Hadoop = 1, 公司类别 = "上市公司", 公司规模 = "1000-5000 人", 学历要求 = "研究生", 工作经验 = 3, 地区 = "北上深")
## 将软件技能转换为 factor 类型
for (i in c(1:12)) {
    testdata[,i] <- as.factor(testdata[,i])
}
logsalary_hat <- predict(lm.bic, newdata = testdata)  # 预测值
sigma_hat2 <- sum(lm.bic$residuals^2)/lm.bic$df.residual  # sigma^2 估计值
y_hat <- exp(logsalary_hat + sigma_hat2/2)  #
cat("薪资水平约为", round(y_hat, 2), "元/月")
## 薪资水平约为 18288.72 元/月
```

根据模型的预测结果，该工作岗位薪资约为 18288.72 元/月。

6.10 小结

回归分析是统计分析中最重要的思想之一，它主要解读变量之间的相关关系，在实际的业务场景中有着非常广泛的应用。

本章中，我们借助数据分析岗位招聘薪酬数据集，详细介绍了线性回归分析的相关内容，包括模型形式、模型理解、回归参数的估计方法、模型评价与回归诊断、模型选择。最后，本章展示了如何在 R 语言中进行完整的线性回归分析，并且对新样本数据进行预测。

6.11 本章习题

1. 推导并证明：在对数线性回归模型中，因变量的期望表达式为：

$$E(Y) = exp(X^{\top}\beta + \sigma^2/2)$$

2. 使用例 2 的案例背景，分析影响二手房单位面积房价的重要因素。数据集 house.csv 为来自某二手房中介网站的北京在售二手房 2016 年 5 月的相关数据，共包括单位面积房价（price）、城区（CATE）、卧室数（bedrooms）、厅数（halls）、房屋面积（AREA）、楼层（floor）、是否临近地铁（subway）、是否是学区房（school）这几个变量。以房价为因变量，在 R 中建立普通线性回归模型，并对模型结果进行诊断。

3. 对于题目 2 中的数据，建立对数线性模型，并加入城区与学区的交互项，对系数进行解读。

4. 使用 BIC 准则对题目 3 的模型进行选择，使用最终的模型进行新样本单位面积房价的预测。其中，预测样本为一间海淀区的两室一厅学区房，其在楼中的低楼层，并且临近地铁，房屋面积为 70 平方米。

6.12 参考答案

1. 通过期望的定义及对数线性模型的形式，推导$E(Y|X)$的表达式。

2. 将分类变量转换为因子型，并在 R 中建立普通线性回归模型与模型诊断，代码略。

3. 加入交互项后，在 R 中建立对数线性回归模型并进行诊断与系数解读，代码略。

4. 使用 BIC 选择后的模型对新样本进行预测，预测的期望值使用题目 1 中推导出的表达式。