第 3 章

基本线性回归

线性回归模型是计量经济学中最经典且最基础的计量模型，能够解释经济现象之间的线性关系。本章首先从最为简单的一元线性回归模型开始，介绍计量经济学模型的设定与估计，帮助学生理解普通最小二乘法（OLS）以及回归分析的基本思想。在此基础上，我们将模型拓展至多元线性回归。作为一元线性回归模型的推广，多元线性回归模型更符合经济现实，具有更强的解释力和应用性，本部分重点介绍多元回归的基本假设、模型估计、OLS 估计量的性质和统计检验；最后，我们进一步放松假设，讨论多元线性回归下 OLS 的渐进性。线性回归模型是经济学和其他社会科学中最广泛使用的经验分析工具，普通最小二乘法在估计线性回归模型参数时也十分常见。

教学目标：通过本章的学习，学生应能理解和掌握一元线性回归和多元线性回归，能够进行线性回归模型的参数估计和检验。

教学要点：

- 理解一元线性回归和多元线性回归的基本概念，掌握两者的联系与区别。
- 理解多元线性回归模型的矩阵表示。
- 掌握普通最小二乘法的基本思想。
- 掌握模型的拟合优度及校正拟合优度。
- 掌握相关统计推断和检验方法，特别是 t 检验与 F 检验的区别。
- 理解大样本下 OLS 的渐进性质。

3.1 一元线性回归

下载资源:\sample\chap03\正文\livebc.dta

3.1.1 一元线性回归模型

1. 回归分析概述

在现实生活中，人们常常遇到各种经济方面的问题，例如，经营者可能想了解广告投入对公司利润的影响，一位农民可能想知道每多施一千克肥料能提高多少大豆产量，学生可能希望了解考研能否为自己将来的求职带来更高的薪资等。计量经济学的一个重要目的就是回答这类问题，采用的方法便是回归分析（regression analysis）。

“回归”（regression）一词最早由弗朗西斯·高尔顿（Francis Galton）引入。高尔顿在研究子女身高与父母身高的关系时发现，尽管高个父母的子女平均较高，矮个父母的子女平均更矮，但极高个父母的子女通常比父母要矮，而极矮个父母的子女则通常比父母高。也就是说，父母的身高越是极端，其子女的身高会回归到人口的平均身高水平。高尔顿称这一现象为回归平庸（regression towards mediocrity），现代更常称为回归均值（regression to the mean）。

在现代计量经济学中，回归分析是研究一个被解释变量 Y 与另一个或多个解释变量 X 之间关系的理论和方法。其主要目的是通过解释变量的已知值来估计被解释变量的均值。回归分析不仅能定量经济变量之间的关系，还可以通过统计分析和检验判断结果的可信度。因此，回归分析是研究经济现象的强有力工具，是计量经济分析的核心。

2. 一元线性回归模型

一元线性回归模型是最简单的计量经济学模型，模型中仅有一个解释变量。假设从总体中随机抽取了 n 个个体，则该模型可以表示为：

$$y_i = \alpha + \beta x_i + \varepsilon_i \tag{3.1}$$

其中，y_i 为被解释变量（explained variable）或因变量（dependent variable），x_i 为解释变量（explanatory variable）或自变量（independent variable）；α 与 β 为待估参数（parameters）或回归系数（regression coefficients），其中 α 为截距项（intercept）或常数项（constant），β 为斜率（slope）；ε_i 为扰动项（disturbance）或误差项（error term），它包含除 x_i 外，其他所有可能影响 y_i 的因素。下标 i 表示第 i 个个体，取值为 1,⋯,n，n 为样本容量。

例 3.1：为了更好地理解一元线性回归模型，我们通过一个实际例子进行讲解。随着网络经济的快速发展，网络直播成为众多年轻人就业的选择，甚至有人成为年入千万甚至过亿的“网红”。为了吸引更多流量和粉丝，很多主播起早贪黑，延长直播时长来提高收入。尤其对初入直播界的新人来说，多露脸才能有机会让更多的人看到。那么，增加直播时长真的能够提高收入吗？

为了回答这个问题，我们通过第三方平台收集了 601 位抖音美妆博主在某月的直播时长和该月

直播收入的相关数据[1]，构建如下回归模型：

$$\text{lnwage}_i = \alpha + \beta \text{hour}_i + \varepsilon_i \tag{3.2}$$

与式（3.1）相对应，lnwage_i 为被解释变量，即直播收入的自然对数值；hour_i 为解释变量，是该月平均每天的直播时长；斜率 β 是我们关心的关键，即主播每延长一小时的直播时间，收入增加的百分比；常数项 α 则表示当直播时长为 0 时的收入。尽管主播不直播，有些博主仍有收入，比如通过签约获得固定底薪、收到追加打赏等。当然，除了直播时长外，直播内容质量、粉丝基数、个人能力等因素都会影响直播收入，均被纳入扰动项ε_i。

为了直观地理解，我们绘制了直播收入对数与直播时长两者关系的散点图，并在图中画出了离样本点最近的拟合线，如图 3.1 所示。从图中可以看出，直播收入对数与直播时长呈正相关，似乎存在线性关系，即直播时长越长，收入越高。

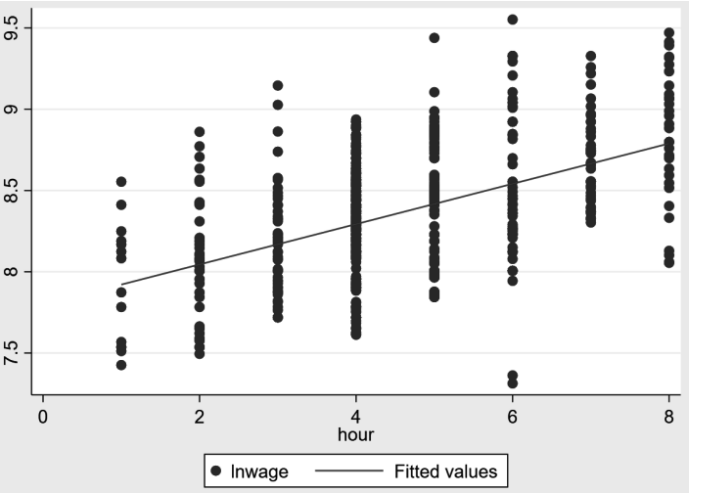

图 3.1　直播收入对数与直播时长的散点图和线性拟合

这里需要强调对“线性”的定义。所谓“线性”，指的是模型对参数是线性的。例如，模型 $Y_i = \alpha + \beta X_i^2 + \varepsilon_i$ 中，尽管解释变量 X 是非线性的，但参数 β 是线性的，这样的模型仍然为线性模型；再如，模型 $Y_i = aX_i^{\beta} e_i^{\mu}$ 经过对数变换转换为 $\ln Y_i = \ln\alpha + \beta \ln X_i + \mu_i$，则经过转换后的模型也是一个线性模型。

3. 总体回归函数与样本回归函数

从图 3.1 中可以看出，尽管不同主播的收入存在差异，但平均来说，随着直播时长的增加，直播收入也在增加。也就是说，对于每一个确定的解释变量值（如小时数 hour），被解释变量（如收入的对数值 lnwage）的条件均值会落在一条正斜率的直线上，这是回归分析中的均值概念。

对于总体来说，被解释变量 Y 的条件均值 $E(Y \mid x_i)$ 随着解释变量x_i的变化而规律性地变化，即每一条件均值 $E(Y \mid x_i)$ 可以看作 x_i 的一个函数，可以用如下公式表示：

$$E(Y \mid x_i) = f\left(x_i\right) \tag{3.3}$$

[1] 该第三方平台所公布的直播时长为每月平均的日直播时长，均为整数值。

将总体被解释变量的条件期望表示为解释变量的某种函数，这个函数被称为总体回归函数（Population Regression Function，PRF）或总体回归线（Population Regression Line）。总体回归函数表明了被解释变量 Y 的均值是如何随解释变量 x 变化的，但并不意味着每个个体的(x, y)都能完美满足这个函数。若 $E(Y \mid x_i)=f\left(x_i\right)=\alpha+\beta x_i$，则每个个体的回归函数为 $y_i=\alpha+\beta x_i+\varepsilon_i$。

尽管总体回归函数揭示了总体中被解释变量与解释变量间的平均变化规律，但在实际情况中，通常无法获得总体的完整信息，总体回归函数也是未知的。因此，我们通常采取抽样方式获得样本，再通过样本的信息来估计总体回归函数。

与总体回归函数类似，通过抽样可以得到样本回归函数（Sample Regression Function，SRF）或样本回归线（Sample Regression Line），其表达式为：

$$\hat{y}_i=f\left(x_i\right)=\hat{\alpha}+\hat{\beta} x_i \tag{3.4}$$

式中，$\hat{y}_i$ 为 $E(Y \mid x_i)$ 的估计值，$\hat{\alpha}$ 、$\hat{\beta}$ 分别为 α、β 的估计值。对于样本中的每一个个体，同样可以写出其回归函数，即 $y_i=\hat{\alpha}+\hat{\beta} x_i+e_i$，$e_i$ 为残差项（residual），代表了该个体影响 y_i 的其他因素的集合。

3.1.2 普通最小二乘法

普通最小二乘法（Ordinary Least Squares，OLS）是回归分析中最经典、应用也最广泛的一种估计方法。已知一组样本观测值 $\left\{\left(x_i, y_i\right): i=1, \cdots, n\right\}$，其在平面坐标上的分布如图 3.2 所示。那么，我们的任务就是找出能够较好反映样本数据趋势的样本回归线，即图中的 $\hat{y}_i=\hat{\alpha}+\hat{\beta} x_i$，以此来估计总体回归线 $E(Y \mid x_i)=\alpha+\beta x_i$。实际上，在这个平面上能画出很多条直线，但关键在于哪一条更好？或者说，我们如何判断哪一条直线能够更好地拟合样本数据的趋势？

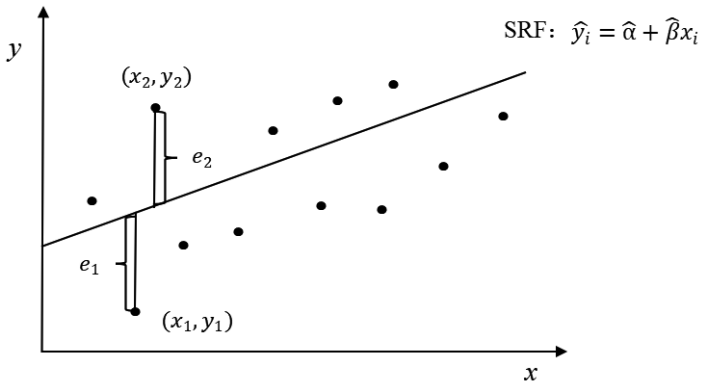

图 3.2　普通最小二乘法示意图

OLS 要求这条直线离所有样本点（即观测值）尽可能近，即每个样本点与样本回归线之间的纵向距离尽可能小，用数学表示为 $e_i \equiv y_i-\hat{y}_i=y_i-\hat{\alpha}-\hat{\beta} x_i$，其中 e_i 为残差，y_i 是某个样本点的纵轴值，$\hat{y}_i$ 是该点在样本回归线上对应的纵轴值。

残差 e_i 越小，说明样本回归线距离该样本点越近；如果所有样本点的残差都较小，即样本回归线整体上离所有样本点都较近，则说明该回归线对样本趋势的拟合程度较好。但由于残差既可能为

正，也可能为负，若直接对残差求和，容易出现正负相抵的情况。因此，OLS 将残差的平方项加总，得到$\sum_{i=1}^{n}e_i^2$，即残差平方和（Residual Sum of Squares，RSS）。普通最小二乘法给出的判断标准是：被解释变量的估计值与实际观测值之差的平方和最小。在数学上，OLS 的目标函数可表示为：

$$\min_{\hat{\alpha},\ \hat{\beta}}\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)^2 \tag{3.5}$$

即在给定样本观测值的条件下，选择$\hat{\alpha}$、$\hat{\beta}$，使残差平方和达到最小化，这就是普通最小二乘法的基本思想。

根据微积分的相关知识，将目标函数式（3.5）分别对$\hat{\alpha}$、$\hat{\beta}$求偏导数，并令其等于零，可以得到一个关于估计量$\hat{\alpha}$、$\hat{\beta}$的二元一次线性方程组，该方程组可表示为式（3.6）或式（3.7），称为正规方程组。

$$\begin{cases}\sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)=0\\ \sum_{i=1}^{n}\left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)x_i=0\end{cases} \tag{3.6}$$

$$\begin{cases}n\hat{\alpha}+\hat{\beta}\sum_{i=1}^{n}x_i=\sum_{i=1}^{n}y_i\\ \hat{\alpha}\sum_{i=1}^{n}x_i+\hat{\beta}\sum_{i=1}^{n}x_i^2=\sum_{i=1}^{n}x_iy_i\end{cases} \tag{3.7}$$

求解上述正规方程组，可得参数估计量的表达式为：

$$\begin{cases}\hat{\alpha}=\overline{y}-\hat{\beta}\overline{x}\\ \hat{\beta}=\dfrac{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2}\end{cases} \tag{3.8}$$

其中，$\overline{x}\equiv\frac{1}{n}\sum_{i=1}^{n}x_i$为$x$的样本均值，$\overline{y}\equiv\frac{1}{n}\sum_{i=1}^{n}y_i$为$y$的样本均值。根据式（3.8）可以解得估计量$\hat{\alpha}$、$\hat{\beta}$，由此得到样本回归线$\hat{y}\equiv\hat{\alpha}+\hat{\beta}x_i$。由于$\hat{\alpha}$、$\hat{\beta}$的估计结果是由 OLS 方法得到的，故称其为 OLS 估计量[1]。

[1] 在对回归模型作出一定基本假设的情况下，OLS 估计量具有良好的统计性质。鉴于本章的目的在于帮助学生更好地理解 OLS 的基本思想，且一元线性回归模型在现实中的应用相对较少，因此，本章不对模型的基本假设和 OLS 估计量的统计性质进行详细介绍，而将相关内容放在应用更为广泛的多元线性回归模型中加以阐述。

3.1.3 OLS 估计量的数学性质

接下来介绍一些关于 OLS 估计量的重要数学性质。

（1）OLS 残差之和以及残差的样本均值均为零。即：

$$\sum_{i=1}^{n} e_i = 0 \tag{3.9}$$

由于 $e_i = y_i - \hat{\alpha} - \hat{\beta} x_i$，因此这一性质可以由式（3.6）的第一个方程直接得到。残差之和为 0，则残差的样本均值自然也为 0。

（2）样本点 $(\bar{x},\ \bar{y})$ 一定位于样本回归线上。即：

$$\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x} \tag{3.10}$$

该性质可以由式（3.8）中 $\hat{\alpha}$ 的表达式直接得到。

（3）OLS 残差与解释变量不相关。即：

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{3.11}$$

这一性质可以由式（3.6）的第二个方程直接得到。

利用 OLS 估计量的上述三个重要数学性质，可以推出很多其他性质。例如，根据性质（1）可知残差的样本均值为 0，而残差的定义为 $e_i \equiv y_i - \hat{y}_i$，对该式两边求均值，便可以得到 $\bar{y} = \bar{\hat{y}}$。再比如，利用性质（1）和性质（3），可以证明被解释变量估计值 $\hat{y}$ 和残差 e_i 之间也不相关，即 $\sum_{i=1}^{n} \hat{y}_i e_i = 0$。对此，可以将 OLS 理解为将被解释变量 y_i 分解为两部分：拟合值 $\hat{y}$ 和残差 e_i，且残差和拟合值彼此不相关。

3.1.4 拟合优度

通过普通最小二乘法，我们使样本回归线尽可能接近所有样本观测点，从而对样本数据的总趋势进行拟合。然而，样本回归线与这些观测点到底有多接近呢？或者说，样本回归线对样本数据的拟合程度究竟如何？拟合优度（goodness of fit）正是用来测度回归模型拟合优良程度的一项指标。

拟合优度的评价标准基于如下思路建立：模型所考察的是因变量 y 与自变量 x 之间的关系，其中，x 是以线性方式决定 y 的最主要因素，而除 x 之外的其他因素（即扰动项 ε）均为次要因素。因此，被解释变量 y 的离差 $y_i - \bar{y}$（反映 y 在样本中的分散程度），应主要由解释变量 x 的离差 $x_i - \bar{x}$（反映 x 在样本中的分散程度）所决定。由此，可以用 y 的离差中由 x 的离差所解释的比例，作为评价样本回归线对数据拟合优良程度的标准。

因此，对 y 的离差 $y_i - \bar{y}$ 进行分解，则有：

$$y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \tag{3.12}$$

式中，第一部分 $(\hat{y}_i - \bar{y}) = \hat{\alpha} + \hat{\beta} x_i - \bar{y} = (\bar{y} - \hat{\beta}\bar{x}) + \hat{\beta} x_i - \bar{y} = \hat{\beta}(x_i - \bar{x})$，由 x 的离差所决定；第二

部分$(y_i-\hat{y}_i)=e_i$为残差，表示样本回归线无法解释的部分。显然，如果y的第i个观测值恰好落在样本回归线上，则y的第i个观测值完全可以由样本回归线解释，说明样本回归线对该点实现了完全拟合。为了考察样本回归线对所有样本点的整体拟合程度，通常采用平方和的形式，对上述离差分解进行汇总，得到平方和分解公式：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}e_i^2 \tag{3.13}$$

其中，$\sum_{i=1}^{n}(y_i-\bar{y})^2$称为总平方和（Total Sum of Squares，TSS）；$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$为可解释平方和（Explained Sum of Squares，ESS），即可由模型解释的部分；$\sum_{i=1}^{n}e_i^2$为残差平方和（Residual Sum of Squares，RSS），即模型无法解释的部分。显然，模型可以解释的变动所占比重越大，样本回归线的拟合程度就越好。

因此，拟合优度R^2为：

$$R^2\equiv\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}\text{或}R^2\equiv\frac{\text{ESS}}{\text{TSS}}=1-\frac{\text{RSS}}{\text{TSS}} \tag{3.14}$$

拟合优度R^2也称为可决系数（coefficient of determination）。显然，$0\leqslant R^2\leqslant 1$，且$R^2$越大，样本回归线对数据的拟合程度越高。当$R^2=1$时，模型与样本观测值完全拟合，此时所有样本点均位于样本回归线上；反之，当$R^2=0$时，解释变量x对y不具有任何解释力。

3.1.5　一元线性回归的 Stata 操作及实例

1. 命令的语法格式

最小二乘线性回归分析在 Stata 中使用的命令为 regress，它的语法格式为：

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

其中，depvar 表示被解释变量（或称因变量），且只能有一个；indepvar 表示解释变量（或称自变量），可以为一个或者多个，在一元线性回归中仅包含一个 indepvar（解释变量）；[if]为条件表达式，[in]用于设置样本范围，[weight]用于设置权重。[, options]为可选项，具体含义如表 3.1 所示。

表 3.1　可选项及其含义

[, options]	含　　义
noconstant	模型不包含常数项
hascons	用户自定义常数项
level(#)	设置置信区间水平，默认值为 95%
beta	标准化回归系数
vce(type)	设置估计量的标准差，常用类型包括 ols、robust、cluster、bootstrap、hc2、hc3 等

2. 获得回归模型系数的相关性矩阵

在执行完回归分析之后，可能需要进一步获得回归模型中回归系数的相关性矩阵，其对应的命令为 vce。具体操作为，在完成回归分析后，直接在命令窗口中输入：

```
vce
```

即可得到回归模型回归系数的相关性矩阵。

3. 绘制回归后估计诊断图

除了前文介绍的 Stata 制图工具外，在回归分析中还可以绘制多种回归后估计诊断图，相关命令及其含义如表 3.2 所示。

表 3.2　命令及含义

命　　令	图形含义
rvfplot	画残差与拟合值的散点图
rvpplot varname	画残差与自变量 x 的散点图
cprplot	分量与残差图
acprplot	增强分量与残差图
lvr2plot	杠杆值与残差平方图

4. 使用回归模型进行预测

在许多情况下，建立回归模型不仅是为了基于历史数据解释已发生的现象，更重要的是利用模型来预测未来。本部分重点介绍在创建单个方程模型之后，如何利用模型进行预测。使用回归模型进行预测的命令及其语法格式如下：

```
predict [type] newvar [if] [in] [, single_options]
```

其中，newvar 表示用于存储预测结果的新变量名，[if]为条件表达式，[in]用于设置样本范围，[weight]用于设置权重，[, single_options]为预测选项，其含义如表 3.3 所示。

表 3.3　可选项及其含义

[, single_options]	含　　义
xb	线性预测拟合值
residual 或者 score	残差
rstandard	标准化的残差
rstudent	学生化的残差
stdp	样本内预测标准差
stdf	样本外预测标准差
stdr	残差的标准差
cooksd	Cook 的 D 影响统计量
covratio	COVRATIO 影响统计量
dfits	DFITS 影响统计量
welsch	Welsch 距离
dfbeta(varname)	变量 varname 的 DFBETA

5. 操作实例

沿用例 3.1 的数据集 livebc.dta，将直播收入对数（lnwage）与直播时长（hour）进行一元线性回归分析。

首先打开数据文件：

```
use livebc.dta, clear
```

在进行回归分析之前，可先对数据的基本特征进行考察，输入命令：

```
sum
```

图 3.3 列出了运行 sum 命令后得到的数据集特征。可以看出，该数据集的样本量（Obs）为 601，共包含三个变量，分别为“hour”“wage”和“lnwage”，并给出了各变量的均值（Mean）、标准差（Std. Dev.）、最小值（Min）、最大值（Max）。

```
. sum

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        hour |        601    4.542429    1.558401         1          8
        wage |        601    4631.299    1974.781   1499.74   14073.52
      lnwage |        601    8.360516    .3951939  7.313047   9.552051
```

图 3.3　livebc.dta 数据集的数据特征

进一步地，可以绘制直播收入对数与直播时长之间关系的散点图，并在图中画出离样本点最近的拟合线，初步考察两个变量之间的关系，输入命令：

```
twoway scatter lnwage hour||lfit lnwage hour
```

得到图 3.1，此处不再展示。

在此基础上，进行一元线性回归分析，输入命令：

```
reg lnwage hour
```

其中，被解释变量为直播收入对数（lnwage），解释变量为直播时长（hour），回归结果如图 3.4 所示。

```
. reg lnwage hour

      Source |       SS           df       MS      Number of obs   =       601
-------------+----------------------------------   F(1, 599)       =    188.36
       Model |  22.4174232         1  22.4174232   Prob > F        =    0.0000
    Residual |  71.2895222       599  .119014227   R-squared       =    0.2392
-------------+----------------------------------   Adj R-squared   =    0.2380
       Total |  93.7069454       600  .156178242   Root MSE        =    .34498

------------------------------------------------------------------------------
      lnwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        hour |   .1240332   .0090374    13.72   0.000     .1062843    .1417821
       _cons |   7.797104   .0433968   179.67   0.000     7.711876    7.882333
------------------------------------------------------------------------------
```

图 3.4　线性回归分析的结果

从回归结果中可以获得多方面信息。首先，右上角显示样本观测值数量为 601（Number of obs =

601），模型的拟合优度（R-squared）为 0.2392；模型的 F 统计量为 F (1，599) =188.36，对应的 p 值（Prob > F）= 0.0000，说明模型整体在统计意义上非常显著。相关统计量的含义将在 3.2 节中进一步讲述。

从左上角可以看到，总平方和 TSS（Total）为 93.7069454，其中可解释平方和 ESS（Model）为 22.4174232，不可解释部分 RSS（Residual）为 71.2895222，由此可以计算得到模型的拟合优度。回归结果下半部分显示了具体的参数估计结果，其中，“Coef.”表示回归系数，“Std. Err.”为系数的标准误，“t”“P>|t|”“[95% Conf.Interval]”分别表示 t 值、p 值及 95%的置信区间，“_cons”表示常数项。

根据回归结果，可以写出样本回归方程为：

$$\text{lnwage} = 7.797 + 0.124*\text{hour} \tag{3.15}$$

该结果表明，美妆博主每增加一小时的直播时间，其直播收入平均提高约 12.4%，说明“多劳多得”的规律在美妆直播界同样适用。

在本例中，还可以通过得到的式（3.15）进行预测，即在给定解释变量 hour 的情况下，预测被解释变量 lnwage 的拟合值，从而将分析结果外推至其他主播或未来时点。对被解释变量的拟合值进行预测，输入命令：

```
predict yhat, xb
```

由此生成被解释变量的拟合值，变量名为 yhat，预测结果如图 3.5 所示。

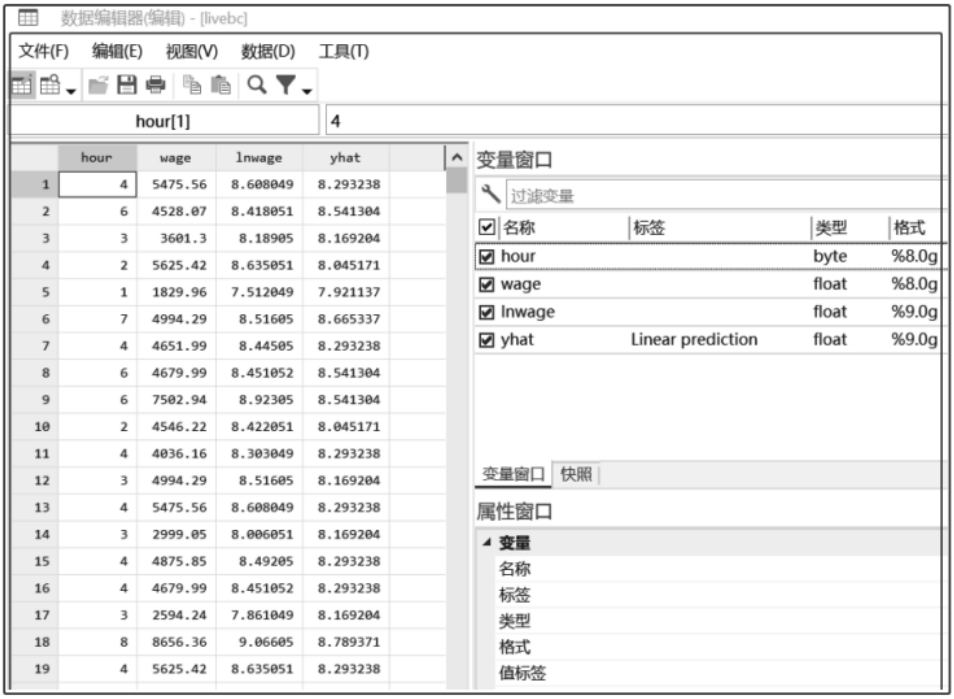

图 3.5　对回归模型的预测结果

3.2　多元线性回归

下载资源:\sample\chap03\正文\earning.dta

3.2.1　多元线性回归模型

现实中的经济问题往往较为复杂，一个经济变量通常会同时受到多个变量的影响。比如在例 3.1 中，直播收入除了会受到直播时长的影响外，显然还会受到直播内容质量、粉丝基数、个人能力等因素的影响。然而，这些因素在式（3.2）中都被纳入了扰动项中，这显然会影响模型的估计结果。因此，我们需要将简单的一元线性回归模型推广到包含多个解释变量的模型，这样的模型被称为多元线性回归模型。

多元线性回归模型参数估计的基本原理与一元线性回归模型相同。因此，多元线性回归模型可以看作是一元线性回归模型的拓展。

多元线性回归模型的一般形式为：

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_K x_{iK} + \varepsilon_i \tag{3.16}$$

其中，y_i 为被解释变量，x_{i1} 为第 1 个解释变量，x_{i2} 为第 2 个解释变量，以此类推，共有 K 个解释变量；β_0 为常数项，$\beta_1,\beta_2,\cdots,\beta_k$ 为回归参数或偏回归系数（partial regression coefficient），表示在其他解释变量保持不变的情况下，$x_j\ (j=1,2,\cdots,K)$ 每变动 1 个单位时，Y 的均值 $E(Y)$的变动幅度；ε_i为扰动项。下标 i 表示第 i 个个体，$i=1,\cdots,n$，其中 n 为样本容量。

由式（3.16）可知，多元线性回归涉及多个变量和多个需要估计的未知参数，而且每个变量都有多个观测数据，其运算和表达式较为复杂和烦琐。因此，在实际应用中，一般采用矩阵方式进行表示。式（3.17）包含 n 个样本观测值，可以将其写成如下方程组的形式：

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_K x_{1K} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_K x_{2K} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_K x_{nK} + \varepsilon_n \end{aligned} \tag{3.17}$$

并进一步写成如下矩阵形式：

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

其中，

$\boldsymbol{Y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n\times 1}$ 为被解释变量构成的 $n\times 1$ 维列向量。

$\boldsymbol{\beta}=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}_{(K+1)\times 1}$ 为未知参数 $\beta_0,\beta_1,\cdots,\beta_K$ 构成的$(K+1)\times 1$ 维列向量。

$\boldsymbol{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n\times 1}$ 为 n 个随机扰动项构成的 $n\times 1$ 维列向量。

$\boldsymbol{X}=\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1K} \\ 1 & x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}_{n\times(K+1)}$ 是解释变量 $x_1,x_2,\cdots,x_K$ 的 n 个观测值构成的 $n\times(K+1)$维数据矩阵。

因此，由式（3.16）表示的 n 个随机方差的矩阵表达式为：

$$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon} \tag{3.18}$$

3.2.2 古典线性回归模型的假定

为了使参数估计量具有良好的统计性质，需要对古典线性回归模型作出若干基本假定。对于一元线性回归模型和多元线性回归模型而言，这些假定在形式和内容上并无实质性差别[1]。

假定 1：参数线性假定。

总体模型为

$$y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_Kx_{iK}+\varepsilon_i$$

即被解释变量 y_i 是参数 $\beta_1,\beta_2,\cdots,\beta_K$ 的线性函数。这意味着每个解释变量 x_{ik} 对 y_i 的边际影响均为常数 β_k。

假定 2：扰动项的条件均值为零。

在给定自变量任取值的条件下，扰动项 ε_i 的期望值为零。即

$$E(\varepsilon_i|\boldsymbol{X})=E(\varepsilon_i|x_1,\cdots,x_K)=0 \quad （i=1,\cdots,n）$$

该假定意味着扰动项 ε_i 与所有解释变量都不相关。

[1] 古典线性回归模型的假定主要参考了陈强《计量经济学及 Stata 应用（第二版）》，高等教育出版社，2023；Wooldridge，《Introductory Econometrics:A Modern Approach》，4th edition. Cengage Learning，2009。

其矩阵形式为：

$$E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = 0 \tag{3.19}$$

假定 3：不存在完全多重共线性。

即在样本中，没有一个解释变量为常数，也不存在某个解释变量是另一个解释变量的倍数，解释变量之间也不存在线性关系。也就是说，不存在多余的解释变量，此时数据矩阵 $\boldsymbol{X}$ 的列秩等于列数，$\boldsymbol{X}$ 为满列秩矩阵。

举个例子，在考察大学班级拔河成绩的影响因素时，如果解释变量中同时包含班级总人数、男生人数和女生人数，则存在完全多重共线性。因为班级总人数等于男生人数与女生人数之和，此时有一个解释变量是多余的：我们总可以用其中任意两个变量线性表示第三个变量。

假定 4：扰动项具有同方差和无序列相关性。即

$$\begin{aligned} &\mathrm{Var}(\varepsilon_i | x_1, \cdots, x_K) = \sigma^2 \qquad (i = 1, \cdots, n) \\ &\mathrm{Cov}(\varepsilon_i, \varepsilon_j | x_1, \cdots, x_K) = 0, \ i \neq j \quad (i, j = 1, \cdots, n) \end{aligned}$$

同方差的假定表明，扰动项 ε_i 的方差不依赖于任何一个解释变量，且不同观测个体的扰动项方差相等；无序列相关性的假定表明，不同观测值的扰动项彼此不相关。

该假定的矩阵形式为：

$$\begin{aligned} \mathrm{Var}(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) &= E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \boldsymbol{X}) \\ &= E\begin{pmatrix} \varepsilon_1^2 & \cdots & \varepsilon_1\varepsilon_n \\ \vdots & & \vdots \\ \varepsilon_n\varepsilon_1 & \cdots & \varepsilon_n^2 \end{pmatrix} \mid \boldsymbol{X} \\ &= \begin{pmatrix} \sigma^2 & \cdots & o \\ \vdots & & \vdots \\ o & \cdots & \sigma^2 \end{pmatrix} \\ &= \sigma^2 \boldsymbol{I}_n \end{aligned} \tag{3.20}$$

其中，$\boldsymbol{I}_n$ 为 n 阶单位矩阵。

假定 5：扰动项服从正态分布。即

$$\varepsilon_i \mid x_1, \cdots, x_K \sim N\left(0, \sigma^2\right)$$

该假定表明，扰动项 ε_i 的条件分布为均值为零、方差为 σ^2 的正态分布。作出该假定后，可以自然推出假定 2 和假定 4。

其矩阵形式表示为：

$$\boldsymbol{\varepsilon} \mid \boldsymbol{X} \sim N(0, \sigma^2 \boldsymbol{I}_n) \tag{3.21}$$

3.2.3　多元线性回归模型的参数估计

对于多元线性回归模型，其参数估计最常用的方法依然是普通最小二乘法（OLS），即寻找最

优的 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_K$，使得残差平方和最小化。随机抽取一组容量为 n 的样本观测值 $\{(x_{i1},x_{i2},\cdots,x_{iK},y_i):i=1,\cdots,n\}$，如果样本回归函数的参数估计值已经得到，则有：

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1x_{i1}+\hat{\beta}_2x_{i2}+\cdots+\hat{\beta}_Kx_{iK} \tag{3.22}$$

OLS 估计量所对应的最小化问题则为：

$$\min_{\hat{\beta}_0,\cdots,\hat{\beta}_K}\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2=\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}\right)^2 \tag{3.23}$$

由微积分知识可知，只需对式（3.23）关于参数向量 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_K$ 求一阶偏导数，并令其值为零，即可得到参数估计值所满足的正规方程组：

$$\begin{cases}\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}\right)=0\\ \sum_{i=1}^{n}x_{i1}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}\right)=0\\ \sum_{i=1}^{n}x_{i2}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}\right)=0\\ \vdots\\ \sum_{i=1}^{n}x_{ik}\left(y_i-\hat{\beta}_0-\hat{\beta}_1x_{i1}-\hat{\beta}_2x_{i2}-\cdots-\hat{\beta}_Kx_{iK}\right)=0\end{cases} \tag{3.24}$$

解由 k+1 个方程组成的线性代数方程组，即可得到 k+1 个参数的估计值 $\hat{\beta}_j$，$j=0,1,2,\cdots,K$。上述正规方程还可以写成矩阵形式：

$$\begin{bmatrix}n & \sum x_{i1} & \cdots & \sum x_{iK}\\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{iK}\\ \vdots & \vdots & \vdots & \vdots\\ \sum x_{1k} & \sum x_{iK}x_{i1} & \cdots & \sum x_{iK}^2\end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_K\end{bmatrix}=\begin{bmatrix}1 & 1 & \cdots & 1\\ x_{11} & x_{21} & \cdots & x_{n1}\\ \vdots & \vdots & \vdots & \vdots\\ x_{1K} & x_{2K} & \cdots & x_{nK}\end{bmatrix}\begin{bmatrix}y_1\\ y_2\\ \vdots\\ y_n\end{bmatrix} \tag{3.25}$$

即

$$(\boldsymbol{X}'\boldsymbol{X})\hat{\boldsymbol{\beta}}=\boldsymbol{X}'\boldsymbol{Y} \tag{3.26}$$

由 $\boldsymbol{X}$ 为满列秩可知，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在，从而可得 OLS 估计量：

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{3.27}$$

3.2.4 校正拟合优度

对于多元线性回归模型，平方和分解公式依然成立，被解释变量的离差平方和 TSS 可以分解为模型可以解释的部分 ESS 和模型无法解释的部分 RSS。根据 3.1.4 节的定义，拟合优度 R^2 为：

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

但在实际应用中，我们会发现，如果在模型中增加解释变量的个数，R^2 往往只增不减，因为至少可以使新增的解释变量对模型没有任何解释力。此时，由解释变量个数所引起的 R^2 上升，已与模型拟合优劣程度无关。因此，对于多元线性回归模型而言，拟合优度不再是衡量模型拟合优良程度的合适指标，必须对其加以调整。

在样本量一定的情况下，增加解释变量的个数必然会减少自由度。为此，我们引入被解释变量的离差平方和与残差平方和的自由度，以剔除解释变量个数对拟合优度的影响。定义校正拟合优度（adjusted R^2）$\overline{R^2}$ 为：

$$\overline{R^2} \equiv 1 - \frac{\mathrm{RSS}/(n-K-1)}{\mathrm{TSS}/(n-1)} \tag{3.28}$$

其中，$(n-K-1)$为残差平方和的自由度，$(n-1)$为离差平方和的自由度。根据式（3.28），如果解释变量个数 K 增加但并未提高模型的解释能力，则校正拟合优度 $\overline{R^2}$ 可能不增反降。因此，校正拟合优度可以看作是对解释变量过多的一种惩罚。

在多元线性回归分析中，校正拟合优度常被用来判断是否应将新的解释变量引入模型中。如果新解释变量的引入使得 $\overline{R^2}$ 增大，则表明该变量对被解释变量的变动具有解释力，应该引入模型；反之，则无须引入模型。

3.2.5　OLS 估计量的性质

通过 OLS 方法得到样本回归函数后，还需要考察 OLS 估计量的统计性质，以分析估计量的估计效果。在古典线性回归模型的各项假定下，OLS 估计量具有以下良好性质。

1. 线性性

线性性是指 OLS 估计量可以表示为被解释变量观测值 $\boldsymbol{Y}$ 的线性组合，也称 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 为线性估计量。根据 OLS 估计量的表达式

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{C}\boldsymbol{Y}$$

其中，$\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ 仅与给定的 $\boldsymbol{X}$ 有关。因此，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 是 Y 的线性函数（线性估计量）。

2. 无偏性

OLS 估计量 $\hat{\boldsymbol{\beta}}$ 具有无偏性，即满足

$$E(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{\beta}$$

这意味着 OLS 估计量不会系统性地高估或低估真实参数 $\boldsymbol{\beta}$。

由于

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$$

可得

$$E(\hat{\boldsymbol{\beta}} \mid \boldsymbol{X}) = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon} \mid \boldsymbol{X}) = \boldsymbol{\beta} \tag{3.29}$$

根据扰动项条件均值为零 $E(\boldsymbol{\varepsilon}\mid\boldsymbol{X})=0$ 的假设，上式成立。

3. 有效性

具有最小方差的估计量被认为是最有效率的估计量。多元回归模型包含多个参数，涉及参数估计量之间的相关性。因此，有效性不仅要求单个参数估计量的方差最小，还要求不同参数估计量之间的协方差也达到最小。

在给定 $\boldsymbol{X}$ 的前提下，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的方差-协方差矩阵为：

$$\begin{aligned}\operatorname{Var}(\hat{\boldsymbol{\beta}}\mid\boldsymbol{X})&=\operatorname{Var}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}\right]\\&=\operatorname{Var}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\left(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}\right)\right]\\&=\operatorname{Var}\left[\boldsymbol{\beta}+\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right]\\&=\operatorname{Var}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}\right]\\&=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\operatorname{Var}\left(\boldsymbol{\varepsilon}\right)\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]'\\&=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\left[\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right]'\\&=\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\end{aligned}\tag{3.30}$$

可以证明，在一定条件范围内，式（3.30）所给出的 OLS 估计量方差是最小的[1]。

4. 高斯-马尔科夫定理（Gauss-Markov Theorem）

在 3.2.2 节假定 4 成立的条件下，最小二乘估计量在所有线性无偏估计中具有最小的方差，即 OLS 估计量是最佳线性无偏估计量（Best Linear Unbiased Estimator，BLUE）。

5. 对扰动项方差的无偏估计

对于扰动项方差 $\sigma^2=\operatorname{Var}(\varepsilon_i)$，在假定 4 之下，有 $E(\hat{\sigma}^2)=\sigma^2$，即对扰动项方差的估计为无偏估计。

但由于 $\{\varepsilon_1,\cdots,\varepsilon_n\}$ 不可观测，实际应用中通常以残差 $\{e_1,\cdots,e_n\}$ 作为其实现值，从而得到 σ^2 的估计值：

$$\hat{\sigma}^2=s^2\equiv\frac{1}{n-K-1}\sum_{i=1}^{n}e_i^2\tag{3.31}$$

其中，n-K-1 为自由度。$s=\sqrt{s^2}$ 被称为回归方程的标准误差，用以衡量回归方程中扰动项的波动程度。

因此，OLS 估计量 $\hat{\boldsymbol{\beta}}$ 的方差-协方差矩阵 $\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ 可以用 $s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ 来估计。设 $\hat{\beta}_k$ 的估计方差为 $s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}$，$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}$ 则为矩阵 $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ 主对角线上的第 k 个元，则 OLS 估计量 $\hat{\beta}_k$ 的标准误差（standard error），简称标准误，记为 $\operatorname{SE}\left(\hat{\beta}_k\right)$，则为：

$$\operatorname{SE}\left(\hat{\beta}_k\right)\equiv\sqrt{s^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}}\tag{3.32}$$

[1] 证明可以参考：《计量经济学（第五版）学习指南与练习》，潘文卿，李子奈编著，高等教育出版社，2021，本书不再证明。

在线性回归模型的估计中，不仅需要给出参数的估计值 $\hat{\beta}_k$，还必须报告相应的标准误，才能判断参数估计的准确程度。

3.2.6　单个系数显著性的 *t* 检验

在多元线性回归模型的参数估计完成后，即得到了样本回归函数。此时，还需要进一步对样本回归函数进行统计检验，以判断参数估计的可靠程度。前面已经介绍了拟合优度，用以考察模型整体拟合的优良程度。本小节将介绍回归参数的显著性检验，即 t 检验，该检验通过考察单个参数 β_k 是否等于 0，从而判断相应的解释变量是否对被解释变量具有显著影响。

1. 参数估计量的概率分布

为了对多元线性回归模型的参数估计量进行统计检验，首先需要明确参数估计量的概率分布特征。根据式（3.29），可得

$$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon} \tag{3.33}$$

显然，$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$ 为扰动项 $\boldsymbol{\varepsilon}$ 的线性函数。根据扰动项服从正态分布的假定 $\boldsymbol{\varepsilon}\,|\,\boldsymbol{X}\sim N\left(0,\ \sigma^2\boldsymbol{I}_n\right)$，可知 $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\,|\,\boldsymbol{X}$ 也服从正态分布。进一步，由于参数 $\boldsymbol{\beta}$ 可视为常数，且 $E(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X})=\boldsymbol{\beta}$，$\mathrm{Var}(\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X})=\sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ 可知，$\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}$ 服从正态分布

$$\hat{\boldsymbol{\beta}}\,|\,\boldsymbol{X}\sim N\left(\boldsymbol{\beta},\ \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right) \tag{3.34}$$

于是，在给定样本的条件下，第 k 个参数估计量服从如下形式的正态分布

$$\hat{\beta}_k\,|\,\boldsymbol{X}\sim N\left(\beta_k,\ \sigma^2\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}\right) \tag{3.35}$$

其中，$\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}$ 为矩阵 $\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}$ 主对角线上的第 k 个元素。

2. 单个系数的显著性检验：*t* 检验

在确定参数估计量的概率分布后，即可依据样本回归函数对总体参数进行显著性检验。单个系数的显著性检验（即 t 检验）的基本步骤如下：

（1）提出原假设。在回归分析中，关注的是某个解释变量 X 是否对解释变量 Y 具有显著影响，因此针对变量 x_k 的原假设设定为

$$H_0:\beta_k=0$$

即检验变量 x_k 的系数 β_k 是否显著地不等于 0。

假设检验的本质是一种反证法：首先假定原假设成立，然后考察在原假设成立的前提下，不太可能发生的“小概率事件”在一次抽样中是否会出现。如果这种小概率事件在样本中被观测到，则说明原假设不可信，应当拒绝原假设，而接受备择假设

$$H_1:\beta_k\neq 0$$

因此，通常所说的“某个系数是显著的”，是指该系数显著地不等于 0，即拒绝了原假设。该检验方法属于沃尔德检验（Wald test）。

（2）计算 t 统计量。已知 $\hat{\beta}_k \mid \boldsymbol{X}$ 服从正态分布，对其进行标准化可得

$$z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}}} \sim N(0,1)$$

由于 σ^2 未知，同样以 σ^2 的估计量 s^2 来替代，可以得到 t 统计量

$$t = \frac{\hat{\beta}_k - \beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{s^2 \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}_{kk}}} \sim t\left(n-K-1\right) \tag{3.36}$$

其中，SE($\hat{\beta}_k$)为 $\hat{\beta}_k$ 的标准误。

t 统计量的分子度量了估计量（$\hat{\beta}_k$）与参数值（β_k）之间的偏离程度，分母以标准误 SE($\hat{\beta}_k$)作为偏离程度的度量单位，因此 t 统计量衡量的是该距离相当于多少个标准误。根据式（3.36）计算 t 统计量，记数值为 t_k。如果原假设 H_0 为真，而$|t_k|$取值较大，则意味着该偏离程度是标准误的较大倍数，从而使原假设 H_0 显得不可信。

（3）规定显著性水平 α，并计算临界值 $t_{\alpha/2}(n-k-1)$。该临界值表示某随机变量大于 $t_{\alpha/2}(n-k-1)$或小于$-t_{\alpha/2}(n-k-1)$的概率均为 $\alpha/2$。在实践中，α 通常取 5%，有时也使用 1%或 10%。

（4）比较$|t_k|$与 $t_{\alpha/2}(n-k-1)$的大小。如果$|t_k| \geqslant t_{\alpha/2}(n-k-1)$，则 t_k 落入拒绝域，故拒绝原假设 H_0；反之，如果$|t_k| < t_{\alpha/2}(n-k-1)$，则 t_k 落入接受域，故接受原假设 H_0，如图 3.6 所示。

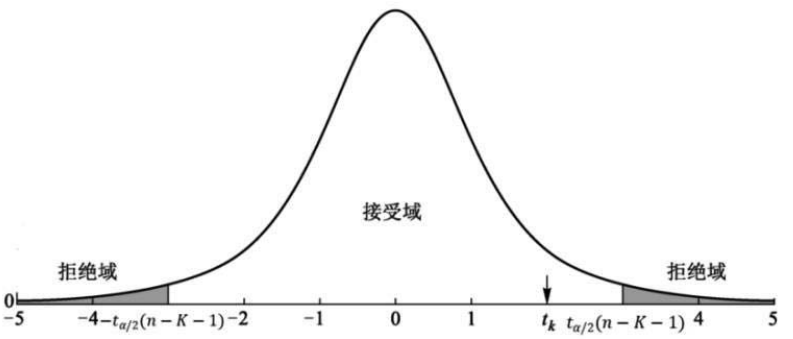

图 3.6　双边 t 检验的临界值与拒绝域

3. 单个系数的显著性检验：p 值

在实际操作中，更多时候通过 p 值来检验单个系数的显著性。在双边检验中，给定 t 统计量的样本观测值 t_k，则 p 值（p-value）为

$$p \equiv P(|T| > |t_k|) \tag{3.37}$$

其中，随机变量 $T \sim t(n-k-1)$。根据定义可知，p 值衡量的是比$|t_k|$更大的 t 分布两端尾部概率之和。如果 p 值为 0.05，则恰好可以在 5%的显著性水平上拒绝原假设，但无法在小于 5%的显著性水平上拒绝原假设。因此，p 值就是原假设可以被拒绝的最小显著性水平。显然，p 值越小，越倾向于拒绝

原假设。例如，p=0.02 时，可在 5%的显著性水平上拒绝原假设，但无法在 1%的显著性水平上拒绝原假设。

4. 参数的置信区间

我们用 t 检验来考察总体参数是否等于某一给定数值，其采用的是点估计方法，但并未指出总体参数可能的取值范围。因此，还可以采用参数的区间估计，以给出在一定概率水平下真实参数的取值范围。假设置信度为$(1-\alpha)$，意味着置信区间（confidence interval）覆盖真实参数β_k的概率为$(1-\alpha)$，该置信区间即为相对应的取值范围。

由于 $t=\dfrac{\hat{\beta}_k-\beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)}\sim t\left(n-K-1\right)$，故 t 统计量落入接受域的概率为$(1-\alpha)$：

$$P\left\{-t_{\alpha/2}<\frac{\hat{\beta}_k-\beta_k}{\mathrm{SE}\left(\hat{\beta}_k\right)}<t_{\alpha/2}\right\}=1-\alpha$$

将上式中的不等式变形可知，在$(1-\alpha)$的置信度下，β_k 的置信区间为

$$\left[\hat{\beta}_k-t_{\alpha/2}\mathrm{SE}\left(\hat{\beta}_k\right),\ \hat{\beta}_k+t_{\alpha/2}\mathrm{SE}\left(\hat{\beta}_k\right)\right]$$

该置信区间以 $\hat{\beta}_k$ 为中心，以 $t_{\alpha/2}\mathrm{SE}(\hat{\beta}_k)$ 为区间半径。显然，标准误 $\mathrm{SE}(\hat{\beta}_k)$ 越大，对 β_k 的估计越不准确，相应的置信区间也就越宽。

3.2.7　方程显著性的 F 检验

除了单个系数的显著性之外，我们通常也希望检验整个线性回归方程是否显著，即考察除常数项外，模型中所有解释变量的回归系数是否都为零。这种检验称为方程的显著性检验——F 检验，其目的在于检验模型中所有解释变量与被解释变量之间的线性关系在总体上是否显著成立。

方程显著性的 F 检验是对模型 $y_i=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\cdots+\beta_Kx_{iK}+\varepsilon_i$ 中各解释变量x_K的参数是否均显著不为零进行检验。按照假设检验的基本原理，其检验步骤如下。

（1）提出原假设。

$$H_0:\beta_1=\cdots=\beta_K=0$$

此原假设等价于对 K 个约束条件进行联合检验，因此，原假设也可写为

$$H_0:\beta_1=0,\beta_2=0,\cdots,\beta_K=0$$

其备择假设为

$$H_1:\beta_k(k=1,2,\cdots,K)\text{不全为}0$$

（2）计算 F 统计量。

F 检验的思想来源于离差平方和的分解公式 TSS=ESS+RSS。我们知道，可解释平方和 ESS 是由模型所解释的部分，反映了所有解释变量 $\boldsymbol{X}$ 对被解释变量 $\boldsymbol{Y}$ 的线性作用的结果。因此，可以考虑 ESS 与残差平方和 RSS 的比值 ESS/RSS。如果该比值较大，意味着所有解释变量 $\boldsymbol{X}$ 对被解释变量 $\boldsymbol{Y}$ 的解释程度较高，可以认为总体存在线性关系；反之，则总体上可能不存在线性关系。因此，可以通过 ESS/RSS 这一比值的大小对总体线性关系进行推断。

因此，构建 F 统计量

$$F=\frac{\mathrm{ESS}/K}{\mathrm{RSS}/(n-K-1)}$$

该统计量服从自由度为（K，n−K−1）的 F 分布。

（3）规定显著性水平 α，并计算临界值 $F_{\alpha}(K,n-K-1)$，该临界值表示某随机变量大于该临界值的概率恰好等于 α。同样，α 通常取 5%，有时也使用 1%或 10%。

（4）比较 F 统计量与临界值 $F_{\alpha}(K,n-K-1)$ 的大小。如果 F 统计量大于临界值，即落入拒绝域，则拒绝原假设 H_0，表明在 α 显著性水平上模型的线性关系在总体上显著成立；反之，如果 F 统计量小于临界值，则接受原假设 H_0。

在实际操作中，Stata 的回归结果通常会给出 F 检验的 p 值。此时仍可根据 p 值与显著性水平α的大小进行判断：如果 p 值小于α，则拒绝原假设，说明方程的线性关系在统计意义上是显著的。

3.2.8 解释变量个数的选择

在模型设定拟合过程中，如果增加解释变量的个数，可以在一定程度上提升拟合效果，或者说提升模型的解释能力，但解释变量的增加也可能带来过度拟合（overfitting）的情况。因此，为了在模型拟合优度与复杂性之间取得平衡，帮助研究者合理选取解释变量的数目，计量经济学家提出了信息准则（information criteria）的概念。

信息准则一方面鼓励模型具有较好的拟合效果，另一方面则对解释变量过多的情形施加惩罚性约束。在实际应用中，无论采用何种信息准则，其基本判别准则都是：信息准则的值越小，说明模型拟合得越好。假设有 n 个备选模型，则可以一次性计算出这 n 个模型对应的信息准则值，并选取信息准则值最小的那个模型作为最优模型。

常用的信息准则包括赤池信息准则（Akaike’s Information Criterion，AIC）和贝叶斯−施瓦茨信息准则（Schwarz’s Bayesian Information Criterion，SBIC 或者 SIC）。二者的计算公式分别为：

$$\mathrm{AIC}=\ln\left(\frac{\mathrm{SSR}}{n}\right)+\frac{2}{n}K$$

$$\mathrm{BIC}=\ln\left(\frac{\mathrm{SSR}}{n}\right)+\frac{\ln n}{n}K$$

从上述公式可以看出，两种信息准则右端的第一项均为对模型拟合优度的奖励项，第二项则为对解释变量过多的惩罚项；二者的主要差别体现在第二项上。一般而言，存在 $\ln n>2$，SBIC 准则对解释变量过多的惩罚力度通常大于 AIC 准则的惩罚力度。因此，SBIC 准则更强调模型的简洁性。

3.2.9　多元线性回归：OLS 的渐进性

为了获得 OLS 估计量的良好性质，我们在 3.2.2 节提出了一系列假定。然而，随着样本容量的不断增大，部分假定与现实情况之间可能存在较大偏差，即这些假定在实际应用中显得过于严格。例如，假定 5 要求扰动项服从正态分布，但在现实中，扰动项不服从正态分布的情况比比皆是。以受教育年限为例，该变量往往受到教育体系设定的影响，当考察受教育年限与工资收入之间的关系时，其扰动项通常很难满足正态分布假定。

因此，随着样本容量趋向于无穷大，有必要对前述假定加以放松，并进一步考察大样本条件下 OLS 估计量的性质，即 OLS 估计量的渐近性质（asymptotic properties）。此时，所依据的理论为渐近理论，也称为大样本理论。

1. 大样本下的 OLS 假定

在大样本条件下，概率统计中的大数定律与中心极限定理是放松原有假定的两个重要理论依据。大数定律表明，当样本容量足够大时，样本均值将趋近于总体均值；中心极限定理则证明，当样本容量足够大时，无论随机序列 $\{x_n\}_{n=1}^{\infty}$ 服从什么分布，其样本均值 $\bar{x}$ 都近似服从正态分布。因此，在大样本条件下，古典线性回归模型中的假定 5——即扰动项服从正态分布的假定——不再是必要条件。

其次，小样本条件下的假定 2 要求扰动项 ε_i 与所有解释变量均不相关；而在大样本下，仅需扰动项与同期解释变量不相关即可。即对于任意个体 i 和变量 k，有 $E\left(x_{ik}\varepsilon_i=0\right)$，此假定称为前定解释变量假定（predetermined regressors assumption）。

再次，对于小样本条件下的假定 4，即扰动项具有同方差且不存在序列相关的假定，在大样本条件下也可以得到放松。在满足大数定律与中心极限定理的前提下，随机序列（也称随机过程），$\{y_i,x_{i1},\cdots,x_{iK}\}$ 为渐近独立的平稳过程。这意味着，随着样本容量的增大，变量之间的相关性在极限处消失，且其概率分布不随时间变化。因此，小样本条件下关于同方差性和无序列相关的假设可以被放松。

此外，小样本条件下的假定 1 和假定 3 在大样本分析中依然需要满足。

综上所述，在大样本条件下，OLS 的基本假定包括 4 条：参数线性假定、不存在完全多重共线性假定、前定解释变量假定，以及随机过程 $\{y_i,x_{i1},\cdots,x_{iK}\}$ 为渐近独立的平稳过程。

2. OLS 的渐近性质

在上述假定下，OLS 估计量依然具有一系列良好的渐近性质。

（1）一致性，对于所有的 k=1,2,⋯,K，OLS 估计量 $\hat{\beta}_k$ 是 β_k 的一致估计量。根据前定解释变量假设，可以推出该渐近性质[1]，其数学表达式为 $\underset{n\to\infty}{p\lim}\hat{\beta}=\beta$，从该表达式可以看出，该性质蕴含了 OLS 估计量在极限处的无偏性和一致性。

（2）OLS 的渐近正态性。OLS 估计量 $\hat{\boldsymbol{\beta}}$ 服从渐近正态分布，即 $\sqrt{n}\left(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\xrightarrow{d}N\left(0,\text{Avar}\left(\hat{\boldsymbol{\beta}}\right)\right)$，其中 Avar($\hat{\boldsymbol{\beta}}$)为 $\hat{\boldsymbol{\beta}}$ 的渐近协方差矩阵。该数学式表明，当样本容量足够大时，$\hat{\boldsymbol{\beta}}$ 近似服从正态分布。

（3）OLS 的渐近有效性。对于所有 k=1,2,⋯,K，OLS 估计量 $\hat{\beta}_k$ 具有最小的渐近方差。该性质

[1] 具体证明参见陈强《计量经济学及 Stata 应用（第二版）》，高等教育出版社，2023，第 6 章 6.8 节。

的严格证明较为复杂，已超出本书讨论范围，我们在此仅作说明：即使在大样本条件下，OLS 估计量依然保持有效性。

3.2.10 多元线性回归的 Stata 操作及实例

1. 命令的语法格式

多元线性回归分析的命令是 regress，它的语法格式为：

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

其中，depvar 表示被解释变量（或称因变量），indepvar 表示解释变量（或称自变量）。depvar 只有一个，indepvar 可包含多个解释变量，在命令中依次输入即可。[if]为条件表达式，[in]用于设置样本范围，[weight]用于设置权重。[, options]为可选项，同表 3.1。

2. 对模型系数进行假设检验

在很多情况下，在执行完回归分析之后，我们有必要对模型进行进一步的假设检验。事实上，在回归分析结果中已经给出了针对模型整体的 F 检验，以及针对各解释变量和常数项回归系数的 t 检验。在此基础上，还可以使用 test 命令，对最近一次拟合模型中的回归系数的简单性假设与复合线性假设进行 Wald 检验，以检验系数是否为 0 或是否满足某些线性约束。

test 命令包括 5 种形式，分别说明如下。

1）test coeflist

该命令用于检验当前方程中 coeflist（系数名称列表）中列出的一个或多个系数是否都等于 0。

2）test exp=exp[=…]

该命令用于检验由表达式 exp=exp[=…]所定义的线性假设是否成立。

3）test [eqno] [: coeflist]

该命令用于在多方程模型中，检验指定方程 eqno 中系数列表（coeflist）所列出的系数是否都等于 0。

4）test [eqno = eqno [= ...]] [: coeflist]

该命令用于在多方程模型中，检验不同方程 eqno 中同一变量列表 coeflist 的系数是否相等。

5）testparm varlist [, testparm_options]

该命令用于批量检验变量列表 varlist 中所有系数是否都为 0，其功能类似于 test coeflist。

此外，如果检验为非线性假设，则需要用到 testnl 命令，该命令的语法格式为：

```
testnl exp=exp[=exp...] [, options]
```

其中，exp=exp[=exp...]表示系数之间的非线性关系式。

3. 使用回归模型进行预测

3.1.5 节介绍了在创建一元线性回归模型之后的预测，接下来介绍创建多元线性回归方程之后的预测，其命令及其语法格式如下：

```
predict [type] newvar [if] [in] [, multiple_options]
```

相关字段含义及可选项与单个方程模型基本一致。

4. 在回归方程中自动剔除不显著变量（逐步回归法）

stepwise 命令的语法格式为：

```
stepwise [, options ] :regress depvar [indepvars]
```

或

```
sw regress depvar [indepvars],[, options ]
```

其中，sw regress depvar [indepvars]为回归分析命令，[, options]选项及其含义如表 3.4 所示。

表 3.4　可选项及其含义

[, options]	含　　义
* pr(#)	删除解释变量的显著性水平
* pe(#)	增加解释变量的显著性水平
forward	前向搜寻法
hierarchical	分层搜寻法
lockterm1	保留第 1 项
lr	使用似然比统计量代替 Wald 统计量

在使用 stepwise 命令时，搜寻的方法和顺序同样较为重要。Stata 提供了 6 种常用的搜寻方法，用户可根据研究需要选择最为恰当的方法，如表 3.5 所示。

表 3.5　6 种常用的搜寻方法

顺序选项	名　　称	功能和计算逻辑
pe(#)	前向搜寻法	一开始建立只包括常数项的原始模型，然后按显著性水平由高到低加入解释变量，只有当解释变量通过所设置的显著性水平检验时，该变量才能被保留
pe(#) hierarchical	前向分层搜寻法	一开始建立只包括常数项的原始模型，然后按排列顺序逐个加入解释变量；当下一个解释变量无法通过设置的显著性水平检验时，该过程停止，并形成最终的模型
pr(#) pe(#) forward	前向分步搜寻法	第一步和前向搜寻法一致；执行结束后，将模型中未通过检验且显著性水平最低的解释变量剔除后继续进行估计，同时将已被剔除但显著性水平最高的且可通过检验的解释变量重新加入并重新估计，不断重复这一过程，直到这两种计算都无法继续进行
pr(#)	后向搜寻法	一开始建立包括所有解释变量的模型，当显著性水平最低的变量无法通过所设定的显著性水平检验时，剔除该变量并重新估计。不断重复这一过程，直到剩余变量均能通过检验
pr(#) hierarchical	后向分层搜寻法	一开始建立包括所有解释变量的模型，然后总是检验解释变量中的最后一个；当最后一个变量不显著时，剔除该变量并重新估计。不断重复这一过程，直到模型中保留下来的最后一个解释变量能够通过检验

（续表）

顺序选项	名　　称	功能和计算逻辑
pr(#) pe(#)	后向分步搜寻法	第一步和后向搜寻法一致；执行结束后，将已被剔除但显著性水平最高的解释变量重新加入进行估计，同时将仍保留但显著性水平最低的解释变量剔除后重新估计，不断重复这一过程，直到两种计算都无法继续进行

5. 操作实例

以数据集 earning.dta 为例，该数据集包括我国 3008 家企业的经营数据。我们希望考察企业盈利能力会受到哪些因素的影响，因此构建以下模型并进行多元线性回归估计：

$$\text{roa}=\beta_0+\beta_1\text{lnasset}+\beta_2\text{lev}+\beta_3\text{turnover}+\beta_4\text{liquid}+\beta_5\text{list}+\varepsilon$$

其中，被解释变量为企业盈利能力roa（总资产收益率），解释变量包括 lnasset（总资产对数）、lev（资产负债率）、turnover（总资产周转率）、liquid（流动比率）和 list（是否上市）。此外，数据集中还包括 indus（企业所属行业）和 province（企业所属省份）。我们可以打开数据编辑器直接查看数据特征，如图 3.7 所示。

图 3.7　数据特征

1）线性回归分析

打开上述数据文件之后，在主界面的命令窗口中依次输入命令：

```
reg roa lnasset lev turnover liquid list
```

从回归结果（见图 3.8）可以得到以下信息。首先，从右上角可以看出，数据集共有 3008 个观测样本（Number of obs = 3008），模型的拟合优度（R-squared）为 0.1247，校正拟合优度（Adj R-squared）为 0.1232，模型的 F 值(5,3002) =85.51，对应的 p 值（Prob > F）= 0.0000，说明模型整体上存在显著的线性关系。其次，从左上角看，TSS（Total）为 33731.3704，其中可解释部分 ESS（Model）为 4205.04893，不可解释部分 RSS（Residual）即残差平方和为 29526.3215，根据这些数值可以计算拟合优度。

图中下半部分为回归结果。第一列展示了解释变量的名称，其中“_cons”表示常数项；与每个解释变量相对应，回归结果汇报了该解释变量的回归系数（Coef.）、系数的标准误（Std. Err.）、t 值（t）、p 值（$P>|t|$）以及 95%的置信区间（[95% Conf.Interval]）。以解释变量 lnasset 为例，其系数为 0.429663（在实际操作中，通常保留小数点后 2 位或 3 位），所对应的 p 值为 0.000，小于 1%，故在 1 显著性水平上显著，表明解释变量 lnasset 与被解释变量 roa 存在显著的正相关关系；其经济学含义为：企业总资产（lnasset）每增加一个单位，总资产收益率（roa）提高约 42.97 个百分点。

Source	SS	df	MS	Number of obs	=	3,008
				F(5, 3002)	=	85.51
Model	4205.04893	5	841.009786	Prob > F	=	0.0000
Residual	29526.3215	3,002	9.83555012	R-squared	=	0.1247
				Adj R-squared	=	0.1232
Total	33731.3704	3,007	11.2176157	Root MSE	=	3.1362

roa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	.429663	.0524272	8.20	0.000	.3268661	.5324599
lev	-.0962699	.0051734	-18.61	0.000	-.1064137	-.0861261
turnover	.0177878	.0018186	9.78	0.000	.0142219	.0213537
liquid	-.087492	.0196736	-4.45	0.000	-.1260672	-.0489169
list	-.0831874	.2163423	-0.38	0.701	-.5073815	.3410066
_cons	-.024133	.7675246	-0.03	0.975	-1.52906	1.480794

图 3.8　多元线性回归分析的结果

从整体来看，除解释变量 list 不显著外（其 p 值为 0.701，大于 10%），其余解释变量均显著，即对总资产收益率（roa）存在显著的线性影响。

2）绘制回归后估计诊断图

在本例中，我们可以通过绘制残差 e 和被解释变量拟合值的散点图，并标识出纵轴值为零的基准线来进行模型诊断，在命令窗口中输入：

```
rvfplot, yline(0)
```

绘制结果如图 3.9 所示。该图展示了模型预测值和残差的分布情况。图中显示，在一开始拟合值较小时残差多为负值，说明此时模型的预测值高于实际值；随后，残差点多分布在纵轴值为零的直线附近，说明模型在这一区间内的拟合效果良好。

3）回归方程中不包含常数项

根据回归结果可以发现，常数项（_cons）并不显著（其 p 值为 0.975，远大于 10%）。在实际分析中，若常数项无显著统计意义，可以考虑将其删除。操作命令可相应修改为：

```
reg roa lnasset lev turnover liquid list,noc
```

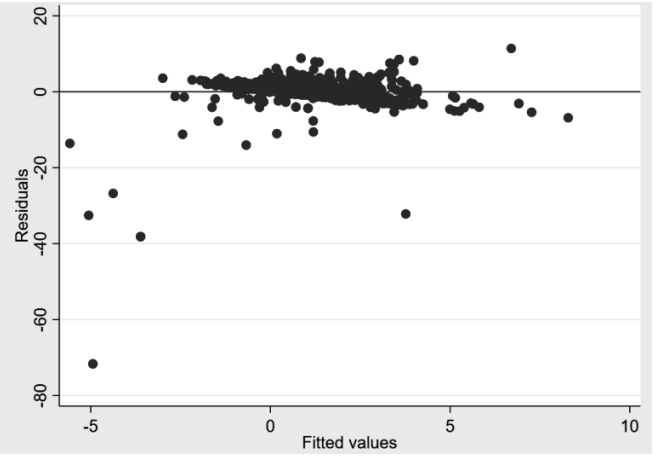

图 3.9　残差对预测值的标绘图

从回归结果图3.10可以发现，模型的校正拟合优度（Adj R-squared）为0.2047，模型的F值(5,3003) =155.80，均比有常数项时的回归模型结果有所提高，说明无常数项的回归模型更符合现实情况。此外，相较于有常数项时的回归模型结果，可以看到，在无常数项的情况下，模型所有解释变量的 t 值均有所提高，尤其是解释变量 lnasset 的 t 值从之前的 8.20 提高到 20.19，说明去掉常数项后，模型中解释变量与被解释变量之间的线性关系变得更加显著。因此，在实际操作中，读者可以通过比较不同回归结果，决定是否在模型中加入常数项。

Source	SS	df	MS	Number of obs	=	3,008
				F(5, 3003)	=	155.80
Model	7659.40343	5	1531.88069	Prob > F	=	0.0000
Residual	29526.3312	3,003	9.83227812	R-squared	=	0.2060
				Adj R-squared	=	0.2047
Total	37185.7346	3,008	12.3622788	Root MSE	=	3.1356

roa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	.4281554	.0212013	20.19	0.000	.3865848	.469726
lev	-.0962665	.0051714	-18.62	0.000	-.1064065	-.0861266
turnover	.0177904	.0018164	9.79	0.000	.0142288	.021352
liquid	-.0877359	.0180774	-4.85	0.000	-.1231812	-.0522906
list	-.0835415	.2160131	-0.39	0.699	-.5070901	.3400071

图 3.10　在回归方程中不包含常数项回归分析结果

4）限定参与回归的样本范围

很多时候，我们希望限定参与回归的样本范围，即仅对部分样本进行回归分析，此时可以使用[if]条件表达式进行设定。例如，我们只想对制造业的企业样本进行回归。首先，从图 3.7 的数据特征中可以发现，indus（企业所属行业）为字符型变量，因此需要先将其转换为数值型变量，操作命令如下：

```
egen ind=group( indus )
```

可以发现，Stata 对不同行业进行了自动编号。通过对比可知，制造业对应的取值为“ind=5”，如图 3.11 所示。

	indus	list	lnasset	lev	turnover	roa	liquid	ind
1	采矿业	1	17.0694	69.1853	83.0373	2.64095	.568882	15
2	制造业	1	16.8161	16.2987	87.6948	18.0971	4.66413	5
3	制造业	1	16.4744	39.3168	43.9175	4.23725	1.33641	5
4	采矿业	1	15.493	66.7976	41.8469	3.04082	.523356	15
5	交通运输、仓储和邮政业	1	15.0786	49.8177	14.325	1.54036	1.21341	1
6	采矿业	1	17.0694	69.1853	83.0373	2.64095	.568882	15
7	交通运输、仓储和邮政业	1	15.0786	49.8177	14.325	1.54036	1.21341	1
8	电力、热力、燃气及水生产和供应业	1	13.6934	52.3032	25.6403	2.04133	.767728	12
9	房地产业	1	14.9151	63.4317	11.9654	3.66513	1.11861	8
10	房地产业	1	15.9724	77.1769	20.8182	.66631	1.79541	8
11	制造业	1	15.1023	119.493	14.9449	-76.6216	.959687	5
12	采矿业	1	14.7268	54.9389	17.5364	.25065	.458202	15
13	房地产业	0	19.039	82.0187	24.0236	2.03921	1.3408	8
14	房地产业	0	16.9289	75.2482	20.6297	1.49447	1.36326	8
15	房地产业	1	16.3909	78.7015	10.5743	.83938	2.47032	8
16	房地产业	0	18.0015	68.9782	26.8177	4.34975	1.54537	8
17	房地产业	0	16.9289	75.2482	20.6297	1.49447	1.36326	8
18	综合	0	16.3108	69.3282	6.25129	.77175	.429678	14
19	交通运输、仓储和邮政业	0	16.0933	76.0249	7.26873	1.35024	1.08394	1
20	综合	0	17.2024	79.9122	3.58667	-2.28181	.982723	14
21	交通运输、仓储和邮政业	1	13.9156	71.4992	50.5354	-2.85155	.826048	1
22	房地产业	1	17.5085	76.5859	20.8293	3.7947	1.40865	8

图 3.11　将字符型变量转换为数值型变量

在此基础上，我们对制造业的企业样本进行回归，输入操作命令：

```
reg roa lnasset lev turnover liquid list if ind==5
```

回归结果如图 3.12 所示，可以发现除了 liquid（流动比率）外，其他解释变量的回归系数均在统计意义上显著。

Source	SS	df	MS	Number of obs	=	178
				F(5, 172)	=	36.52
Model	9343.79182	5	1868.75836	Prob > F	=	0.0000
Residual	8800.62471	172	51.1664228	R-squared	=	0.5150
				Adj R-squared	=	0.5009
Total	18144.4165	177	102.510828	Root MSE	=	7.1531

roa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	2.899537	.4584917	6.32	0.000	1.994542	3.804532
lev	-.4641958	.0389836	-11.91	0.000	-.5411437	-.3872479
turnover	.0453479	.0117841	3.85	0.000	.0220879	.0686079
liquid	-.5156865	.822629	-0.63	0.532	-2.139435	1.108061
list	-2.393411	1.186963	-2.02	0.045	-4.7363	-.050522
_cons	-18.16404	7.619468	-2.38	0.018	-33.20374	-3.124337

图 3.12　限定企业行业的子样本分析结果

同样地，我们也可以对资产负债率高于 60%的公司样本进行回归，输入命令：

```
reg roa lnasset lev turnover liquid list if lev>=60
```

回归结果如图 3.13 所示，可以发现解释变量 lnasset（总资产对数）、lev（资产负债率）和 turnover（总资产周转率）的回归系数均为显著。

进一步地，也可以同时采用多个条件对样本进行筛选。例如，我们希望对资产负债率不低于 60%的制造业公司样本进行回归分析，可输入命令：

```
reg roa lnasset lev turnover liquid list if lev>=60&ind==5
```

Source	SS	df	MS	Number of obs	=	1,546
				F(5, 1540)	=	108.29
Model	7284.32587	5	1456.86517	Prob > F	=	0.0000
Residual	20718.7286	1,540	13.4537199	R-squared	=	0.2601
				Adj R-squared	=	0.2577
Total	28003.0545	1,545	18.1249544	Root MSE	=	3.6679

roa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	.5913801	.081119	7.29	0.000	.4322648	.7504955
lev	-.2995054	.0137497	-21.78	0.000	-.3264756	-.2725352
turnover	.0123053	.0023948	5.14	0.000	.0076079	.0170027
liquid	-.1160158	.0772489	-1.50	0.133	-.26754	.0355083
list	.1942038	.3219181	0.60	0.546	-.4372403	.8256479
_cons	11.81917	1.587593	7.44	0.000	8.705099	14.93325

图 3.13　限定企业资产负债率的子样本分析结果

回归结果如图 3.14 所示。可以发现，在同时限定两个条件后，由于样本量仅有 91 个观测值，部分回归系数的显著性有所降低。

Source	SS	df	MS	Number of obs	=	91
				F(5, 85)	=	63.57
Model	11550.6874	5	2310.13747	Prob > F	=	0.0000
Residual	3089.02897	85	36.3415174	R-squared	=	0.7890
				Adj R-squared	=	0.7766
Total	14639.7163	90	162.663515	Root MSE	=	6.0284

roa	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	1.334359	.5796925	2.30	0.024	.1817748	2.486943
lev	-.9788463	.0651104	-15.03	0.000	-1.108303	-.8493893
turnover	.0123995	.0120081	1.03	0.305	-.0114758	.0362748
liquid	-4.021413	1.684929	-2.39	0.019	-7.371503	-.6713229
list	-4.635694	1.542234	-3.01	0.003	-7.702068	-1.569321
_cons	51.91269	12.81205	4.05	0.000	26.4389	77.38648

图 3.14　限定多个条件的子样本分析结果

5）大样本下的线性回归分析

对于样本量较大的数据，我们通常采用 OLS 估计的稳健标准误。对于 Stata 而言，只需要在回归命令的最后加入“robust”选择项，在实际操作中可简写为“r”。本例中，样本量为 3008，显然属于大样本，因此在回归时采用稳健标准误。在主界面的命令窗口中依次输入命令：

```
reg roa lnasset lev turnover liquid list,r
```

回归结果如图 3.15 所示。可以发现，在采用 OLS 估计的稳健标准误后，所得到的解释变量回归系数与图 3.8 中未使用“robust”选项时的回归系数完全相同，但其标准误变为稳健标准误（Robust Std. Err.）；与此同时，相应的 t 值和 p 值也发生了变化。在大样本情况下，通常应当使用稳健标准误，这是因为经济数据在多数情况下存在异方差。在存在异方

```
. reg roa lnasset lev turnover liquid list,r

Linear regression                               Number of obs     =      3,008
                                                F(5, 3002)        =      24.01
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1247
                                                Root MSE          =     3.1362
```

roa	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnasset	.429663	.1032813	4.16	0.000	.2271537	.6321723
lev	-.0962699	.0206519	-4.66	0.000	-.1367633	-.0557765
turnover	.0177878	.0026314	6.76	0.000	.0126283	.0229472
liquid	-.087492	.0226651	-3.86	0.000	-.1319328	-.0430513
list	-.0831874	.5196846	-0.16	0.873	-1.102161	.9357864
_cons	-.024133	.7315035	-0.03	0.974	-1.458432	1.410166

图 3.15　大样本下回归分析的结果

差的情况下，使用普通标准误会降低 OLS 估计结果的准确性。关于异方差的相关内容，将在第 4 章中进一步介绍。

3.3　本章小结与习题

3.3.1　本章小结

本章所讲解的内容是计量经济学中最核心的基础知识。在 3.1 节中，我们首先介绍了简单一元线性回归模型，以及总体回归函数和样本回归函数的基本概念；其次，介绍了普通最小二乘法（OLS）的基本思想，并证明了 OLS 估计量所具有的优良数学性质；最后，讲解了拟合优度这一用于衡量模型拟合程度的重要指标。

3.2 节介绍了更为常用的多元线性回归模型，这是本章的重点内容。首先，我们介绍了多元线性回归模型，并指出其与一元线性回归模型在形式上的相似性及在分析上的复杂性；其次，介绍了古典线性回归模型的基本假定，这是 OLS 估计量具有优良性质的前提条件；再次，讲解了普通最小二乘的估计方法及其矩阵表示形式，其中回归模型的矩阵表示和计算是本章的重点内容，也是学习中的难点之一；接下来，介绍了校正拟合优度以及 OLS 估计量的相关性质，并进一步讲解了单个回归系数的显著性检验和回归模型总体显著性的检验方法；最后，在放松部分假设的基础上，讨论了多元线性回归模型下 OLS 估计量的渐进性质。

最后，基于前述理论知识及其现实应用背景，本章通过具体案例分析，介绍了如何利用 Stata 软件进行回归分析及相关统计检验。

关键术语：总体回归函数、样本回归函数、普通最小二乘法、残差平方和（RSS）、拟合优度、总平方和（TSS）、可解释平方和（ESS）；多重共线性、同方差、无序列相关、校正拟合优度、t 检验、t 统计量、F 检验、F 统计量、置信区间、显著性水平；前定解释变量、渐近性质、渐近正态性、渐近有效性。

3.3.2　本章习题

下载资源:\sample\chap03\习题\pgdp.dta、bond.dta

1. 阐述总体回归函数与样本回归函数的概念，并辨析二者之间的联系与区别。

2. 普通最小二乘法进行参数估计的基本思想是什么？

3. 为什么需要对回归模型的拟合程度进行评价？评价标准为什么选择可决系数，而不是残差平方和？

4. 考虑以下函数：其中 $T_i=\alpha+\beta Y_i+\varepsilon_i$，其中 T_i、Y_i 分别表示某年地区 i 的税收和国内生产总值。假设 OLS 回归所得的样本回归线为 $\hat{T}_i=\hat{\alpha}+\hat{\beta}Y_i$。试问答：

（1）斜率 $\hat{\beta}$ 的经济含义是什么？

（2）截距项 $\hat{\alpha}$ 的经济含义是什么？

（3）请通过国家统计局网站查找我国 2023 年 31 个省（自治区、直辖市）的税收和国内生产总值数据，并使用 Stata 软件中的 regress 命令进行回归分析。

5. 古典线性回归模型的基本假定有什么？

6. 比较拟合优度 R^2 与校正拟合优度 $\overline{R^2}$ 的区别与联系，并进一步说明为什么多元线性回归模型中需要采用校正拟合优度 $\overline{R^2}$ 。

7. 阐述 OLS 估计量具有哪些性质？

8. 按步骤解释以下检测过程：

（1）单个回归系数显著性的 t 检验。

（2）回归方程整体显著性的 F 检验。

9. 数据集 pgdp.dta 包含我国 2020 年 31 个省（市、自治区）的如下变量：lnpgdp（人均 GDP 的自然对数值）、gdpr（GDP 增速）、ind3（第三产业 GDP 占比）、finance（金融业 GDP 占比）、lntrade（进出口贸易额的自然对数值）。请使用 Stata 软件完成以下操作：

（1）以 lnpgdp 为被解释变量，对其他变量进行 OLS 回归。

（2）根据回归结果，判断哪些解释变量是显著的，并说明判断依据。

（3）根据回归结果，说明拟合优度 R^2 与校正拟合优度 $\overline{R^2}$ 分别是多少？并结合回归结果写出这两个数值的计算过程。

（4）根据回归方程是否整体显著，并说明判断依据。

（5）给出 lntrade 的回归系数，并解释其经济含义。

10. 债券发行已成为企业的融资方式之一，那么债券的融资成本受到哪些因素的影响？数据集 bond.dta 包含我国债券市场 2782 只债券的相关信息：spread（债券利差，代表了债券的融资成本）、size（债券发行规模）、rating（债券评级，评级越高表示债券风险越大）、rate（债券发行的企业评级，评级越高表示企业信誉越高）、maturity（债券期限）、coupon（票面利率）、province（债券所属省份）、indus（债券所属行业）。请使用 Stata 软件执行以下操作：

（1）使用全样本，以 spread 为被解释变量，对变量 size、rating、rate、maturity、coupon 进行 OLS 回归分析。

（2）使用全样本，以 spread 为被解释变量，对变量 size、rating、rate、maturity、coupon 使用稳健标准误进行回归分析，并与普通标准误的回归结果进行对比。

（3）判断稳健标准误回归中各解释变量及回归方程的显著性，并说明判断标准。

（4）使用债券所属省份为江苏省的子样本，对模型进行估计。

（5）使用债券所属行业为建筑业的子样本，对模型进行估计。

（6）比较子样本与全样本的估计结果，并分析产生差异的原因。