

基于 OLS 一元线性回归方程显著性检验之数学推导



王伟*

广西财经学院中国-东盟统计学院, 广西南宁 530003

摘要: 一元线性回归在统计学和计量经济学的教学内容体系中具有独特的重要地位, 它是理解和掌握多元线性回归的重要基础。然而, 作为一元线性回归的重点和难点, 关于 OLS 估计量性质与回归方程显著性检验往往缺乏系统的数学推导。为了对回归方程显著性检验进行系统的数学推导, 论文首先梳理了高斯的五个基本假定和三大分布 (卡方分布、t 分布和 F 分布) 的定义与数学特征; 接着论证了回归方程 OLS 估计量满足线性特性、无偏性和有效性的特点, 是最优线性无偏估计量 (BLUE); 然后阐述了关于回归直线拟合优度判断的判定系数法和估计标准误差法, 两种方法具有内在的一致性, 估计标准误差越小, 判定系数随之越大, 意味着回归直线的拟合程度越好; 最后通过证明 SSR/σ^2 和 SSE/σ^2 分别服从自由度为 1 和 $(n-2)$ 的卡方分布, 由此证明了 t 统计量和 F 统计量分别服从 t 分布和 F 分布, 而且 F 检验与 t 检验、相关系数 r 显著性检验具有一致性。因此, 在一元线性回归方程的统计检验中, 不需要对整体线性关系、回归系数和相关系数同时进行统计显著性检验, 只需要对三者之一进行统计检验即可。

关键词: 一元线性回归; OLS 估计量; 三大分布; 显著性检验; 拟合优度

DOI: [10.57237/j.wjms.2025.01.002](https://doi.org/10.57237/j.wjms.2025.01.002)

Mathematical Derivation of Significance Test of Unitary Linear Regression Equation Based on OLS

Wang Wei*

China-Asean Institute of Statistics, Guangxi University of Finance and Economics, Nanning 530003, China

Abstract: Unitary linear regression plays a unique and important position in the teaching content system of statistics and econometrics. It is an important basis for understanding and mastering multiple linear regression. However, as the key and difficult point of unitary linear regression, the properties of OLS estimators and the significance test of regression equations often lack of systematic mathematical derivation. In order to derive systematically the significance test of regression equation, the paper first combs the five basic assumptions of Gauss and the definition and mathematical characteristics of three major distributions (Chi-square distribution, t-distribution and F-distribution). Then, it is

基金项目: 2022 年度全国统计科学研究项目 (2022LY065);

广西财经学院博士启动项目 (BS2021012);

2024 年度广西高等教育本科教学改革工程项目 (2024JGA324);

广西财经学院校级本科教学改革工程项目 (2024XJJGYB16、2024XJJGYB22).

*通信作者: 王伟, 2020110014@gxufe.edu.cn

收稿日期: 2025-01-19; 接受日期: 2025-05-20; 在线出版日期: 2025-06-30

<http://www.wjoms.com>

demonstrated that the OLS estimator of the regression equation satisfies the characteristics of linearity, unbiased and effectiveness, and it is the optimal linear unbiased estimator (BLUE). Then, expounds the determination coefficient method and the estimation standard error method about the judgement on goodness of fit for regression line. The two methods have intrinsic consistency. The smaller the estimation standard error is, the larger the determination coefficient is, which means the better the fitting degree of regression line. Finally, by proving that SSR/σ^2 and SSE/σ^2 obey the Chi-square distribution of freedom of 1 and (n-2) respectively, it is proved that t statistic and F statistic obey the t distribution and F distribution respectively, and the F test is consistent with the t test and the significance test of correlation coefficient r. Therefore, in the statistical testing of unitary linear regression equations, it is not necessary to test the statistical significance of the correlation coefficient, regression coefficient, and overall regression equation simultaneously. Only one of the three needs to be tested.

Keywords: Unitary Linear Regression; OLS Estimator; Three Major Distributions; Test of Significance; Goodness of Fit

1 引言

一元线性回归的参数估计方法和显著性检验原理是多元线性回归等相关内容的重要基础。大多数教材对于一元线性回归方程的参数估计和显著性检验往往直接给出统计量的抽样分布, 未提供系统的数学证明, 使学生不得其解。本文以一元线性回归的基本假定与三大分布的定义及数学特征作为研究起点, 依次展开关于一元线性回归方程参数估计、回归直线拟合优度和回归方程显著性检验的数学证明, 最后证明在一元线性回归中 F 检验、t 检验与样本相关系数 r 显著性检验的内在一致性。

2 一元线性回归的基本假定与三大分布的定义及数学特征

2.1 基本假定

假设一元线性回归模型为: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, 其中, β_0 和 β_1 为模型未知参数, 分别称其为回归常数和回归系数; ε_i 为随机误差项或随机扰动项。为了估计模型未知参数 β_0 和 β_1 , 往往随机抽取 n 个样本观测点 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 。为了保证关于线性回归模型的参数估计量具有良好的统计性质, 高斯对随机误差项 ε_i 提出以下假定 (称为高斯假定^①) [1]。

^①严格来说, 高斯假定并不是针对一元回归模型本身, 而是针对普通最小二乘法 OLS。如果实际模型满足高斯假定, OLS 则

(1) 零均值假定, 即:

$$E(\varepsilon_i | x_i) = 0 \Rightarrow E(\varepsilon_i) = E[E(\varepsilon_i | x_i)] = 0$$

(2) 同方差假定, 即:

$$\begin{aligned} \text{var}(\varepsilon_i | x_i) &= E\left\{\left[\varepsilon_i - E(\varepsilon_i | x_i)\right]^2 | x_i\right\} = E(\varepsilon_i^2 | x_i) = \sigma^2, \\ \text{var}(\varepsilon_i) &= E(\varepsilon_i^2) = E[E(\varepsilon_i^2 | x_i)] = \sigma^2 \end{aligned}$$

(3) 正态分布假定, 即 $\varepsilon_i \sim N(0, \sigma^2)$

(4) 无自相关假定, 即:

$$\text{cov}(\varepsilon_i, \varepsilon_j) = E\left[\varepsilon_i - E(\varepsilon_i)\right]\left[\varepsilon_j - E(\varepsilon_j)\right] = E(\varepsilon_i \varepsilon_j) = 0$$

由于在二维正态分布下, “ ε_i 与 ε_j 不相关”与“ ε_i 与 ε_j 相互独立”等价。因此, 综合以上四个假定可知, 随机误差项 ε_i 相互独立且均服从零均值、同方差 σ^2 的正态分布, 即:

$$\varepsilon_i : \text{iid}N(0, \sigma^2)$$

(5) 随机误差项与解释变量不相关假定, 即:

$$\begin{aligned} \text{cov}(\varepsilon_i, x_i) &= E\left\{\left[\varepsilon_i - E(\varepsilon_i)\right]\left[x_i - E(x_i)\right]\right\} \\ &= E(\varepsilon_i x_i) - E(\varepsilon_i)E(x_i) \\ &= E\left[E(\varepsilon_i x_i | x_i)\right] - 0 \times E(x_i) \\ &= E\left[x_i E(\varepsilon_i | x_i)\right] = 0 \end{aligned}$$

是好的估计方法; 如果不满足高斯假定, OLS 则不再适用。

由于 $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, 根据高斯假定可知, y_i 相互独立且均服从均值为 $\beta_0 + \beta_1 x_i$ 、方差为 σ^2 的正态分布, 即 $y_i: \text{iid}N(\beta_0 + \beta_1 x_i, \sigma^2)$ 。

2.2 三大分布的定义与数学特征

2.2.1 卡方分布的定义与数学特征

假定随机变量 $X_1, X_2, \dots, X_n$ 相互独立, 且 $X_i: N(0,1)$, 则 $Y = \sum_{i=1}^n X_i^2: \chi^2(n)$, 而且:

$$E[Y] = \sum_{i=1}^n E(X_i^2) = n, D[Y] = \sum_{i=1}^n D(X_i^2) = 2n$$

$$E[1/Y] = 1/(n-2), E[1/Y^2] = 1/[(n-2)(n-4)]$$

而且, 卡方分布满足可加性 [2], 如果 $Y_1: \chi^2(n_1), Y_2: \chi^2(n_2)$, 则 $(Y_1 + Y_2): \chi^2(n_1 + n_2)$ 。

2.2.2 t 分布的定义与数学特征

假定随机变量 $X: N(0,1), Y: \chi^2(n)$, 且 X 与 Y 相互独立, 则 $t = \frac{X}{\sqrt{Y/n}}: t(n)$, 而且:

$$E(t) = \sqrt{n}E(X)E\left(\frac{1}{\sqrt{Y}}\right) = 0,$$

$$D(t) = E(t^2) = E\left(\frac{X^2}{Y/n}\right) = nE(X^2)E\left(\frac{1}{Y}\right) = \frac{n}{n-2} (n > 2)$$

2.2.3 F 分布的定义与数学特征

假定随机变量 $Y_1: \chi^2(n_1), Y_2: \chi^2(n_2)$, 且 Y_1 和 Y_2 相互独立, 则 $F = \frac{Y_1/n_1}{Y_2/n_2}: F(n_1, n_2)$, 而且:

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{l_{xy}}{l_{xx}}, \beta_0 = \bar{y} - \beta_1 \bar{x}$$

由此可知, 在直角坐标系中, 根据估计的回归方程拟合的回归直线必然经过 $(0, \beta_0)$ 和 $(\bar{x}, \bar{y})$ 两点。而且, $\bar{\hat{y}} = \bar{y}$ 。

3.2 OLS 估计量的性质

OLS 估计量是否是一个好的估计量呢? 下面从估计量的分布出发, 依次证明 OLS 估计量的线性特性、无偏性和有效性。

$$E(F) = \frac{n_2}{n_1} E(Y_1) E\left(\frac{1}{Y_2}\right) = \frac{n_2}{n_1} g_1 g_2 \frac{1}{(n_2 - 2)} = \frac{n_2}{n_2 - 2} (n_2 > 2)$$

$$D(F) = E(F^2) - [E(F)]^2 = \frac{2n_2(n_1 + n_2 - 2)}{n_1(n_2 - 2)^2(n_2 - 4)} (n_2 > 4)$$

3 一元线性回归 OLS 估计量的求解及其性质

3.1 OLS 估计量的求解

为了估计一元线性回归模型的参数 β_0 和 β_1 , 采用不同的估计方法会得到不同的样本统计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$, 进而得到不同的回归估计值 $\hat{y}_i = \beta_0 + \beta_1 x_i$ 。但是, 理想的估计方法应使得所有的 $\hat{y}_i$ 与 y_i 的距离越小越好 [3]。通常, 采用普通最小二乘法 OLS 进行估计, 使得 $Q = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ 最小化 [4]。根据微积分的极值定理, 可得:

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = -2 \sum_{i=1}^n e_i = 0,$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = -2 \sum_{i=1}^n x_i e_i = 0$$

经过整理, 可得到以下正规方程组:

$$\sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i, \sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2$$

通过消元法求解, 可得 [5]:

3.2.1 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 估计量的分布

$$\begin{aligned}\beta_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} y_i = \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} y_i \\ \therefore \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} &= 0, \frac{\sum_{i=1}^n (x_i - \bar{x}) x_i}{l_{xx}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x})}{l_{xx}} = 1\end{aligned}$$

β_1 的期望和方差分别如下：

$$\begin{aligned}\therefore E(\beta_1) &= \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} E(y_i) = \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} (\beta_0 + \beta_1 x_i) = \beta_1 \\ D(\beta_1) &= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} D(y_i) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sigma^2 / l_{xx}\end{aligned}$$

由于 $y_i : iidN(\beta_0 + \beta_1 x_i, \sigma^2)$ ，因此 β_1 是 n 个独立正态随机变量 y_i 的线性组合，同样服从正态分布。因此， $\beta_1 \sim N(\beta_1, \sigma^2 / l_{xx})$ 。

同样，可求出 β_0 的期望和方差如下：

$$\begin{aligned}\therefore \beta_0 &= \bar{y} - \beta_1 \bar{x} = \frac{\sum_{i=1}^n y_i}{n} - \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} y_i \bar{x} = \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right] y_i \\ \therefore E(\beta_0) &= E(\bar{y}) - E(\beta_1 \bar{x}) = E\left(\sum_{i=1}^n y_i / n\right) - \bar{x} E(\beta_1) = (\beta_0 + \beta_1 \bar{x}) - \beta_1 \bar{x} = \beta_0 \\ D(\beta_0) &= \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x}) \bar{x}}{l_{xx}} \right]^2 D(y_i) = \left\{ \sum_{i=1}^n \frac{1}{n^2} - \frac{2 \sum_{i=1}^n (x_i - \bar{x}) \bar{x}}{n l_{xx}} + \sum_{i=1}^n \left[\frac{\bar{x} (x_i - \bar{x})}{l_{xx}} \right]^2 \right\} \sigma^2 = \left[\frac{1}{n} - 0 + \frac{\bar{x}^2}{l_{xx}} \right] \sigma^2 = \frac{\sum_{i=1}^n x_i^2}{n l_{xx}} \sigma^2\end{aligned}$$

因此， $\beta_0 \sim N(\beta_0, \frac{\sum_{i=1}^n x_i^2}{n l_{xx}} \sigma^2)$ 。

3.2.2 估计量的线性特性

$$Q \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})}{l_{xx}} y_i = \sum_{i=1}^n k_{1i} y_i$$

其中， $k_{1i} = \frac{x_i - \bar{x}}{l_{xx}}$ ， $\sum_{i=1}^n k_{1i} = 0$ ， $\sum_{i=1}^n k_{1i} x_i = 1$

$$\hat{\beta}_0 = \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} \right] y_i = \sum_{i=1}^n k_{0i} y_i$$

$$\text{其中, } k_{0i} = \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} \right], \sum_{i=1}^n k_{0i} = 1, \sum_{i=1}^n k_{0i} x_i = 0$$

因此, OLS 估计量均是正态随机变量 y_i 的线性组合, 也服从正态分布。

但是, 由于估计量的无偏性, 保证了估计值对参数真值的平均偏离为零。

3.2.3 估计量的无偏性

根据估计量的分布可知, $E(\beta_1) = \beta_1, E(\beta_0) = \beta_0$ 。样本的随机性导致了估计值偏离参数真值的随机性。

3.2.4 估计量的有效性

现假设 β_1^* 和 β_0^* 分别是模型参数 β_1 和 β_0 的任意一个线性无偏估计量, 即:

$$\begin{cases} \beta_1^* = \sum_{i=1}^n w_{1i} y_i (w_{1i} \neq k_{1i}) \Rightarrow E(\beta_1^*) = \sum_{i=1}^n w_{1i} E(y_i) = \sum_{i=1}^n w_{1i} (\beta_0 + \beta_1 x_i) \\ \beta_0^* = \sum_{i=1}^n w_{0i} y_i (w_{0i} \neq k_{0i}) \Rightarrow E(\beta_0^*) = \sum_{i=1}^n w_{0i} E(y_i) = \sum_{i=1}^n w_{0i} (\beta_0 + \beta_1 x_i) \end{cases}$$

$$\begin{cases} \sum_{i=1}^n w_{1i} \beta_0 = 0, \sum_{i=1}^n w_{1i} \beta_1 x_i = \beta_1 \Rightarrow \sum_{i=1}^n w_{1i} = 0, \sum_{i=1}^n w_{1i} x_i = 1 \\ \sum_{i=1}^n w_{0i} \beta_0 = \beta_0, \sum_{i=1}^n w_{0i} \beta_1 x_i = 0 \Rightarrow \sum_{i=1}^n w_{0i} = 1, \sum_{i=1}^n w_{0i} x_i = 0 \end{cases}$$

则 β_1^* 的方差为:

$$\begin{aligned} D(\beta_1^*) &= \sum_{i=1}^n w_{1i}^2 D(y_i) = \sigma^2 \sum_{i=1}^n (w_{1i} - k_{1i} + k_{1i})^2 = \sigma^2 \sum_{i=1}^n [(w_{1i} - k_{1i})^2 + k_{1i}^2 + 2(w_{1i} - k_{1i})k_{1i}] \\ &\because \sum_{i=1}^n (w_{1i} - k_{1i})k_{1i} = \sum_{i=1}^n w_{1i} k_{1i} - \sum_{i=1}^n k_{1i}^2 = \sum_{i=1}^n w_{1i} \frac{(x_i - \bar{x})}{l_{xx}} - \sum_{i=1}^n \left[\frac{(x_i - \bar{x})}{l_{xx}} \right]^2 = \frac{1}{l_{xx}} - \frac{1}{l_{xx}} = 0 \\ &\therefore D(\beta_1^*) = \sigma^2 \sum_{i=1}^n (w_{1i} - k_{1i})^2 + \sigma^2 \sum_{i=1}^n k_{1i}^2 = \sigma^2 \sum_{i=1}^n (w_{1i} - k_{1i})^2 + \sigma^2 / l_{xx} \geq D(\beta_1) = \sigma^2 / l_{xx} \end{aligned}$$

同理, β_0^* 的方差为:

$$\begin{aligned} D(\beta_0^*) &= \sum_{i=1}^n (w_{0i} - k_{0i} + k_{0i})^2 D(y_i) = \sigma^2 \sum_{i=1}^n [(w_{0i} - k_{0i})^2 + k_{0i}^2 + 2(w_{0i} - k_{0i})k_{0i}] \\ &\because \sum_{i=1}^n k_{0i}^2 = \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} \right]^2 = \left(\frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} \right) = \frac{\sum x_i^2}{nl_{xx}}, \sum_{i=1}^n w_{0i} k_{0i} = \sum_{i=1}^n w_{0i} \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} \right] = \frac{1}{n} + \frac{\bar{x}^2}{l_{xx}} = \frac{\sum x_i^2}{nl_{xx}} \\ &\therefore D(\beta_0^*) = \sigma^2 \sum_{i=1}^n (w_{0i} - k_{0i})^2 + \frac{\sum x_i^2}{nl_{xx}} \sigma^2 \geq D(\beta_0) = \frac{\sum x_i^2}{nl_{xx}} \sigma^2 \end{aligned}$$

可知,在所有满足线性无偏性质的估计量中, β_1 和 β_0 均具有最小的方差性质,即满足有效性[6]。相比于其他的线性无偏估计量,有效性保证了 OLS 估计量具有更大的概率离参数真值更近。因此,OLS 估计量是最优线性无偏估计量 (BLUE) [7]。

y 之间的线性关系。但是回归方程所拟合的回归直线并不一定经过所有的样本观测点 (x_i, y_i) 。各样本观测点离回归直线越近,回归直线的拟合程度越好。为了反映回归直线的拟合优度,可以使用判定系数 R^2 和估计的标准误差 S_e 来衡量。

4 回归直线拟合优度的判断

4.1 判定系数 R^2

根据 OLS 估计所得到的回归方程描述了变量 x 与

$$\text{令 } SST = \sum_{i=1}^n (y_i - \bar{y})^2, SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2, SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$\because \sum_{i=1}^n (y_i - \hat{y}_i)(y_i - \bar{y}) = \sum_{i=1}^n e_i (y_i - \bar{y}) = \sum_{i=1}^n e_i [\beta_0 + \beta_1 x_i - (\beta_0 + \beta_1 \bar{x})] = 0$$

$$\therefore SST = SSR + SSE$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = r^2$$

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}} = \beta_1 \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

因此,相关系数 r 与 β_1 的正负号一致。而且,相关系数 r 与判定系数 R^2 一样也可以反映回归直线的拟合优度。相关系数 r 的绝对值越大,判定系数 R^2 越大,拟合优度越好。

4.2 估计标准误差 S_e

估计标准误差是一个度量回归估计值与实际值之间偏离程度的统计量,它反映了用回归估计值 $\hat{y}_i$ 预测 y_i 时误差的大小[8]。残差平方和越小,回归直线的拟合程度越好。

$$\begin{aligned} SSE &= \sum_{i=1}^n \left[(\beta_0 - \beta_0) + (\beta_1 - \beta_1)x_i + \varepsilon_i \right]^2 \\ &= \sum_{i=1}^n \left[(\beta_0 - \beta_0)^2 + (\beta_1 - \beta_1)^2 x_i^2 + \varepsilon_i^2 + 2(\beta_0 - \beta_0)(\beta_1 - \beta_1)x_i \right] + \sum_{i=1}^n \left[2(\beta_0 - \beta_0)\varepsilon_i + 2(\beta_1 - \beta_1)x_i \varepsilon_i \right] \end{aligned}$$

$$\begin{aligned} E(SSE) &= \sum_{i=1}^n D(\beta_0) + \sum_{i=1}^n D(\beta_1)x_i^2 + \sum_{i=1}^n D(\varepsilon_i) + 2 \sum_{i=1}^n \text{cov}(\beta_0, \beta_1)x_i - 2 \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} + \frac{(x_i - \bar{x})x_i}{l_{xx}} \right] E(y_i \varepsilon_i) \\ &= \left[\frac{\sum_{i=1}^n x_i^2}{l_{xx}} + \frac{\sum_{i=1}^n x_i^2}{l_{xx}} + n - \frac{2n\bar{x}^2}{l_{xx}} \right] \sigma^2 - 2 \sum_{i=1}^n \left[\frac{1}{n} - \frac{(x_i - \bar{x})\bar{x}}{l_{xx}} + \frac{(x_i - \bar{x})x_i}{l_{xx}} \right] \sigma^2 \\ &= (n+2)\sigma^2 - 4\sigma^2 = (n-2)\sigma^2 \end{aligned}$$

由此, SSE 的大小与其自由度 $(n-2)$ 相关。因此,

$SSE / (n-2)$ [9], 其数学期望为:

为了剔除自由度的影响,将 SSE 除以其自由度得到

$$E(SSE / n - 2) = E(MSE) = \sigma^2$$

因此, $SSE/(n-2)$ 和均方残差 MSE 是误差方差 σ^2 的无偏估计量。在实际回归分析中, 由于标准差 σ 的未知性, 常用标准误差 se 来代替误差标准差 σ 的估计值 σ [10], 即 $\sigma = se = \sqrt{SSE/(n-2)} = \sqrt{MSE}$ 。

在样本量 n 一定的情况下, 衡量回归直线拟合优度的判定系数 R^2 和估计的标准误差 Se 具有内在的一致性。当估计的标准误差 Se 越小, 判定系数 R^2 随之越大, 意味着回归直线的拟合程度越好。

5 回归方程的显著性检验

5.1 SSE/σ^2 与 SSR/σ^2 服从卡方分布之证明

首先, 对 $y_1, y_2, \dots, y_n$ 作正交变换 $Z = AY$, 其中, A 为正交矩阵, 具有如下形式[11]:

$$A = \begin{bmatrix} 1/\sqrt{n} & 1/\sqrt{n} & L & 1/\sqrt{n} \\ (x_1 - \bar{x})/\sqrt{l_{xx}} & (x_2 - \bar{x})/\sqrt{l_{xx}} & L & (x_n - \bar{x})/\sqrt{l_{xx}} \\ a_{31} & a_{32} & L & a_{3n} \\ M & M & M & M \\ a_{n1} & a_{n2} & L & a_{nn} \end{bmatrix}$$

由 A 的正交性, 满足以下条件:

$$\sum_{j=1}^n a_{ij}^2 = 1, \sum_{j=1}^n a_{ij} = 0, \sum_{j=1}^n a_{ij} x_j = 0 (i \geq 3); \sum_{j=1}^n a_{ij} a_{kj} = 0 (k > i \geq 3)$$

$$Z = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ \vdots \\ z_n \end{bmatrix} = A \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sqrt{n}y \\ l_{xy}/\sqrt{l_{xx}} \\ \sum_{j=1}^n a_{3j}y_j \\ \vdots \\ \sum_{j=1}^n a_{nj}y_j \end{bmatrix} = \begin{bmatrix} \sqrt{n}(\beta_0 + \beta_1\bar{x}) \\ \beta_1\sqrt{l_{xx}} \\ \sum_{j=1}^n a_{3j}(\beta_0 + \beta_1x_j + \varepsilon_j) \\ \vdots \\ \sum_{j=1}^n a_{nj}(\beta_0 + \beta_1x_j + \varepsilon_j) \end{bmatrix}$$

由于 $y_i: iidN(\beta_0 + \beta_1x_i, \sigma^2)$, 因而, $z_1, z_2, z_3, \dots, z_n$ 相互独立且满足: $z_1: N(\sqrt{n}(\beta_0 + \beta_1\bar{x}), \sigma^2)$, $z_2 \sim N(\beta_1\sqrt{l_{xx}}, \sigma^2)$, $z_i: N(0, \sigma^2) (i \geq 3)$ [12]。

根据卡方分布的定义, 可知:

$$\therefore SSR = \sum_{i=1}^n (y_i - \bar{y})^2 = \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \beta_1^2 l_{xx} = z_2^2, \text{ if } \beta_1 = 0$$

$$\therefore SSR / \sigma^2 = z_2^2 / \sigma^2 = \left[(z_2 - \beta_1\sqrt{l_{xx}}) / \sigma \right]^2 \sim \chi^2(1)$$

$$\therefore SSE = SST - SSR = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 - \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 - \frac{l_{xy}^2}{l_{xx}} = \sum_{i=1}^n z_i^2 - z_1^2 - z_2^2 = \sum_{i=3}^n z_i^2$$

$$\therefore SSE / \sigma^2 = \sum_{i=3}^n z_i^2 / \sigma^2 = \sum_{i=3}^n \left(\frac{z_i - 0}{\sigma} \right)^2 \sim \chi^2(n-2)。$$

根据卡方分布的数学特征, 可知:

$$E(SSE / \sigma^2) = n - 2$$

5.2 t 统计量服从 t 分布之证明

根据 t 分布的定义，可知：

$$\begin{aligned} \therefore \frac{\beta_1 - \beta_1}{\sigma / \sqrt{l_{xx}}} &\sim N(0,1) \\ SSE / \sigma^2 = \sum_{i=3}^n z_i^2 / \sigma^2 = \sum_{i=3}^n \left(\frac{z_i - 0}{\sigma} \right)^2 &\sim \chi^2(n-2) \\ \therefore t = \frac{\frac{\beta_1 - \beta_1}{\sigma / \sqrt{l_{xx}}}}{\sqrt{(SSE / \sigma^2) / (n-2)}} &= \frac{(\beta_1 - \beta_1) \sqrt{l_{xx}}}{\sqrt{SSE / (n-2)}} = \frac{(\beta_1 - \beta_1)}{se / \sqrt{l_{xx}}} = \frac{(\beta_1 - \beta_1)}{se(\beta_1)} \sim t(n-2) \end{aligned}$$

5.3 F 统计量服从 F 分布之证明

$$\begin{aligned} \therefore \text{if } \beta_1 = 0, SSR / \sigma^2 &\sim \chi^2(1), SSE / \sigma^2 \sim \chi^2(n-2) \\ \therefore F = \frac{(SSR / \sigma^2) / 1}{(SSE / \sigma^2) / (n-2)} &\sim F(1, n-2) \end{aligned}$$

5.4 F 检验与 t 检验、相关系数 r 显著性检验的一致性

在一元线性回归中，根据样本观测点经过 OLS 所估计的回归方程是否真实反映了总体变量之间的关系，需要进行回归方程的显著性检验，包括回归方程整体线性关系的 F 检验和回归系数的 t 检验。

$$\begin{aligned} \therefore F = \frac{(SSR / \sigma^2) / 1}{(SSE / \sigma^2) / (n-2)} &= \frac{\beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2}{SSE / (n-2)} = \frac{\beta_1^2 l_{xx}}{SSE / (n-2)}, t = \frac{\beta_1 \sqrt{l_{xx}}}{\sqrt{SSE / (n-2)}} \\ \therefore F &= t^2 \\ \therefore R^2 = SSR / SST &= SSR / (SSR + SSE), R^2 = r^2 \\ \therefore F = \frac{SSR}{SSE / (n-2)} &= \frac{(n-2) SSR}{(1-r^2) SSR / r^2} = \frac{r^2 (n-2)}{(1-r^2)} \end{aligned}$$

因此， $F = t^2 = \frac{r^2 (n-2)}{(1-r^2)}$ [13, 14]。对于一元线性

回归方程，因只有一个自变量 x，回归方程的显著性检验与回归系数的显著性检验具有一致性。同时，为了考察样本相关系数 r 的可靠性，同样需要进行样本相关系数 r 的显著性检验。由于相关系数 r 显著性的检验统

计量为 $t = r \sqrt{\frac{(n-2)}{1-r^2}}$ 同样服从自由度为 (n-2) 的 t 分布。

而且，回归系数检验统计量的值等于相关系数检验统

计量的值，即：

$$t = \frac{\beta_1}{se(\beta_1)} = \frac{\beta_1}{se / \sqrt{l_{xx}}} = \frac{r \sqrt{l_{yy} / l_{xx}}}{se / \sqrt{l_{xx}}} = \frac{r \sqrt{l_{yy}}}{se} = r \sqrt{\frac{n-2}{1-r^2}}$$

因此，在一元线性回归中，F 检验与 t 检验、相关系数 r 显著性检验具有一致性。

6 结论

在满足高斯基本假定的前提下,证明了 OLS 估计量满足线性特性、无偏性和有效性的特点,是最优线性无偏估计量 (BLUE)。阐述了回归直线拟合优度判断的判定系数法和估计标准误差法,两者具有内在的一致性,估计标准误差越小,判定系数随之越大,意味着回归直线的拟合程度越好。最后,通过证明 SSR/σ^2 和 SSE/σ^2 分别服从自由度为 1 和 $(n-2)$ 的卡方分布,由此证明了 t 统计量和 F 统计量分别服从 t 分布和 F 分布,而且一元线性回归中的 F 检验与 t 检验、相关系数 r 显著性检验具有一致性。因此,在一元线性回归的统计检验中,不需要同时对整体线性关系、回归系数和相关系数同时进行统计显著性检验,只需要对其中三者之一进行检验即可[15]。

参考文献

- [1] 李子奈, 潘文卿. 计量经济学 [M], 高等教育出版社, 2005 年 4 月: 30-43.
- [2] 贾俊平, 何晓群, 金勇进. 统计学 [M], 中国人民大学出版社, 2021 年 10 月: 100-103.
- [3] 唐薇, 冯长焕. 一元线性回归模型的参数估计法的误差分析 [J], 宜宾学院学报, 2014 年 12 期: 18-21.
<https://doi.org/10.3969/j.issn.1671-5365.2014.12.005>
- [4] 张晓峒. 应用数量经济学 [M], 机械工业出版社, 2009 年 3 月: 88-94.
- [5] 李秀兰. 一元线性回归参数的估计 [J], 山西大同大学学报 (自然科学版), 2020 年第 6 期: 24-26.
<https://doi.org/10.3969/j.issn.1674-0874.2020.06.008>
- [6] 肖翔, 许伯生, 张颖. 一元线性回归中最小二乘估计量的有效性 [J], 上海工程大学学报, 2011 年第 3 期: 266-268.
<https://doi.org/10.3969/j.issn.1009-444X.2011.03.018>

- [7] 赵国庆. 应用计量经济学 (第二版) [M], 中国人民大学出版社, 2017 年 1 月: 26.
- [8] 贾俊平, 何晓群, 金勇进. 统计学 [M], 中国人民大学出版社, 2021 年 10 月: 228-229.
- [9] 程海奎, 章建跃. 通过成对数据的统计分析发展学生的数据分析素养 [J], 数学通报, 2022 年第 3 期: 7-17.
<https://doi.org/10.3969/j.issn.0583-1458.2022.03.002>
- [10] 陈武, 向杭, 陈尘. 一元线性回归方程估计标准误差的新认识 [J], 统计与决策, 2017 年 12 期: 2+189.
<https://doi.org/10.13513/j.cnki.tjyjc.2017.12.003>
- [11] 周纪芄. 回归分析 [M], 1993 年 3 月: 10-12.
- [12] 王蕊, 陈桂景. 一元线性回归检验理论的证明 [J], 大学数学, 2009 年第 6 期: 149-151.
<https://doi.org/10.3969/j.issn.1672-1454.2009.06.031>
- [13] 唐亚宁, 周丙常, 谢文贤. 一元线性回归方程的三种显著性检验的等价性分析 [J], 高等数学研究, 2022 年第 6 期: 76-78. <https://doi.org/10.3969/j.issn.1008-1399.2022.06.027>
- [14] 陈立强. 一元线性回归四种检验等效性证明 [J], 柳州职业技术学院学报, 2019 年第 4 期: 92-94.
<https://doi.org/10.16221/j.cnki.issn1671-1084.2019.04.020>
- [15] 王天营. 一元线性回归分析中三种检验的等价性研究 [J], 统计与决策, 2011 年第 3 期: 8-11.
<https://doi.org/10.13513/j.cnki.tjyjc.2011.03.008>

作者简介

王伟

1980 年生, 男, 江西吉安人, 博士后, 副教授. 研究方向是金融统计和风险管理的教学与研究.

E-mail: 2020110014@gxufe.edu.cn