数据科学导论第 6 讲—— 回归方法

王小宁

中国传媒大学数据科学与智能媒体学院

2025 年 04 月 14 日

中国传媒大学
COMMUNICATION UNIVERSITY OF CHINA

目录

问题的提出

例子

- 为了研究某社区家庭月消费支出与家庭月可支配收入之间的关系，随机抽取并调查了 12 户家庭的相关数据，见下表。

Income	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
Consume	594	638	1122	1155	1408	1595	1969	2078	2585	2530

- 注：数据来自李子奈、潘文卿《计量经济学》（第三版）
- 通过调查所得的样本数据能否发现家庭消费支出与家庭可支配收入之间的数量关系，以及如果知道了家庭的月可支配收入，能否预测家庭的月消费支出水平呢?

消费与收入的散点图

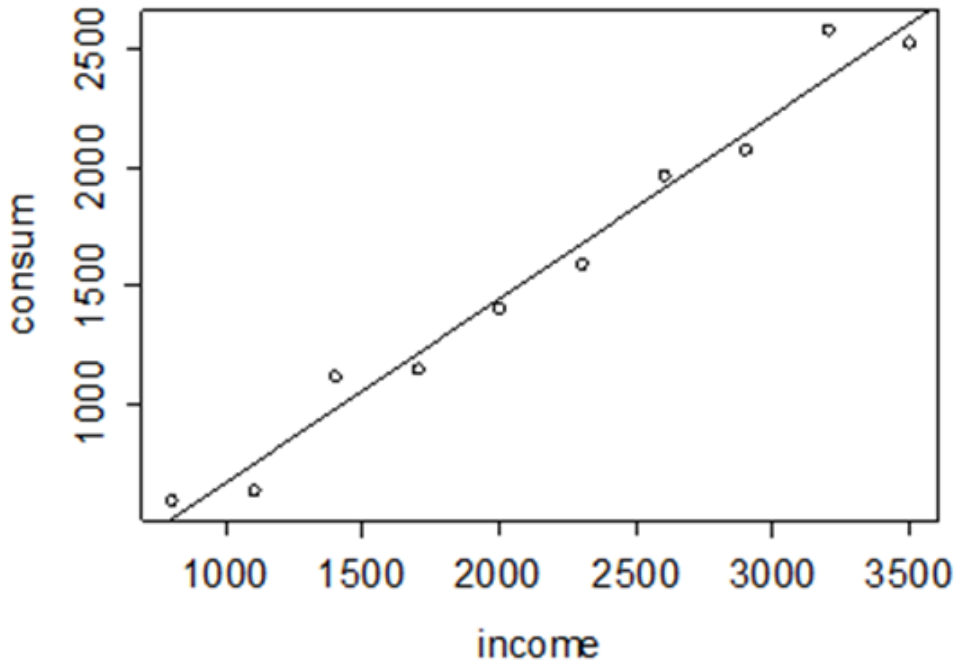


图 1: Scatter Plot

- 我们首先对数据进行探索性分析，发现消费与收入具有很强的正相关关系，pearson 相关系数为 0.988。通过图 1 的散点图可以看出，两者有着明显的线性关系，但是还无法确定收入具体是如何影响消费支出的呢?

最大心率研究

- 医学上认为一个人的最大心率和年龄是有很大关系的，一般有这样的经验公式 MaxRate = 220 - Age 来决定的。现在收集了 15 个来自不同年龄层的人接受了最大心率测试的数据，如下表所示。

表2 最大心率与年龄的调查数据

Age(x)	Max Rate(y)	Age(x)	Max Rate(y)	Age(x)	Max Rate(y)
18	202	54	169	23	193
23	186	34	174	42	174
25	187	56	172	18	198
35	180	72	153	39	183
65	156	19	199	37	178

图 2: 调查数据

探索分析

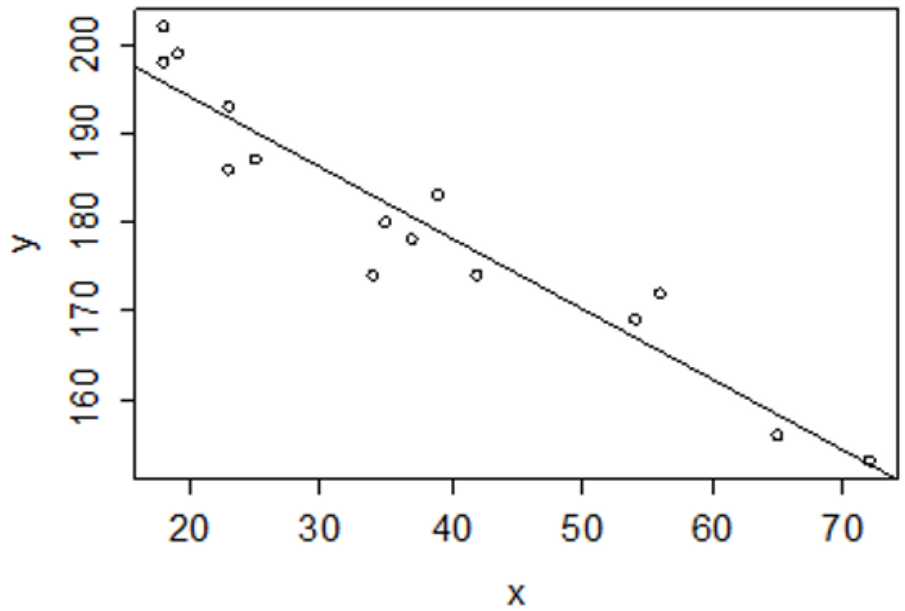


图2 最大心率与年龄散点图

图 3: 调查数据

- 通过探索性分析，我们发现最大心率与年龄具有很强的负相关关系，pearson 相关系数为-0.953。通过图 2 的散点图可以看出，两者有着明显的线性的关系，但同样也无法确定年龄具体是如何影响最大心率的呢?

一元线性回归

例子

- 在一个假想的由 100 户家庭组成的社区中，我们想要研究该社区每月家庭消费支出与每月家庭可支配收入的关系（参见下表），例如随着家庭月收入的增加，其平均月消费支出是如何变化的?

收入支出数据

表3 某社区家庭月可支配收入和消费支出

	每月家庭可支配收入X（元）									
	800	1100	1400	1700	2000	2300	2600	2900	3200	3500
每月家庭消费支出Y（元）	561	638	869	1023	1254	1408	1650	1969	2090	2299
	594	748	913	1100	1309	1452	1738	1991	2134	2321
	627	814	924	1144	1364	1551	1749	2046	2178	2530
	638	847	979	1155	1397	1595	1804	2068	2266	2629
		935	1012	1210	1408	1650	1848	2101	2354	2860
		968	1045	1243	1474	1672	1881	2189	2486	2871
			1078	1254	1496	1683	1925	2233	2552	
			1122	1298	1496	1716	1969	2244	2585	
			1155	1331	1562	1749	2013	2299	2640	
			1188	1364	1573	1771	2035	2310		
			1210	1408	1606	1804	2101			
				1430	1650	1870	2112			
				1485	1716	1947	2200			
						2002				
总计	2420	4950	11495	16445	19305	23870	25025	21450	21285	15510

注：数据来自李子奈、潘文卿《计量经济学》（第三版）

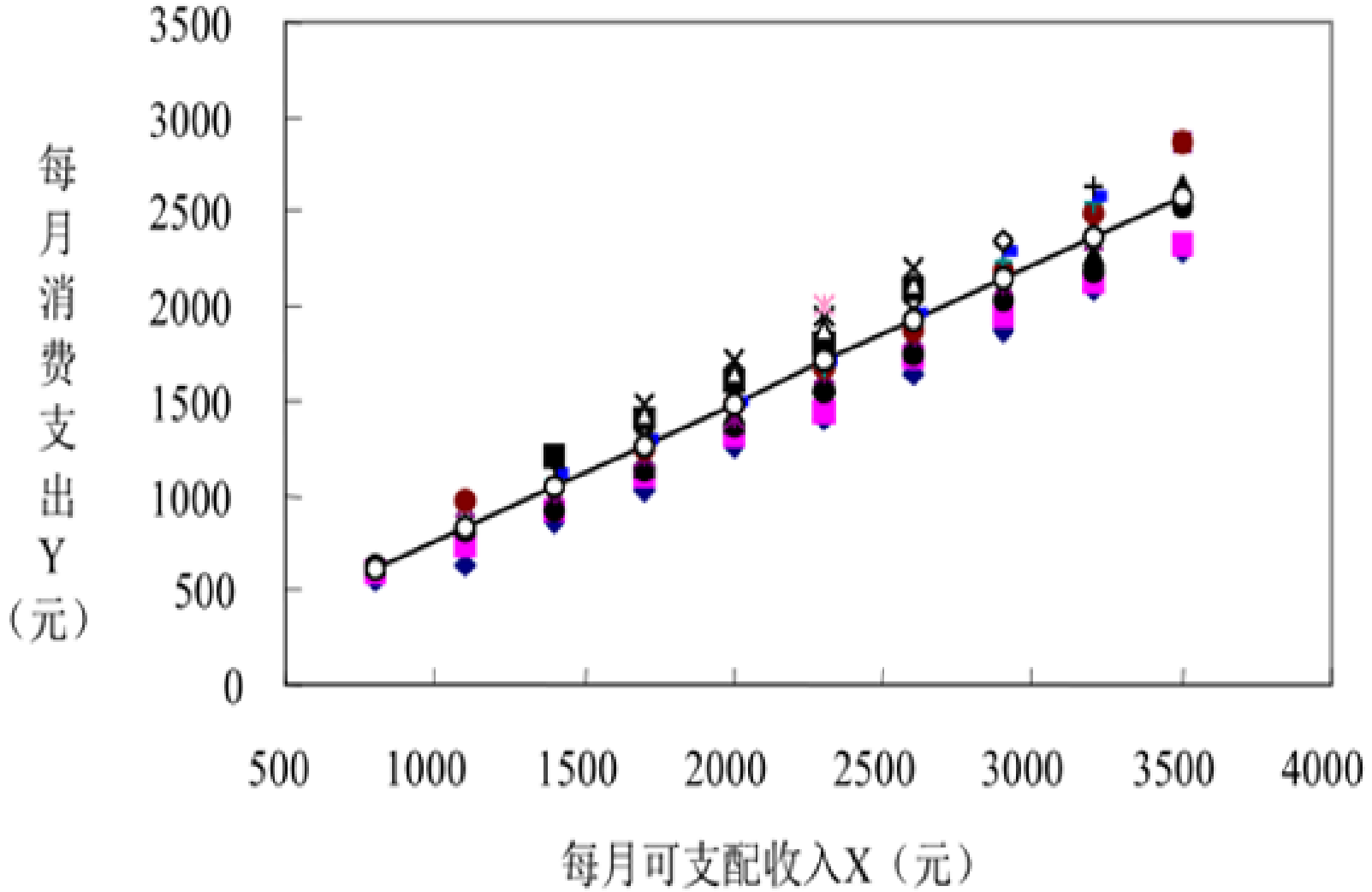
图 4: 调查数据

初步分析

* 从上表计算可以看出：

1. 可支配收入相同的家庭，其消费支出不一定相同，即收入和消费支出的关系不是完全确定的；
2. 由于是假想的总体，给定收入水平的消费支出的分布是确定的，即在给定下的条件分布（Conditional distribution）是已知的，如 $P(Y = 638|X = 800) = 1/4$

可视化


3500
3000
2500
2000
1500
1000
500
0
每月消费支出Y（元）
500
1000
1500
2000
2500
3000
3500
4000
每月可支配收入X（元）

- 从图 3 可以发现家庭消费支出的平均值随着收入的增加而增加，且的条件均值和收入近似落在一条直线上。我们称这条直线为总体回归线，相应的函数为，称为总体回归函数（population regression function, PRF），刻画了因变量的平均值随自变量变化的规律。
- 可以是线性的也可以是非线性的。例 3 中，将居民消费支出看成是其可支配收入的线性函数时，总体回归函数为 $E(Y|X_I) = \beta_0 + \beta_1 X_i$。
- 其中，β_0, β_1 是未知参数，也称为回归系数（regression coefficients）。

总体回归函数

- 总体回归函数描述了在给定的收入水平下，家庭的平均消费支出水平。但对某一个别的家庭，其消费支出可能与该平均水平有偏差。
- $\mu_i = Y_i - E(Y|X_i)$, 这是一个不可观测的随机变量，称为随机误差项（error term）或随机干扰项（disturbance）。
- 例 3 中，个别家庭的消费支出为：

$$y_i = E(y_i|x_i) + \mu_i = \beta_0 + \beta_1 x_i + \mu_i$$

- 即给定收入水平，个别家庭的消费支出可表示为两部分之和：

1. 该收入水平下所有家庭的平均消费支出，称为**系统性（systematic）或确定性（deterministic）**部分
2. μ_i 称为**其他随机或非确定性（nonsystematic）**部分。

- 例 1 的数据实际上是从例 3 的总体中抽取出来的样本数据。
- 从图 1 的样本散点图可以看出这些散点近似于一条直线，自然的想法是能否画一条直线尽可能好地拟合这些散点，这条直线称为样本回归线 (sample regression lines)。
- $\hat{y}_i = f(x_i) = \hat{\beta}_0 + \hat{\beta_1}_1 x_i$，称为样本回归函数（sample regression function, SRF)。
- 样本回归函数也有如下的随机形式：$y_i = \hat{y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$。
- 其中，e_i 称为残差（residual)，代表了其他影响 y_i 的随机因素的集合，可以看成是 μ_i 的估计量 $\hat{\mu}_i$。

参数估计

- 回归分析的主要目的是要通过样本回归函数（模型）SRF 尽可能准确地估计总体回归函数（模型）PRF，即根据 $y_i = \hat{y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta_1}_1 x_i + e_i$ 去估计，或者说利用 y_i 去估计。
- 参数估计方法有多种，其中使用最广泛的是普通最小二乘估计法（Ordinary Least Squares, OLS）和极大似然估计法（Maximum Likelihood Estimation, MLE）。
- 回归分析的主要目的是要通过样本回归函数（模型）SRF 尽可能准确地估计总体回归函数（模型）PRF，即根据 $y_i = \hat{y}_i + \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i$ 去估计，或者说利用 y_i 去估计。
- 参数估计方法有多种，其中使用最广泛的是普通最小二乘估计法（Ordinary Least Squares, OLS）和极大似然估计法（Maximum Likelihood Estimation, MLE）。

模型假设

- 为保证参数估计量具有良好的性质，通常要求模型满足若干基本假设：
  - 假设 1 自变量 X 是确定的，不是随机变量；
  - 假设 2 随机误差项 μ 具有零均值、同方差和无序列相关性，即：

$$E(\mu_i) = 0, i = 1, 2, \cdots, n$$

$$Var(\mu_i) = \sigma_\mu^2, i = 1, 2, \cdots, n$$

$$Cov(\mu_i, \mu_j) = 0, i \neq j \quad i, j = 1, 2, \cdots, n$$

- 假设 3 随机误差项与自变量之间不相关，即：

$$Cov(X_i, \mu_i) = 0, i = 1, 2, \cdots, n$$

- 假设 4 服从正态分布，即

$$\mu_i \ N(0, \sigma^2), i = 1, 2, \cdots, n$$

- 以上假设也称为线性回归模型的经典假设或高斯（Gauss）假设，满足该假设的线性回归模型，也称为经典线性回归模型（Classical Linear

普通最小二乘估计（OLS)

- 普通最小二乘法（Ordinary least squares, OLS）是求解参数，使得样本观测值和拟合值之差的平方和最小，即：

$$\min: Q = \sum_{i=1}^{n}(Y_i - \hat{Y})^2 = \sum_{i=1}^{n}(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \tag{3}$$

求解可得：

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum x_i y_i}{\sum x_i^2} \\ \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \end{cases} \tag{4}$$

其中 $\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{1}{n}(\sum X_i)^2$

$$\sum x_i y_i = \sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum X_i Y_i - \frac{1}{n}\sum X_i \sum Y_i$$

普通最小二乘估计（OLS)

- 当模型参数估计出后，需考察参数估计量的统计性质，可从如下几个方面考察其优劣性:
  1. 线性性，即它是否是另一随机变量的线性函数;
  2. 无偏性，即它的期望值是否等于总体的真实值;
  3. 有效性，即它是否在所有线性无偏估计量中具有最小方差。
- 这三个准则也称作估计量的小样本性质。拥有以上性质的估计量称为最佳线性无偏估计量（best liner unbiased estimator, BLUE)。
- 最小二乘法估计量具有高斯——马尔可夫定理（Gauss-Markov theorem)：在给定经典线性回归的假定下，最小二乘估计量是具有最小方差的线性无偏估计量（best linear unbiased estimator, BLUE)。。

参数估计量的概率分布及随机干扰项方差的估计

参数估计量 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的概率分布

普通最小二乘估计量 $\hat{\beta}_0$ 、$\hat{\beta}_1$ 分别是 Y_i 的线性组合，所以 $\hat{\beta}_0$ 与 $\hat{\beta}_1$ 的分布取决于Y的分布。在μ是正态分布的假设下，Y也是正态分布，则 $\hat{\beta}_0$ 、$\hat{\beta}_1$ 也服从正态分布，分别为

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2}) \quad , \quad \hat{\beta}_0 \sim N(\beta_0, \frac{\sum X_i^2}{n\sum x_i^2}\sigma^2)$$

随机误差项μ的方差σ^2的估计。

$\hat{\beta}_0$与$\hat{\beta}_1$的方差中，都含有随机扰动项μ的方差σ^2。由于σ^2实际上是未知的，因此$\hat{\beta}_0$与$\hat{\beta}_1$的方差实际上无法计算，这就需要对其进行估计。由于随机项μ_i不可观测，只能从μ_i的估计（残差e_i）出发，对σ^2进行估计。

σ^2的最小二乘估计量为$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$，可以证明它是$\sigma^2$的无偏估计量。

参数$\hat{\beta}_0$与$\hat{\beta}_1$的方差和标准差的估计量分别是：

$\hat{\beta}_0$ 的样本方差：$S^2_{\hat{\beta}_0} = \hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2$

$\hat{\beta}_0$ 的样本标准差：$S_{\hat{\beta}_0} = \hat{\sigma} \sqrt{\sum X_i^2 / n \sum x_i^2}$

$\hat{\beta}_1$ 的样本方差：$S^2_{\hat{\beta}_1} = \hat{\sigma}^2 / \sum x_i^2$

$\hat{\beta}_1$ 的样本标准差：$S_{\hat{\beta}_1} = \hat{\sigma} / \sqrt{\sum x_i^2}$

模型检验

- 回归分析的目的是要通过样本所估计的参数 $(\hat{\beta}_0, \hat{\beta}_1)$ 来代替总体的真实参数 (β_0, β_1) 或者说是用样本回归线代替总体回归线。
- 尽管从统计性质上可以保证如果有足够多的重复抽样，参数的估计值的期望（均值）就等于其总体的参数真值，即具有无偏性。
- 但在一次抽样中，估计值不一定就等于该真值。那么，在一次抽样中，参数的估计值与真值的差异有多大、是否显著，这就需要进一步进行统计检验，主要有拟合优度检验、变量的显著性检验。

拟合优度检验

拟合优度检验是对回归拟合值与观测值之间拟合程度的一种检验。度量拟合优度的指标主要是判定系数（可决系数）R^2。要理解R^2需先理解总离差平方和的分解。

Y的第i个观测值与样本均值的离差 $y_i = \left(Y_i - \bar{Y}\right)$ 可分解为两部分之和：

$$y_i = \left(Y_i - \bar{Y}\right) = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right) = e_i + \hat{y}_i \tag{5}$$

其中，$y_i = \left(Y_i - \bar{Y}\right)$ 是样本回归拟合值与观测值的平均值之差，可认为是由回归直线解释的部分；$e_i = \left(Y_i - Y_i\right)$ 是实际观测值与回归拟合值之差，是回归直线不能解释的部分。如果 $Y_i = \bar{Y}_i$ 即实际观测值落在样本回归“线”上，则拟合得最好。对于所有样本点，则需考虑这些点与样本均值离差的平方和,可以证明

$$\sum_{i=1}^{n} y_i^2 = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n} \hat{y}_i e_i = \sum_{i=1}^{n} \hat{y}_i^2 + \sum_{i=1}^{n} e_i^2 \tag{6}$$

$TSS=\sum_{i=1}^{n}y_i^2=\sum_{i=1}^{n}(Y_i-\bar{Y})^2$ 为**总体平方和**（Total Sum of Squares）

$ESS=\sum_{i=1}^{n}\hat{y}_i^2=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$ 为**回归平方和**（Explained Sum of Squares）

$RSS=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$ 为**残差平方和**（Residual Sum of Squares）

三者之间有如下关系：$TSS=ESS+RSS$，所以，Y 的观测值围绕其均值的总离差（total variation）可分解为两部分：一部分来自回归（ESS），另一部分则来自随机因素（RSS）。在给定样本下，TSS不变，如果实际观测点离样本回归线越近，则ESS在TSS中占的比重越大。

记 $R^2=\frac{ESS}{TSS}=1-\frac{RSS}{TSS}$，称 R^2 为**可决系数**（coefficient of determination）。R^2 的取值范围为[0，1]，R^2越接近1，说明实际观测点离样本线越近，拟合优度越高。

变量显著性检验

回归分析的目的之一是要判断X 是否是Y 的一个显著影响因素。这就需要进行变量的显著性检验。我们已经知道回归系数估计量 $\hat{\beta}_1$ 服从正态分布，即 $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum x_i^2})$ 。又由于真实的σ^2 未知，利用它的无偏估计量$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ 替代时，可构造检验统计量

$$t = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{\sigma}^2 / \sum x_i^2}} = \frac{\hat{\beta}_1 - \beta_1}{S_{\hat{\beta}_1}} \sim t(n-2) \qquad (7)$$

进行检验。

变量显著性检验步骤

对总体参数提出假设 $H_0: \beta_1 = 0,\ H1: \beta_1 \neq 0$

在原假设H_0 成立下，构造t 统计量 $t = \dfrac{\hat{\beta}_1}{S_{\hat{\beta}_1}}$ ；

给定显著性水平α，查 t分布表，得临界值 $t_{1-\alpha/2}(n-2)$;

比较，判断：若 $|t| > t_{1-\alpha/2}(n-2)$，则拒绝 H_0，接受H_1；

若 $|t| \leq t_{1-\alpha/2}(n-2)$，则不拒绝 H_0

对于一元线性回归方程中的截距项 $\hat{\beta}_0$，同理可构造如下 统计量：

$$t = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{\sigma}^2 \sum X_i^2 / n \sum x_i^2}} = \frac{\hat{\beta}_0}{S_{\hat{\beta}_0}} \sim t(n-2) \qquad (8)$$

线性回归预测

- 对于拟合得到的一元线性回归模型 $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$，给定样本以外的自变量观测值 x_0 ，可以得到因变量的预测值 $\hat{y}_0$，并以此作为其条件均值 $E(y|x = x_0)$ 或个别值 y_0 的一个近似估计，我们称之为点预测。
- 给定显著性水平下，可以求出 y_0 的预测区间，我们称之为区间预测。

点预测

对总体回归函数 $E(Y|X)=\beta_0+\beta_1X$，当$X=X_0$ 时，$E(Y|X=X_0)=\beta_0+\beta_1X_0$ 。通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得拟合值为 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$，两边取期望可得，

$$E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0=E(Y|X=X_0) \quad (9)$$

可见，$\hat{Y}_0$ 是 $E(Y|X=X_0)$ 的无偏估计。

对总体回归模型 $Y=\beta_0+\beta_1X+\mu$，当$X=X_0$ 时，$Y_0=\beta_0+\beta_1X_0+\mu$，两边取期望可得

$$E(Y_0)=E(\beta_0+\beta_1X_0+\mu)=\beta_0+\beta_1X_0+E(\mu)=\beta_0+\beta_1X_0 \quad (10)$$

而通过样本回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1X$，求得拟合值为 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$ 的期望为

$$E(\hat{Y}_0)=E(\hat{\beta}_0+\hat{\beta}_1X_0)=E(\hat{\beta}_0)+X_0E(\hat{\beta}_1)=\beta_0+\beta_1X_0\neq Y_0 \quad (11)$$

可见 $\hat{Y}_0$不是个值Y_0 的无偏估计。

区间预测

由于 $\hat{Y}_0=\hat{\beta}_0+\hat{\beta}_1X_0$，$\hat{\beta}_1\sim N(\beta_1,\frac{\sigma^2}{\sum x_i^2})$，$\hat{\beta}_0\sim N(\beta_0,\frac{\sum X_i^2}{n\sum x_i^2}\sigma^2)$，可以证明：

$$\hat{Y}_0\sim \mathrm{N}(\beta_0+\beta_1X_0,\sigma^2(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}x_i^2}))$$

t 统计量：$$t=\frac{\hat{Y}_0-(\beta_0+\beta_1X_0)}{s_{\hat{Y}_0}}=\frac{\hat{Y}_0-(\beta_0+\beta_1X_0)}{\sqrt{\hat{\sigma}^2(\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum_{i=1}^{n}(X_i-\bar{X})^2})}}\sim t(n-2) \quad (12)$$

在给定显著性水平 下，总体均值 $E(Y_0\mid X_0)$ 的置信区间为：

$$\hat{Y}_0-t_{1-\frac{\alpha}{2}}\times S_{\hat{Y}_0}<E(Y_0\mid X_0)<\hat{Y}_0+t_{1-\frac{\alpha}{2}}\times S_{\hat{Y}_0} \quad (13)$$

这也称为 $E(Y_0\mid X_0)$ 的区间预测。

区间预测

由于 $Y_0 = \beta_0 + \beta_1 X_0 + \mu$ 可得 $Y_0 \sim N(\beta_0 + \beta_1 X_0, \sigma^2)$，进而我们可以得到 $\hat{Y}_0 - Y_0$ 的分布为：

$$\hat{Y}_0 - Y_0 \sim N\left(0, \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum x_i^2}\right)\right) \tag{14}$$

t 统计量：

$$t = \frac{\hat{Y}_0 - Y_0}{s_{\hat{Y}_0 - Y_0}} = \frac{\hat{Y}_0 - Y_0}{\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)}} \sim t(n-2) \tag{15}$$

在给定显著性水平 下，总体均值 Y_0 的置信区间为：

$$\hat{Y}_0 - t_{1-\frac{\alpha}{2}} \times S_{\hat{Y}_0 - Y_0} < Y_0 < \hat{Y}_0 + t_{1-\frac{\alpha}{2}} \times S_{\hat{Y}_0 - Y_0} \tag{16}$$

这也称为 Y_0 的区间预测。

多元线性回归

中国税收收入相关数据研究

- 为了研究影响中国税收收入增长的主要原因，预测中国税收未来的增长趋势，需要建立回归模型。
- 影响中国税收收入增长的因素很多，选择包括中央和地方税收的“国家财政收入”中的“各项税收”（简称“税收收入”）作为因变量，以反映国家税收的增长；选择“国内生产总值（GDP）”作为经济整体增长水平的代表；选择中央和地方“财政支出”作为公共财政需求的代表；选择“商品零售物价指数”作为物价水平的代表。（如表 4 所示）

数据

表4　中国税收收入相关数据

年份	tax	GDP	expand	CPI	年份	tax	GDP	expand	CPI
1978	519.28	3645.22	1122.09	100.7	1996	6909.82	70142.49	7937.55	106.1
1979	537.82	4062.58	1281.79	102	1997	8234.04	78060.85	9233.56	100.8
1980	571.7	4545.62	1228.83	106	1998	9262.8	83024.33	10798.18	97.4
…	…	…	…	…	…	…	…	…	…
1994	5126.88	48108.46	5792.62	121.7	2012	100614.3	516282.1	125953	102
1995	6038.04	59810.53	6823.72	114.8					

例4中自变量个数不止一个，该如何建模分析？

利用多元回归分析方法。可以建立模型

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \varepsilon_i \qquad (17)$$

图 6: 税收数据

多元线性回归模型及假定

线性模型的一般形式是 $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_k X_{ki} + \varepsilon_i$， i=1, 2, ⋯, n （18）

其中，Y_i 为因变量，$X_{2i}, X_{3i}, \ldots, X_{ki}$ 为自变量，ε_i是随机误差项，β_1 为模型的截距项，β_j（j=2, K, k ）为模型回归系数。

我们还可将上述模型用矩阵形式记为$Y = X\beta + \varepsilon$ （19）

总体回归方程为 $E(Y|X) = X\beta$。其中，X 是由1组成的列向量和$X_2, X_3, \ldots X_k$构成的设计矩阵，其中截距项可视为取值为1的自变量。

图 7: 模型假定

样本回归模型为：$\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{e}$ (20)

样本回归方程为：$\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ (21)

这里$\hat{\boldsymbol{Y}}$表示Y的样本估计值向量；$\hat{\boldsymbol{\beta}}$表示回归系数β估计值向量；e表示残差向量。

图 8: 模型假定

- 零均值。假定随机干扰项 的期望向量或均值向量为零，即 $E(\boldsymbol{\varepsilon}) = E\begin{pmatrix}\varepsilon_1\\ \varepsilon_2\\ \vdots\\ \varepsilon_n\end{pmatrix} = \begin{pmatrix}E\varepsilon_1\\ E\varepsilon_2\\ \vdots\\ E\varepsilon_n\end{pmatrix} = \begin{pmatrix}0\\ 0\\ \vdots\\ 0\end{pmatrix} = \mathbf{0}$
- 同方差和无序列相关。假定随机干扰项 不存在序列相关且方差相同，即

$$Var(\boldsymbol{\varepsilon}) = E\left[(\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} - E\boldsymbol{\varepsilon})'\right] = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \boldsymbol{\sigma}^2\boldsymbol{I}_n$$

- 随机干扰项 与自变量相互独立。即$E(X'\varepsilon) = 0$
- 无多重共线性。假定数据矩阵X列满秩，即*Rank (X) =k*
- *正态性。假定*$\varepsilon \sim N(0, \sigma^2 I_n)$

图 9: 模型假定条件

参数估计

对于总体回归模型$Y = X\beta + \varepsilon$，求参数β的方法是最小二乘（OLS法）。即求$\hat{\beta}$使得残差平方和$\sum e_i^2 = e'e$达到最小。令：

$$\begin{aligned} Q(\hat{\beta}) &= e'e \\ &= (Y - X\hat{\beta})'(Y - X\hat{\beta}) \\ &= Y'Y - \hat{\beta}'X'Y - Y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \end{aligned} \tag{22}$$

对上式关于$\hat{\beta}$求偏导，并令其为零，可得$\hat{\beta} = (X'X)^{-1}X'Y$ (23)

这就是线性回归模型参数的最小二乘估计量。

图 10: 参数估计

参数估计

- 在线性模型经典假设的前提下，线性回归模型参数的最小二乘估计具有优良的性质，满足高斯—马尔可夫（Gauss–Markov）定理，即在线性模型的经典假设下，参数的最小二乘估计量是线性无偏估计中方差最小的估计量（BLUE 估计量）

参数 σ^2 的估计量可以用 $s^2 = (\frac{e'e}{n-k})$，可以证明 s^2 为 σ^2 的无偏估计量，即

$$E(s^2) = E\left(\frac{e'e}{n-k}\right) = \sigma^2$$

模型检验

- 拟合优度检验
- 方程整体显著性检验
- 单个变量的显著性检验

R 实现

一元线性

- lm1 <- lm (consum ~ income)
- coef (lm1)
- coef (lm (consum ~ - 1 + income))
- summary (lm1)

一元预测

- conf <- predict (lm1 , data.frame (income = sx) , interval = "confidence") # 区间
- pred <- predict (lm1 , data.frame (income = sx) , interval = "prediction") # 单点
- plot (income , consum) # 画散点图
- abline (lm1) # 添加回归线
- lines (sx , conf [, 2]) ; lines (sx , conf [, 3])
- lines (sx , pred [, 2] , lty = 3) ; lines (sx , pred [, 3] , lty = 3)

本周推荐

1. 一本书：《离心力》，Johnny Ryan，电子工业出版社，2018（哈佛，斯坦福学生读物）
2. 一部电影：《三傻大脑宝莱坞（3 idiots）》，2009
3. 练习：《R 语言实战（第 2 版）》，第 8 章代码实现
4. 作业：生成一个数据（包含多个 x，一个 y，构建一个线性回归模型）

谢　　谢!