第九章 线性回归

(Linear Regression)

导论

- 统计分析：根据统计数据提供的资料，揭示变量之间的关系，并由此推演为事物之间内在联络的规律性

为何学习回归分析

- 回归分析探讨客观事物之间的联络，体现为变量之间的统计关系
- 建立在对客观事物进行大量试验和观察的基础上，用来寻找隐藏在看起来不拟定的现象中的统计规律的统计措施
- 因因变量衡量方式的不同，回归分析可分为线性回归和非线性回归
  - 线性回归合用于因变量为连续衡量的场合
  - 非线性回归多合用于因变量为虚拟变量、多分类变量、计数变量等场合
  - 即便在这两大类中，分析措施又可区别为许多不同的类型
- 根据处理的变量多少来看，回归分析又分为：
  - 简朴有关和一元回归：研究的是两个变量之间的关系
  - 多元有关或多元回归：研究的是多种变量之间的关系

本章主要内容

9.1. 变量间的有关关系（correlation）
9.2. 线性回归概述
9.3. 一元线性回归
9.4. 多元线性回归

9.1. 变量间的有关关系（correlation）

1、函数关系

- 回归分析前，首先必须掌握变量之间是否有关；只有变量之间存在关系，才有必要进行回归分析
- 假若x增长时，y的取值发生相应变化，则x和y之间是有关的
- 假若x增长时，y的取值没有拟定的变化，x则y和之间是不有关的，或是没有有关关系的
- 变量之间的有关关系归纳起来能够分为两种：**函数关系**和**统计关系**

● 函数关系：当一个变量 x 取一定的值时，另一变量 可以按照确定的函数公式 y 取一个确定的值，记为

$y = f(x)$ ，则称 y 是 x 的函数，换言之，y 与 x 之间存在函数关系，函数关系是一一对应的确定性关系，比较 度（比如，路程 = 时间

$\times$ 速度）；长方形的边长与面积之间的关系，等

2、统计有关

- 现实事物之间的联络不像函数关系那样轻易拟定
- 现象之间存在关联；但无法拟定详细关系，不能像函数关系那样，用一种公式将它们的关系精确地描述出来；当一种变量取一定的值时，另一种变量可能有多种取值
- **当一种变量的值不能由另一种变量的值唯一拟定时，这种关系称为统计关系**
- 统计关系不如函数关系直接和明确；但经过对大量数据的观察和研究，就会发觉许多变量之间确实存在着某种关联，强弱各不相同

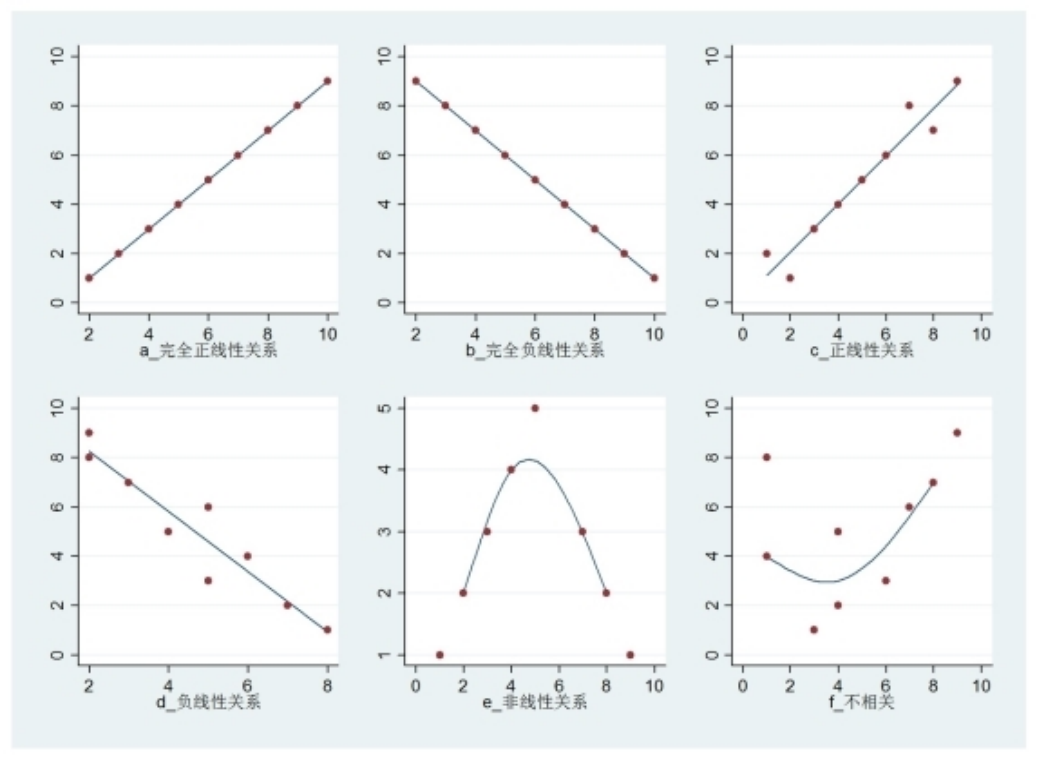
3、有关关系的特点

- 双向变化关系；
- 一种变量的取值不能由另一种变量的取值唯一拟定；当x取一定的值时，y可能有多种取值，因为还受到其他原因的影响；
- 不拟定关系难以用函数关系来衡量和描述，但这并不表达x和y之间无规律可循；
- 类似定性描述
- 有关分为线性有关和非线性有关。

4、有关分析

- 对两（多）个变量之间的关系进行描述，分析它们
  - 是否有关
  - 关系是否亲密
  - 关系的性质是什么（是正有关还是负有关）
    - 伴随x的变化，y值的变化程度就拟定两者是否有关和有关的强度
    - 当x增长（降低）时，y的取值也随之增长（降低），则x和y之间呈正有关关系；相反，当x增长（降低）时，y的取值却随之降低（增长），则和之间呈负有关关系
- 有关分析的措施涉及散点图和有关系数

有关散点图（ scatter plot ）



有关系数

- 图形虽然直观，但不够精确；对散点图的视觉分析带有很大的主观性；需要更精确和更客观的度量；
- 有关系数可精确地描述变量之间的线性有关程度；
- 线性**有关系数**是衡量变量之间有关程度的统计量，是描述两变量线性关系强度及方向的数值；
- 若有关系数是根据总体数据计算的，称为总体有关系数，记为ρ；若是根据样本计算出来的，则称为样本有关系数，记为**r**；
- 在统计学中，一般用样本有关系数来推断总体有关系数。

有关系数：性质与方向

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right)$$

- 相关系数 r 的取值在 -1 和 1 之间，变量的线性关系强度随 r 从

r

接近 0 表示两变量的线性关系薄弱

r

接近 -1 或 1 表示散点图的点对于直线

r

等于-1 或 1 表示散点图的点全在直线上

- 相关系数为正值时，表示两变量正向关联
- 相关系数为负值时，表示两变量负向联系

有关系数：程度

- 根据经验，能够将有关程度分为几等：
  - $|r| \geq$ 0.8，变量之间高度相关
  - $\leq |r| <$ 0.8，变量之间中度相关
  - 0.3
  - $\leq |r| <$ 0.5，变量之间低度相关
  - $|r| <$
- 但这种解释必须建立在对有关系数进行明显性检验的基础之上。

有关系数：其他特征

（a） r

(b) 无区分 DV 和 IV；相关关系不一定就是因果关系

r

(c) 的计算以数值型变量为主，不适用于类别变量

r

(d) 的计算使用 Z 值，与各数值型变量的度量单位无关

r

仅能衡量变量的线性关系，无法衡量曲线关系强度

i.

r

=0 并不等于 变量间无任何关系，而仅仅表示它们之间不存在线性相关关系；二者可能有非线性关系

ii. 当变量间的非线性关系程度较大时，可导致

r

=0 或很小时，不能轻易说二者不相关，而应该结合散点图做出合理的解释（陈 81e）

（e） 相关系数值受离群点(outliers)影响很大，基于平均的数据会中和个体之间的差异，并可能令人相关系数

r

=0

iii. 当

有关系数的计算

- **. correlate [变量名]**
-
- ①　　②
- ①：**. correlate**也可写为**corr**，是生成变量之间有关系数矩阵、协有关矩阵、回归系数有关矩阵的基本命令；
- ②：需要生成有关关系的变量名称
- **如：corr age edu weight height**

- **若要给出有关系数（每个变量的上行）及其假定检验的P指，使用命令：**
- **pwcorr [变量名],sig**

9.2. 线性回归概述

“回归”一词来自英国学者、优生学的创始人S. F. Galton（1822-1911）。Galton在对遗传现象进行研究后发觉，当高个的夫妻或矮个的夫妻有了孩子时，这些孩子的身高趋于回归到更经典的、同一性别的人的平均身高。

1、回归分析

- 经过找出代表变量之间关系的直线图形或直线方程来描述变量之间的数学关系
  - 这条直线称为回归直线；
  - 该直线方程称为回归方程。
- 一元线性回归是回归分析中最简朴、最基本的回归分析，描述两个变量之间的关系。
- 它是根据统计资料，谋求一种变量与另一种变量关系的恰当数学体现式的经验方程，来近似地表达变量间的平均变化关系的一种统计分析措施：
  - 其中一种变量作为DV或被解释变量，一般用y表达；
  - 另一种变量IV（预测变量或解释变量）一般用x表达。

2、有关分析与回归分析之别

- **依存关系与平等关系。**回归反应两个变量的依存关系，一种变量的变化引起另一种变量的变化，是一种单向的关系；其y变量称为因变量，被解释变量；在有关分析中，变量和变量处于平等地位：双向关系
- **关系程度与影响大小**。有关分析主要是刻画两类变量间线形有关的亲密程度；而回归分析不但要揭示自变量对因变量的影响大小，还能够由回归方程进行预测和控制。所以，回归是对两（多）个变量作定量描述，研究变量之间的数量关系，从已知的一种变量的取值预测另一种变量的取值，得到定量成果。

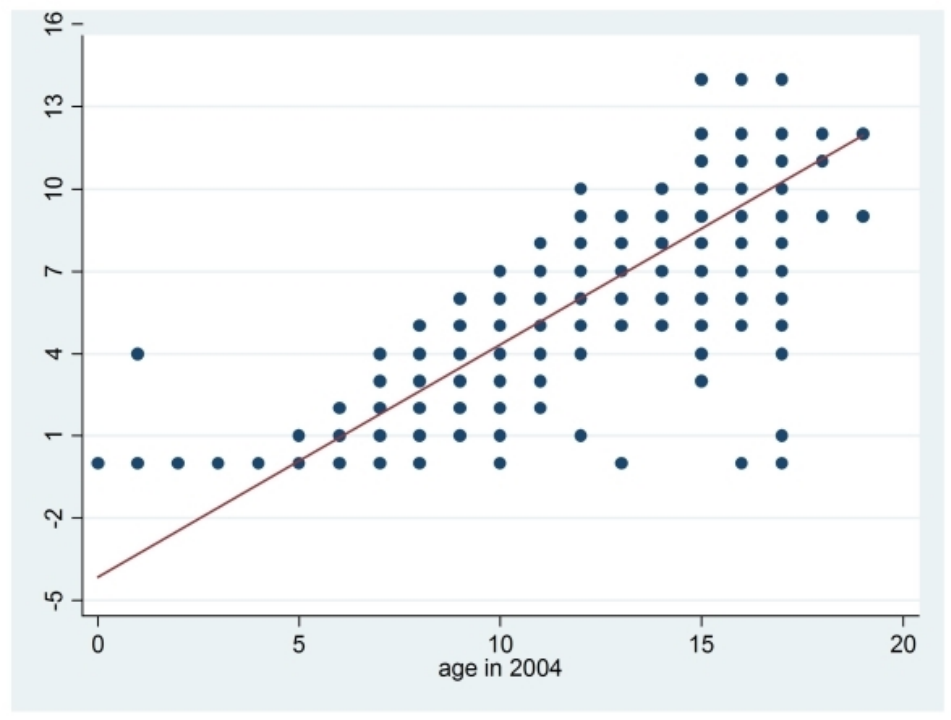
3、回归分析的目的

- 从一组样本数据出发，拟定解释变量（ IV ）与被解释变量（DV）之间的数学关系式；回归方程就是要找出一条最佳地描述两个变量之间关系的直线方程。
- 对该关系式的可信程度进行多种统计检验；从影响DV的一组IV中找出哪些变量的影响是明显的，那些是不明显的。
- 利用直线方程（即所求的关系），根据一种或几种变量的取值来估计或预测DV的取值，并给出这种估计或预测的置信度。
  - 预测是有规律的。如，
    - 利用汽车的速度来预测它刹车所需要的距离
    - 利用学生的中学成绩来预测考上大学的成功率
  - 精确的y值是不可预测的，接近实际值。

4、回归分析的用途

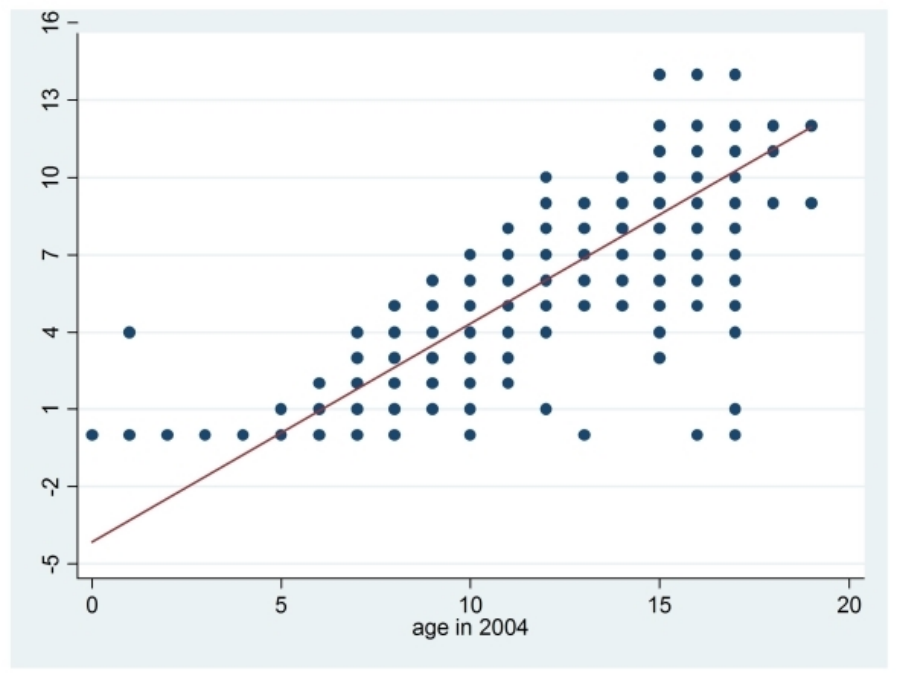
- **用于研究一种IV对一种数值型DV在数量上的影响程度。设有两个变量，x，y，其中，y的取值随x取值的变化而变化，故y是DV，x是IV。**

← 对于这两个变量，经过观察或试验得到若干组数据，记为1，2，…，n)。若x以代表年龄，以y代表教育，则从散点图中，能够清楚地确认x与y存在线性关系



线性回归模型：回归直线

- 小朋友的年龄与教育之间存在很强的正向有关关系
- 线条就是**回归直线**（regression line）
- 怎样将变量之间的有关关系用数学关系的代数体现式体现出来



线性回归的理论模型

$$y = \beta_0 + \beta_1 x + \varepsilon \qquad (9.1)$$

- 等式（9.1）称为一元线性回归模型，描述因变量y怎样依赖于自变量x和误差项e而异。在该模型中，y是x的线性函数（β0+β1x 部分）加上误差项e。其中，
  - β0和β1是模型的未知参数，前者称为回归常数项（或截距，intercept）；后者称为回归系数（coefficient）；β0+β1x反应了因为x的变化而引起的y的变化，也称为**边际变化（当变量x变化一种单位时，变量y变化的数量）**
  - e是被称为误差项的随机变量，代表因主观和客观原因而不可观察的随机误差，反应了除x和y之间的线性关系之外的随机原因对y的影响，是不能由x和y的线性关系所揭示的变异性。

线性回归模型的基本假定

（1）零均值，即 $E(\varepsilon)=0$ 。误差项是一种期望值=0 的随机变量

- 在自变量取一定值的条件下，其总体各误差项的条件平均值为0。这意味着，在等式（9.1）中，因为β0和β1都是常数或系数，故有

$$E(\beta_0)=\beta_0$$

- 所以，对于一种给定的x值，y的期望值为 $E(\beta_1)=\beta_1$

（2）等方差，即对于全部的x值，e的方差 σ^2 都相同 $E(y)=\beta_0+\beta_1 x$

（3）误差项服从正态分布，且相互独立，即

$$\varepsilon \sim N(0,\sigma^2) \qquad (9.2)$$

独立性意味着，对于一个特定的 x 值，它所对应的 ε 值，它所对应的 y 与其它 x 值，它所对应的 y 不相关

ε 与其它 x 值对应的 ε 对于一个特定的 x

总体回归方程 (equation)（Ⅰ）

- 根据回归模型的假定，误差项的期望值为0；所以，y的期望值等于

$$E(y) = \beta_0 + \beta_1 x \qquad (9.3)$$

- 一元**总体回归方程**：

$$y = \beta_0 + \beta_1 x \qquad (9.4)$$

- 方程是一条直线，故也称直线回归方程（linear regression equation）
  - β_0 是回归系数在 y 轴上的截距，是当 x 为 0 时 y 的取值
  - β_1 是直线的斜率，表示当 x 每变动一个单位时，y 的变化值

总体回归方程（II）

- 等式（9.4）从平均意义上表达了变量 y 与 x

β_0 β_1

x

β_0 β_1

（预测的）回归方程

- 一元线性回归模型的**样本**回归方程可以表示为：

$$\cdot\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \quad (9.5)$$

$\hat{y}$ 读成 y hat）代表对应于一个特殊的 x 值的 y 的预测值； $\hat{y}$ 的估计值 是

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是样本的统计量，分别代表总体的未知参数 β_0 和 β_1

$\hat{\beta}_0$ 为估计的回归直线在 y 轴上的截距

$\hat{\beta}_1$ 为估计的回归直线的斜率，它表示对于一个给定的 x 值，y 的平均变动值

回归分析的三种检验

- F检验——用于检验回归方程的明显性
- R^2——用于测度回归直线对观察数据的拟合程度；也称鉴定系数、可决系数（coefficient of determination）
- t检验——用于检验自变量回归系数的明显性

reg edu age

Source	SS	df	MS	Number of obs =	1749
				F(1, 1747) =	10416.64
Model	14779.6461	1	14779.6461	Prob > F =	0.0000
Residual	2478.73126	1747	1.41885018	R-squared =	0.8564
				Adj R-squared =	0.8563
Total	17258.3774	1748	9.87321359	Root MSE =	1.1912

edu	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	.846604	.008295	102.06	0.000	.8303349	.8628732
_cons	-4.141329	.1031883	-40.13	0.000	-4.343714	-3.938943

- **上部分分为左右两个区域**
  - 左边是方差分析。方差部分给出回归平方和（Model）、残差平方和（Residual），总平方和（Total），自由度（df），回归和残差的均方（MS）
  - 右边是回归统计量。涉及检验统计量（F），F检验的明显水平（Prob>F），R2（R-square）（鉴定系数），Adj R-squared（调整后的R2），观察值的个数（即样本量），估计原则误差（Root MSE）
- **下部分是参数估计的内容**。涉及回归方程截距（_cons）和斜率（Coef）的参数估计、原则误、t 统计量，P值（P>|t|），置信区间（[95% Conf. Interval]）

（1）鉴定系数R^2：概念与计算

- 对估计的回归方程拟合优度的度量，也就是要检验样本数据汇集在样本回归直线周围的密集程度，从而鉴定回归方程对样本数据的代表程度；
- 该指标是建立在对总离差（deviation）平方和进行分解的基础之上。显然，各样本观察点（散点）与样本回归直线靠得越紧，SSR/SST则越大，直线拟合得越好。将该百分比定义为鉴定系数或可决系数，记为R^2

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$$

（1）鉴定系数R^2：意义

- 若全部观察值都落在回归直线上，则 $R^2 = 1$，拟合是完全的，模型具有完全解释能力；假如回归直线没有解释任何离差，则y的总离差全部归于残差平方和，即SST=SSE， $R^2 = 0$，表达自变量x对因变量y的变异缺乏解释能力
- 一般观察值都是部分地落在回归线上，即$0< R^2 <1$
- R^2 越接近1，表白回归直线的拟合优度越好；反之， R^2 越接近于0，则回归直线的拟合程度就越差。
- 就上面的例子而言，鉴定系数的实际意义是，在教育水平取值的离差中，有将近86%能够由年龄与教育之间的线性关系来解释；即86%的差别是由孩子的年龄决定的——孩子年龄能够解释教育差别的86%。

（2）F检验：概念

- 对总体回归方程的明显性检验，是对因变量与全部自变量之间的线性关系是否明显的一种假设检验；
- 回归分析的主要目的是，根据所建立的估计方程用自变量x来估计、预测因变量y的取值；
- 当我们建立了估计方程后，还不能立即进行估计或预测，因为该估计方程是根据样本数据得出的，它是否真实地反应了变量x和y之间的关系，需要经过检验后才干证明；
- 该检验利用方差分析的措施进行。F统计量定义为：平均的回归平方和与平均的残差平方和（均方误差）之比。