计量经济学STAT30021

第三讲：简单线性回归分析(2)

肖志国

复旦大学管理学院

2025年9月

简单线性回归模型

基本框架：$\{(X_i, Y_i), i = 1, \ldots, n\}$为数据，且满足：对于所有的$i = 1, \ldots, n$，

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \tag{1}$$

以及

$$E[u_i | X_i] = 0. \tag{2}$$

- 参数β_0, β_1的估计：

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

- 最小二乘法的性质
- $\hat{\beta}_1$与相关系数的关系
- 模型的解释能力:R^2
- R^2与相关系数的关系

回归模型的应用

考虑大学成绩(colgpa)与高中成绩(hsgpa)的关系。通过一组30个学生的数据，我们得到如下的回归方程：

$$\text{colgpa} = 1.41 + 0.52 \times \text{hsgpa}$$

我们可能关心的问题有如下几类：

1. Y的变化对X的变化的敏感度：某学生A的高中成绩比学生B的高中成绩多0.5个绩点，那么A的大学成绩比B高出多少？
2. 给定X的取值，预测Y的取值：如果某学生的高中成绩为3.5，那么其大学成绩为多少？
3. 模型的拟合程度：高中成绩到底能解释多少的大学成绩？

回归系数的大小

是不是回归系数$\hat{\beta}_1$的值越大，就表示X对Y的影响就越大？

- $\hat{\beta}_1$: 当X增加一个单位时，Y平均的变化量；或者说，是Y对X的敏感度。
- $\hat{\beta}_0$: 当X为零时，Y的平均值？
- 严格说来，回归分析的结果只适用于当前数据。也就是说，他们对于新的数据来说未必成立。
- 但是我们通常假定我们通过当前数据$\{(X_i, Y_i), i = 1, \ldots, n\}$得到的结果能够近似的刻划$X$与$Y$的关系。
- 实际问题中，X与Y的线性回归关系可能只对于特定范围取值的(X, Y)成立(我们常常忽视这一点)。
- 不能过度解释回归分析的结果。要对回归分析结果成立的范围有个大致的概念。

Y的测量单位的变化对回归结果的影响

假设将所有的Y_i都乘以一个常数λ，用新的Y来和原来的X做回归。回归系数会有什么变化？R^2呢？记新的结果为$\hat{\beta}_{0\lambda}, \hat{\beta}_{1\lambda}$以及$R^2_\lambda$。

$$
\begin{aligned}
\hat{\beta}_{1\lambda} &= \frac{\lambda \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \lambda \hat{\beta}_1 \\
\hat{\beta}_{0\lambda} &= \lambda \bar{Y} - \hat{\beta}_{1\lambda} \bar{X} = \lambda \bar{Y} - \hat{\beta}_{1\lambda} \bar{X} = \lambda \hat{\beta}_0 \\
R^2_\lambda &= r^2_{X(\lambda Y)} = \frac{\lambda^2 \sum_{i=1}^{n} \left[(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2 \lambda^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2} = R^2
\end{aligned}
$$

X的测量单位的变化对回归结果的影响

假设将所有的X_i都乘以一个常数γ，用新的X来和原来的Y做回归。回归系数会有什么变化？R^2呢？记新的结果为$\hat{\beta}_{0\gamma}, \hat{\beta}_{1\gamma}$以及$R^2_\gamma$。

$$
\begin{aligned}
\hat{\beta}_{1\gamma} &= \frac{\gamma \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\gamma^2 \sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{1}{\gamma} \hat{\beta}_1 \\
\hat{\beta}_{0\gamma} &= \bar{Y} - \hat{\beta}_{1\gamma}(\gamma \bar{X}) = \bar{Y} - \hat{\beta}_1 \bar{X} = \hat{\beta}_0 \\
R^2_\gamma &= r^2_{(\gamma X)Y} = \frac{\gamma^2 \sum_{i=1}^{n} \left[(X_i - \bar{X})(Y_i - \bar{Y})\right]^2}{\gamma^2 \sum_{i=1}^{n} (X_i - \bar{X})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2} = R^2
\end{aligned}
$$

取log还是不取log

基本模型是

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

这个模型描述的是X的变化对Y的变化的影响。出于以下一些原因，我们可能需要对X，或者Y，或者同时做一个非线性变换,比如，取log:

- 原模型的拟合程度不好，我们希望通过变换增加拟合程度
- 我们关心的是X的变化对Y的百分比变化的影响
- 我们关心的是X的百分比变化对Y的变化的影响
- 我们关心的是X的百分比变化对Y的百分比变化的影响，也就是说，Y对X的弹性

取log的方式

我们可以有三种方式取log，每一种方式代表一种不同的模型，从而其参数的解释也完全不同：

1. 对Y取log:
$$\log(Y_i) = \beta_0 + \beta_1 X_i + u_i,$$
此模型中β_1表示X的单位变化对Y的百分比变化的影响
2. 对X取log:
$$Y_i = \beta_0 + \beta_1 \log(X_i) + u_i,$$
此模型中β_1表示X的百分比变化对Y的单位变化的影响
3. 对X和Y都取log:
$$\log(Y_i) = \beta_0 + \beta_1 \log(X_i) + u_i,$$
此模型又称为一个常弹性模型，其中β_1表示Y对X的弹性

百分比变化

某变量y增长了10%，则我们称其百分比变化为$+10$，记作$\% \triangle y = 10$；

某变量y减少了20%，则我们称其百分比变化为-20，记作$\% \triangle y = -20$。

依定义，

$$\% \triangle y = 100 \frac{\triangle y}{y}$$

取与不取log：实际运用的原则

- log变化的效应：将大的值缩小，将小的值放大

$$\log(100) = 4.6, \ \log(0.02) = -3.9$$

- 一般原则：相关变量必须全部都是正数时才能取log
- 通常情况下，如果某变量的所有值都是正数，而且其跨度很大时(也就是说，方差很大时)，取log可能是一个合适的选择
- 通常会取log的变量：GDP，工资，销售额，等(我们更关心这些变量的增长率而不是其绝对数值)
- 如果某变量既有正值，又有负值，而我们又非得考虑对于它的弹性怎么办？回归原始定义，并注意可能需要去掉它的极端值：

$$\frac{\triangle Y_i}{Y_i} = \beta_0 + \beta_1 \frac{\triangle X_i}{X_i} + u_i,$$

关于模型设定

模型设定是计量经济分析的首要问题，也是一个非常难的问题。

对于模型设定，一般来说，我们无法找到真实的模型是什么。我们只能尽可能找到和真实模型接近的模型。

一个一般形式的可加模型形如：

$$\psi(Y) = \phi(X, \beta) + u, \quad E[u|X] = 0, \tag{3}$$

其中$\psi(), \phi()$为给定函数。

当$\psi(Y) = Y$，$\phi()$既是X的线性函数，又是β的线性函数时，这就是最常见的线性模型：

$$Y_i = \beta_0 + \beta_1 X_i + u_i.$$

关于模型设定

一个稍一般的情形是，$\phi()$是X的非线性函数，但它是β的线性函数。这种时候我们仍然叫线性模型，如：

$$
\begin{aligned}
\log(Y_i) &= \beta_0 + \beta_1 X_i + u_i, \\
Y_i &= \beta_0 + \beta_1 \log(X_i) + u_i, \\
\log(Y_i) &= \beta_0 + \beta_1 \log(X_i) + u_i, \\
\sqrt{Y_i} &= \beta_0 + \beta_1 \frac{1}{X_i} + u_i,
\end{aligned}
$$

如果$\phi()$是β的非线性函数，那这种模型就叫非线性模型，如：

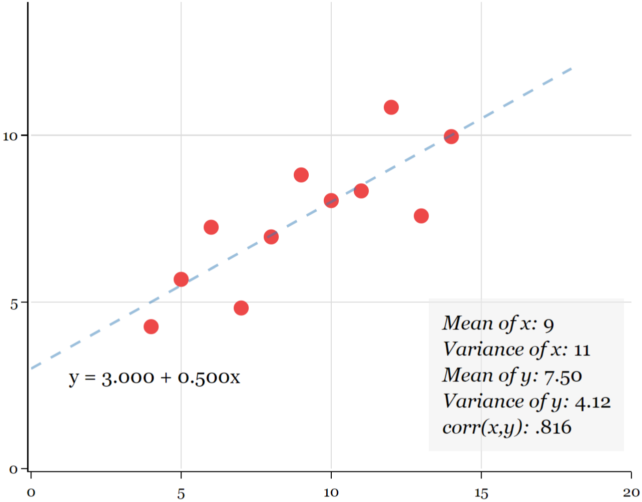
$$
\begin{aligned}
Y_i &= \frac{1}{\beta_0 + \beta_1 X_i} + u_i, \\
\log(Y_i) &= (\beta_0 + \beta_1 X_i)^2 + u_i,
\end{aligned}
$$

Anscombe 的回归例子

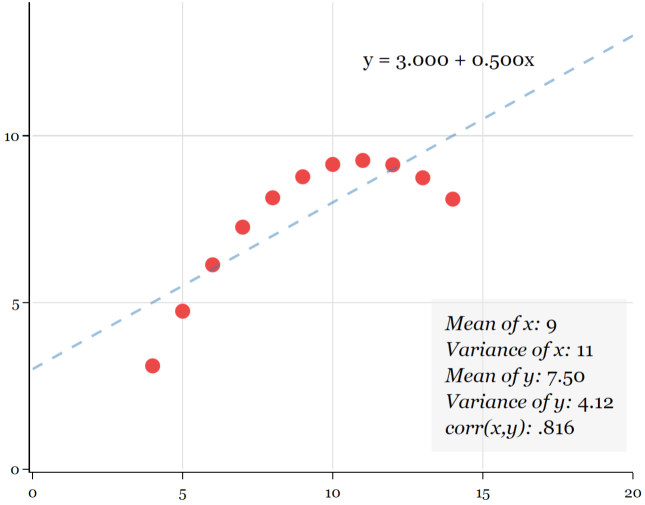
Anscombe (1973)给出了下面的一组回归数据：

X1	Y1	X2	Y2	X3	Y3	X4	Y4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.71
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.1	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.1	4	5.39	19	12.5
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

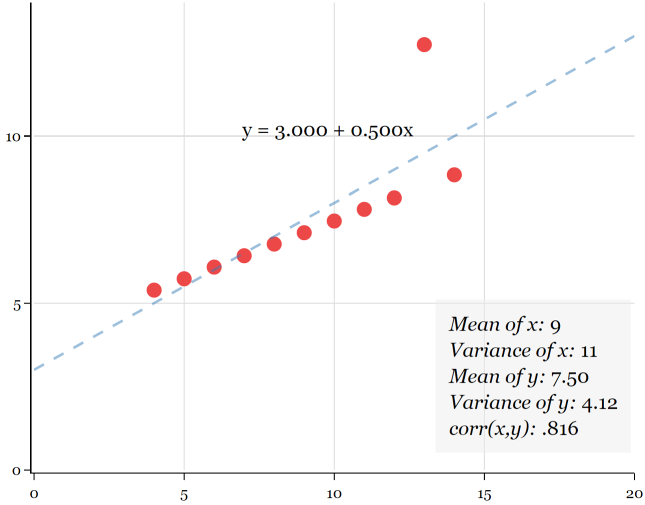
Anscombe 的回归例子: $Y_1 \sim X_1$



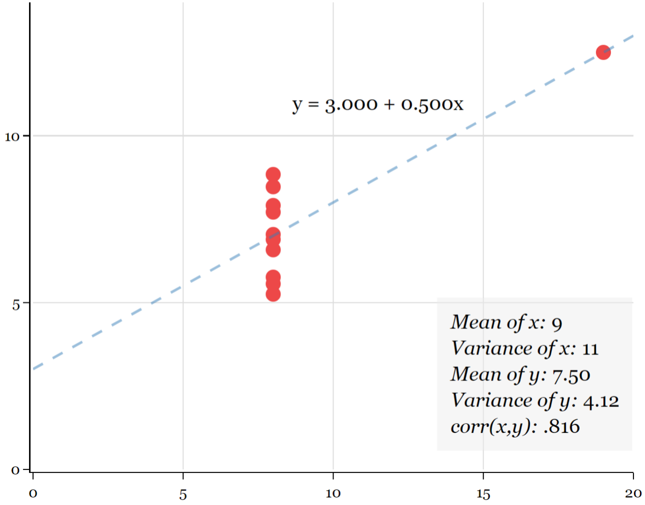
Anscombe 的回归例子: $Y_2 \sim X_2$



Anscombe 的回归例子: $Y_3 \sim X_3$



Anscombe 的回归例子：$Y_4 \sim X_4$



Anscombe 的回归例子: 讨论

四组不同的数据，得到的回归结果，包括各种统计量，都是一模一样的!

Number of observations (n) = 11
Mean of the x's $(\bar{x})$ = 9.0
Mean of the y's $(\bar{y})$ = 7.5
Regression coefficient (b_1) of y on x = 0.5
Equation of regression line: $y = 3 + 0.5\,x$
Sum of squares of $x - \bar{x}$ = 110.0
Regression sum of squares = 27.50 (1 d.f.)
Residual sum of squares of y = 13.75 (9 d.f.)
Estimated standard error of b_1 = 0.118
Multiple R^2 = 0.667

Anscombe 的回归例子：讨论

四组不同的数据，得到的回归结果，包括各种统计量，都是一模一样的！

系数[a]

group	模型		未标准化系数 B	未标准化系数 标准错误	标准化系数 Beta	t	显著性
1	1	(常量)	3.000	1.125		2.667	0.026
		x	0.500	0.118	0.816	4.241	0.002
2	1	(常量)	3.001	1.125		2.667	0.026
		x	0.500	0.118	0.816	4.239	0.002
3	1	(常量)	3.002	1.124		2.670	0.026
		x	0.500	0.118	0.816	4.239	0.002
4	1	(常量)	3.002	1.124		2.671	0.026
		x	0.500	0.118	0.817	4.243	0.002

a. 因变量：y

Anscombe 的回归例子: 讨论

但事实上，只有第一组数据$Y_1 \sim X_1$，直接进行线性回归才是最合理的选择。

第二组数据，$Y_2 \sim X_2$，明显存在曲线关系，也就是说，更好的模型可能是一个多项式形式。

第三组数据，$Y_3 \sim X_3$，在完美的线性关系之外，存在一个异常值。直接进行线性回归未必是最合理的选择。

第四组数据，回归直线的斜率完全由$x = 19$那一个观测值所决定。我们需要了解为什么$x = 19$的这个点这么重要。同时，我们需要确保这个观测值是可靠的。如果不可靠，那么整个回归分析结果的可靠性就存疑了。

启示，进行回归分析之前，要先画图看一看，这样可能会避免模型选择一开始就进入误区。

最小二乘估计量的合理性基础：Gauss-Markov假设

什么情况下最小二乘估计量才具有合理性质？

1. 假设SLR.1 (线性模型)：总体变量Y,X满足线性模型关系

$$Y = \beta_0 + \beta_1 X + u, \tag{4}$$

2. 假设SLR.2 (随机抽样)：样本数据$\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$为来自总体的简单随机样本
3. 假设SLR.3 (解释变量的样本数据必须有波动性)：所有X_i的值不是恒为常数
4. 假设SLR.4 (条件期望为零)：$E[u_i|X_i] = 0$
5. 假设SLR.5 (条件不相关性与同方差性)：

$$Var(u_i|X_i) = \sigma^2,\ \forall i;\ \ Cov(u_i, u_j|X_i, X_j) = 0, \forall i \neq j$$

关于Gauss-Markov假设

假设SLR.1-SLR.3相对比较容易理解。假设SLR.1关注的是线性模型的合理性；假设SLR.2关注的是数据的采样方式；假设SLR.3在实际情况下都是满足的。

假设SLR.4-SLR.5相对比较复杂。也是关键的假设。这两个都是关于u_i关于X_i的条件分布的特征：假设SLR.4关注的是这个条件分布的均值（一阶矩），假设SLR.5关注的是这个条件分布的方差（二阶矩）。

这两个假设后期我们都会进行适当放松。但无论如何放松，本质上涉及的都是这两点：一个是u_i与X_i的相关性，一个是（在考虑了X_i以后）u_i与u_j的相关性。

条件期望及其性质

设Y为一个随机变量，X为一个随机向量。给定$X=x$，Y的条件分布的期望$E[Y|X=x]$是x的函数。不妨记作

$$\phi(x) = E[Y|X=x] = \int y f_{y|x}(y|x)dy$$

我们称随机变量$E[Y|X]=\phi(X)$为Y关于X的条件期望。
条件期望的基本性质：

- 性质1：
$$E[E[Y|X]] = E[Y].$$
- 性质2：对任意函数h，
$$E[h(X)Y|X] = h(X)E[Y|X]$$
- 性质3：
$$E[h_1(X)Y_1 + h_2(X)Y_2|X] = h_1(X)E[Y_1|X] + h_2(X)E[Y_2|X]$$

条件方差及其性质

Y关于X的条件方差定义为：

$$Var(Y|X) = E\Big((Y - E[Y|X])^2|X\Big)$$

条件方差的基本性质：

- 性质1：全方差公式

$$Var(Y) = E[Var(Y|X)] + Var(E[Y|X])$$

- 性质2：

$$Var(h(X)Y|X) = h(X)^2 Var(Y|X)$$

- 性质3：

$$Var(h_1(X)Y_1 + h_2(X)Y_2|X) = h_1(X)^2 Var(Y_1|X)$$
$$+h_2(X)^2 Var(Y_2|X) + 2h_1(X)h_2(X)Cov(Y_1, Y_2|X)$$

其中

$$Cov(Y_1, Y_2|X) = E\Big((Y_1 - E[Y_1|X])(Y_2 - E[Y_2|X])|X\Big)$$

最小二乘估计量的无偏性

记$X = (X_1, \ldots, X_n)$。在上述假设条件下：

1. $\hat{\beta}_1$是β_1的无偏估计量

$$E[\hat{\beta}_1 | X] = \beta_1$$

因而

$$E[\hat{\beta}_1] = \beta_1$$

2. $\hat{\beta}_0$是β_0的无偏估计量

$$E[\hat{\beta}_0 | X] = \beta_0$$

因而

$$E[\hat{\beta}_0] = \beta_0$$

最小二乘估计量的方差

在上述假设条件下：

1. $\hat{\beta}_1$的条件方差与无条件方差分别为

$$Var[\hat{\beta}_1|X] = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2, \quad Var[\hat{\beta}_1] = E\left[\frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\sigma^2$$

2. $\hat{\beta}_0$的条件方差与无条件方差分别为

$$Var[\hat{\beta}_0|X] = \frac{\frac{1}{n}\sum_{i=1}^{n}X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sigma^2, \quad Var[\hat{\beta}_0] = E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]\sigma^2$$

σ^2的估计及性质

- σ^2的一个估计量如下：

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$$

- 定理2.3: s^2是σ^2的无偏估计量：

$$E[s^2|X] = \sigma^2$$

- σ的一个估计量为：

$$s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$$

- s不是σ的无偏估计量

$\hat{\beta}_0, \hat{\beta}_1$的标准差的估计量

- $\hat{\beta}_1$的标准差的估计量为：

$$se(\hat{\beta}_1) = \frac{s}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

- $\hat{\beta}_0$的标准差的估计量为：

$$se(\hat{\beta}_0) = s\sqrt{\frac{n^{-1}\sum_{i=1}^{n}X_i^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

注：以上标准差估计量是条件标准差（而不是无条件标准差）的估计量。为什么这样做是合理的？教材Theorem 3.2下面的文字进行了讨论。我们在大样本理论那一部分再回来讨论这个问题。

评论

简单线性回归模型的局限性：

- 现实中影响Y的变量有多个
- 简单线性模型没有控制住除X之外的变量对Y的影响
- 遗漏变量可能导致推断结果出现严重偏差

多变量线性回归模型：

- 引入多个影响Y的变量
- 其分析结果可以提供类似于“保持其他变量不变”情形下的解释