条件期望

&线性模型

[高中统计学·第柒课]

[何濯羽·2024年2月23日]

目录

条件期望函数是大多数回归模型的基石。

第一节 条件期望函数

- 条 件 期 望 的 定 义
- 期 望 迭 代 法 则
- 条 件 定 理
- 条 件 期 望 函 数 的 定 义

条件期望的定义

条件期望是什么？

假设随机变量 Y 在给定 $X = x$ 时的条件 PMF 或 条件 PDF 为 $f_{Y|X}(y|x)$。Y 在给定 $X = x$ 时的条件期望是

$$E(Y|X = x) = \sum_{x \in \mathcal{X}} y \cdot f_{Y|X}(y|x)$$

期望迭代法则

Law of Iterated Expectation (LIE)

如果 $E(|Y|) < \infty$，那么对于任意随机变量 X，以下等式恒成立：

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

这条法则说明了“条件期望的期望是非条件期望”。

条件定理

条件期望的性质之一是：$E(X|X) = X$，即 X 在给定其本身取值时的数学期望是其本身。

这一性质的第一条推论是

$$E[g(X)|X] = g(X)$$

其中，$g(\cdot)$ 是以随机变量 X 为自变量的任意函数。

这一性质的第二条推论就是“条件定理”（conditioning theorem）：当 $E(|Y|) < \infty$ 时，

$$E[g(X) \cdot Y|X] = g(X) \cdot E(Y|X)$$

当 $E(|Y|) < \infty$ 且 $E[|g(x)|] < \infty$ 时，

$$E[g(X) \cdot Y] = E[g(X) \cdot E(Y|X)]$$

条件期望是函数

Hán Shù

函数

这节课的重点：我们必须认识到

- 数学期望 $E(Y)$ 是一个定值；
- 条件期望 $E(Y|X)$ 是一个以 X 为自变量的函数！

条件期望是函数

例： 已知随机变量 X 和 Y 拥有如下的联合分布列，请计算 $E(X|Y=2)$ 。

		Y=1	Y=2	Y=3	边际概率
X	2	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{3}$
	3	$\frac{1}{3}$	0	$\frac{1}{6}$	$\frac{1}{2}$
	4	0	$\frac{1}{6}$	0	$\frac{1}{6}$
	边际概率	$\frac{5}{12}$	$\frac{1}{3}$	$\frac{1}{4}$	

先计算条件概率：

$$\Pr(Y=y|X=2)=\frac{\Pr(X=2,Y=y)}{\Pr(X=2)}=\begin{cases}\frac{1}{4} & (y=1)\\ \frac{1}{2} & (y=2)\\ \frac{1}{4} & (y=3)\end{cases}$$

再计算条件期望：

$$E(Y|X=2)=\sum_{y=1}^{3}y\cdot\Pr(Y=y|X=2)=1\times\frac{1}{4}+2\times\frac{1}{2}+3\times\frac{1}{4}=2$$

相似地，我们可以算出： $E(Y|X=3)=\frac{5}{3}$

$E(Y|X=4)=2$

显然，$E(Y|X=x)$ 是一个以 x 为自变量的函数。

条件期望函数

我们已知 $E(Y|X)$ 是一个以随机变量 X 为自变量的函数，可记为

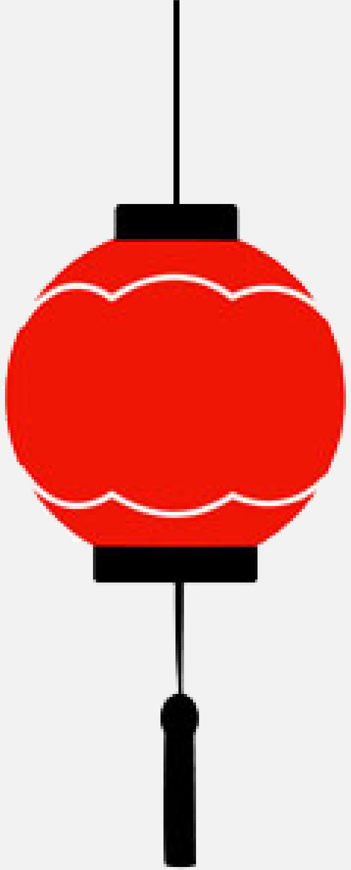
$$m(X) = E(Y|X)$$

这一函数被称为“条件期望函数”（conditional expectation function，简记为CEF）。这是一个含有重要信息的函数——当我们知道 CEF 时，我们就知道了对应每一个 X 取值的 Y 的期望值。

例如，X 为某位成年人的学位（0代表本科以下，1代表本科，2代表硕士，3代表博士及以上）而 Y 代表成年人的年收入。如果我们知道中国全体人口的 X 与 Y，我们就可以得到中国全体人口的 CEF。这样，当我们知道某人的学位时，通过 CEF，我们可以轻易地算出该人的期望年收入——这大概是无数应届毕业生都想知道的函数吧……为了实现“年入百万”的期望，一个人需要获得多高的学位?

现实是残酷的。

我们无法知晓这个 CEF！因为我们无法直接观测总体（即 X 与 Y 的联合概率分布），所以我们无从得知真实的条件期望函数 $E(Y|X)$。我们唯一可以尝试的事就是：抽样，然后估计 CEF ……



第二节 误差项

- 误差项的定义与性质
- 最优描述器
- 误差项的方差

误差项的定义

Wù Chā Xiàng

误 差 项

误差项（error term）被定义为“现实与期望的差距”，即

$$e = Y - E(Y|X) = Y - m(X)$$

误差项 e 的随机变量 Y 与它在给定 X 时的数学期望的差。注意：e 是一个随机变量!

通过移项，我们得到 Y 的表达式：$Y = E(Y|X) + e$。

误差项的性质

性质一： $E(e|X) = 0$

性质二： $E(e) = 0$

证明：

$$E(e|X) = E[Y - m(X)|X] = E(Y|X) - E[m(X)|X] = E(Y|X) - m(X) = 0$$

该等号为什么成立?

该等号为什么成立?

$$E(e) = E[E(e|X)] = E(0) = 0$$

注意：以上两个误差项 e 的性质是由概率论的定理推导而来的，不是假设！

分解性质

随机变量的“分解性质”（decomposition properties）是指任何一个随机变量 Y 都可以被表达为如下的分解形式：

$$Y = E(Y|X) + e$$

这一等式被称为“回归模型”（regression model）。其中，e 具有以下性质：

- $E(e|X) = 0$；
- $E(e) = 0$；
- $E[h(X) \cdot e] = 0$，对于任意满足 $E[|h(X) \cdot e|] < \infty$ 的函数 $h(\cdot)$ 恒成立。

换言之，随机变量的“分解性质”告诉我们，任一随机变量都可以被分解为两个部分：

1) 可以被 X 解释的部分；
2) 不可以被 X 解释的部分。

注： $E(e|X) = 0 = E(e)$ 说明了误差项 e 均值独立于 X，但不等同于“e 与 X 相互统计独立”。

最优描述器

假设我们已知一个随机变量 X，而我们想利用 X 构建一个函数 $g(X)$ 去描述 Y。请问最优的 $g(X)$ 应该是什么函数?

在回答这个问题之前，我们需要解答另一个问题——什么是“最优”？

自然地，$g(X)$ 与 Y 的差距，即 $Y-g(X)$，越小越好。然而，$g(X)$ 与 Y 均是随机变量，所以它们的差也是随机变量。于是，一种可行的思路是：我们希望找到一个 $g(X)$ 能使得 $Y-g(X)$ 的数学期望最小。

为了避免正差与负差相互抵消，也为了计算机的运算快捷，我们还可以给 $Y-g(X)$ 套上一个平方运算符。

因此，最优的 $g(X)$ 应该是一个能够将如下所示的式子最小化的函数：

$$E\{[Y-g(X)]^2\}$$

最优描述器

根据“分解性质”，随机变量 Y 可以被分解为

$$Y = m(X) + e$$

于是，

$$
\begin{aligned}
E\{[Y - g(X)]^2\} &= E\{[e + m(X) - g(X)]^2\} \\
&= E(e^2) + 2E\{e \cdot [m(X) - g(X)]\} + E\{[m(X) - g(X)]^2\} \\
&= E(e^2) + E\{[m(X) - g(X)]^2\} \\
&\geq E(e^2)
\end{aligned}
$$

当且仅当 $g(X) = m(X)$ 时等号成立

因此，最优的 $g(X)$ 正是 Y 的条件期望函数 $m(X) = E(Y|X)$。

误差项的方差

误差项的非条件方差（unconditional variance）是

$$D[Y - E(Y|X)] = D(e) = E\{[e - E(e)]^2\} = E(e^2)$$

我们通常将其记作 $D(e) = \sigma^2$。

σ^2 反映的是有多少 Y 的变化没有被它的条件期望 $E(Y|X)$ 解释。

显然，σ^2 的大小与给定的信息 X 相关。数学上可证：

$$D(Y) \geq D[Y - E(Y|X_1)] \geq D[Y - E(Y|X_1, X_2)]$$

即，误差项的方差随着给定信息的增加而呈现弱单调递减（weakly decreasing）的现象。

注：证明过程需要使用“琴生不等式”（Jensen's Inequality），在此不做展示。

误差项的方差

误差项的条件方差（conditional variance）是

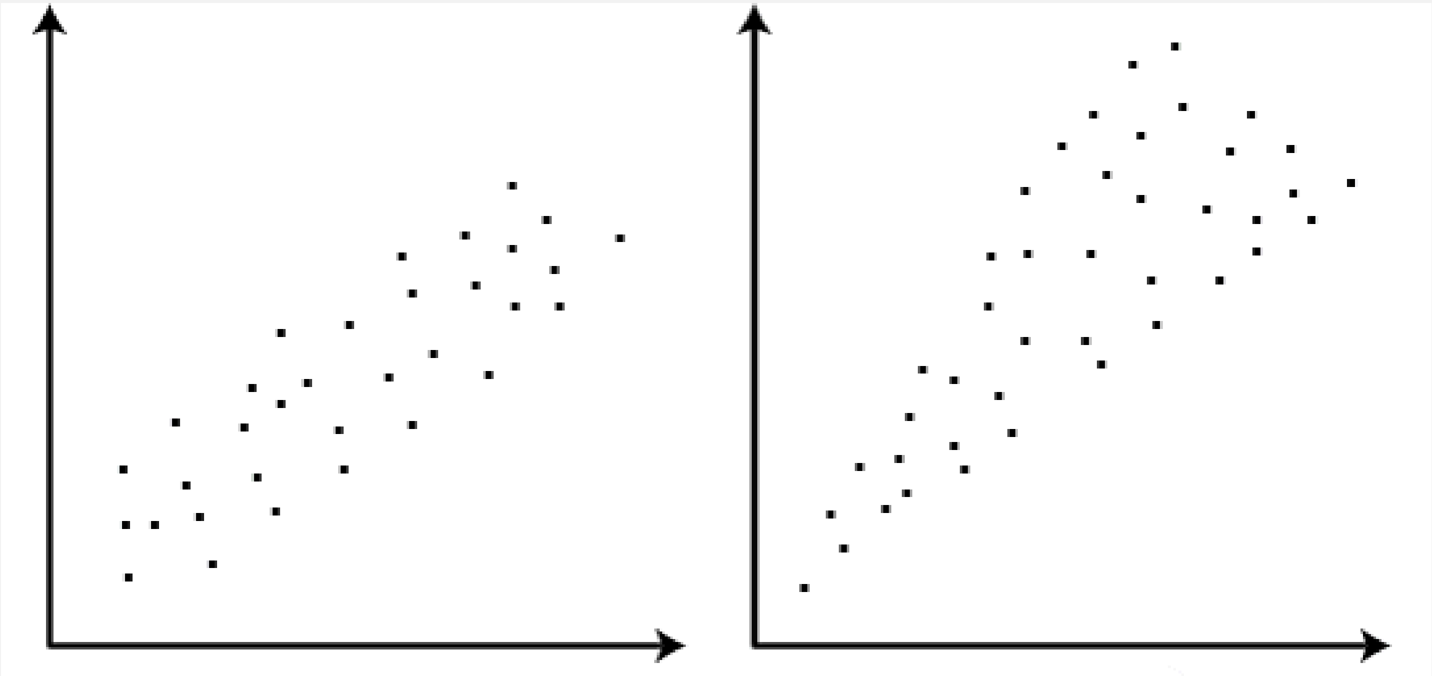
$$D(e|X) = E\{[e - E(e)]^2|X\} = E(e^2|X)$$

与条件期望相同，这也是一个以 X 为自变量的函数。我们通常将其记作 $\sigma^2(X)$。

当 $\sigma^2(X) = \sigma^2$，即误差项的条件方差不随 X 变化时，我们称“误差项是同质的（homoskedastic）”。
当误差项的条件方差 $\sigma^2(X)$ 随 X 变化时，我们称“误差项是异质的（heteroskedastic）”。

同质的误差项

异质的误差项



第三节 线性模型

- 线性模型的构造
- 模型系数的识别

线性条件期望函数

一般情况下，我们仅知晓 Y 的条件期望函数 $m(X) = E(Y|X)$ 是一个以 X 为自变量的函数，但无法知晓其具体的函数解析式。

在一些研究中，基于理论知识，或者出于计算简便的考虑，研究者们会“假设”条件期望函数是一个线性函数（linear function）：$m(X) = \alpha + \beta X$。这样，整个回归模型可以被写作

$$Y = m(X) + e = \alpha + \beta X + e$$

常数项　X 的系数

回归系数

这一等式被称为“线性回归模型”（linear regression model）。其中，Y 被称作因变量（dependent variable）、结果变量（outcome variable）、响应变量（response variable）、被回归量（regressand），X 被称作自变量（independent variable）、解释变量（explanatory variable）、预测量（predicator）或回归量（regressor）。

常见问题

问：在线性回归模型 $Y=\alpha+\beta X+e$ 中，我们为什么要添上一个常数项 α?

答：不在模型中添加 α 意味着我们的假设为 $m(X)=\beta X$。这一假设强于 $m(X)=\alpha+\beta X$，因为假设 $m(X)=\beta X$ 要求 $m(0)=0$。

在研究中，如果研究者们相信 $m(0)=0$ 是一定成立的，那么可以不在线性回归模型中添加常数项 α。然而，现实中，研究者们往往不容易证明 $m(0)=0$ 一定成立，所以往往会在模型中添上常数项。

问：在线性回归模型 $Y=\alpha+\beta X+e$ 中，$E(e|X)=0$ 和 $E(e)=0$ 依然成立吗?

答：根据“分解性质”，$Y=m(X)+e$，其中 $E(e|X)=0$ 和 $E(e)$ 恒成立。但是，在线性回归模型中，我们假设了 $m(X)=\alpha+\beta X$。因此，只有当假设 $m(X)=\alpha+\beta X$ 真实成立时，我们才有 $E(e|X)=0$ 和 $E(e)=0$。

回归系数的识别

在我们假设线性回归模型 $Y=\alpha+\beta X+e$ 成立之后，我们下一个问题就是：如何计算回归系数 α 和 β?
倘若我们知晓 X 与 Y 的联合概率分布，我们能否用 X 与 Y 的矩表示 α 和 β 呢?

根据“迭代期望法则”，我们有

$$E(Y)=E[E(Y|X)]=E(\alpha+\beta X)=\alpha+\beta\cdot E(X)$$

$$E(XY)=E[E(XY|X)]=E[X\cdot E(Y|X)]=E[X\cdot(\alpha+\beta X)]=\alpha\cdot E(X)+\beta\cdot E(X^2)$$

于是，

$$\begin{aligned}Cov(X,Y)&=E(XY)-E(X)\cdot E(Y)\\&=\alpha\cdot E(X)+\beta\cdot E(X^2)-E(X)\cdot[\alpha+\beta\cdot E(X)]\\&=\beta\cdot E(X^2)-\beta\cdot[E(X)]^2\\&=\beta\cdot D(X)\end{aligned}$$


因此，

$$\beta=\frac{Cov(X,Y)}{D(X)}\quad\Longrightarrow\quad\alpha=E(Y)-\beta\cdot E(X)=E(Y)-\frac{Cov(X,Y)}{D(X)}\cdot E(X)$$

回归系数的识别

综上，我们发现在线性回归模型 $Y=\alpha+\beta X+e$ 中，

$$\beta=\frac{Cov(X,Y)}{D(X)}$$

$$\alpha=E(Y)-\beta\cdot E(X)=E(Y)-\frac{Cov(X,Y)}{D(X)}\cdot E(X)$$

这种用 X 和 Y 的矩表示回归模型系数的操作叫做“识别”（ identification ）。
在识别回归系数之后，我们是否完成了所有任务呢？显然没有！在现实中，总体是不可直接观测的，这意味着 $E(X)$、$E(Y)$、$D(X)$、$Cov(X,Y)$ 这些矩虽然存在，但都不可直接观测！
于是，我们只能通过抽样与估计去推断 α 和 β 的数值了。

那么，根据我们在第陆课所学的知识，我们应该如何构建 α 和 β 的估计量呢？
这些估计量是无偏的、有效的吗？
这将是我们下一节课的核心问题。

线性模型的解释

在本节课结束之前，不得不强调的一点：在没有额外假设或改进的情况下，前文介绍的线性模型仅是一个描述模型（descriptive model）——描述样本中自变量 X 与因变量 Y 的关系（association）。

例如，假设我们发现真实的线性模型是 $Y = 2 + 3X + e$，即 $\alpha = 2$，$\beta = 3$，那么我们可以说：

- 在 $X = 0$ 时，Y 的期望值为 2。
- 随着 X 增加/降低 1 个单位，Y 的期望值增加/降低 3 个单位。

谨记：

- 描述模型没有因果推断的能力——我们不能说 X 的变化造成了 Y 的变化。
- 描述模型没有预测未来的能力——我们不能凭借 X 的取值预测 Y 的取值。

谢谢聆听

[T H A N K S]