

MAT220 Lecture 19: Linear Regression (线性回归)

Contents

1	线性回归	2
1.1	相关系数	2
1.2	例: 距离与通勤时间	2
2	最小二乘法	3
2.1	作为算法的回归	3
2.2	最小二乘准则	4
2.3	回归直线的公式	6
2.4	影响点	7
2.5	例: 计算回归直线	7
3	使用回归直线进行预测	8
3.1	进行预测	8
3.2	插值与外推	9
3.3	决定系数	10
3.4	决定系数的其他形式	12
3.5	作为 r^2 的决定系数	12
3.6	例: 通勤数据的决定系数	13

在第18讲中, 我们通过在散点图上绘制观测值 (x_i, y_i) , 研究了成对定量数据。随后, 我们使用样本相关系数 r 来描述线性关系的方向和强度。相关性本身只能告诉我们各点是否大致符合一条直线, 但它并不会实际构造出这条直线。今天, 我们将构造这条直线, 它被称为最小二乘回归直线。我们首先确定回归直线的主要特征, 然后使用它进行预测。最后, 我们将引入决定系数, 它用于衡量回归直线作为预测模型的质量。

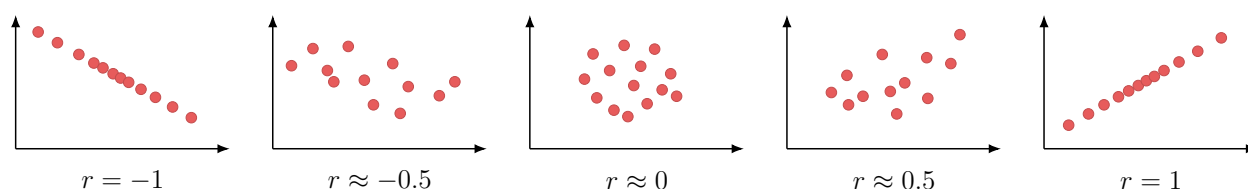
1 线性回归

1.1 相关系数

假设我们有 n 组成对观测值:

$$(x_1, y_1), \quad (x_2, y_2), \quad \dots, \quad (x_n, y_n).$$

在上一讲中，我们开始研究成对数据的描述统计。特别地，我们研究了线性相关，它描述成对数据在多大程度上沿着一条直线分布。在此过程中，我们引入了相关系数，它是一个用于量化 x 与 y 之间线性相关程度的数。我们用 r 表示这个数，它的取值范围是 -1 到 1 。 r 的值决定了线性相关的强度和方向：



从代数角度看，样本相关系数可以使用以下公式计算：

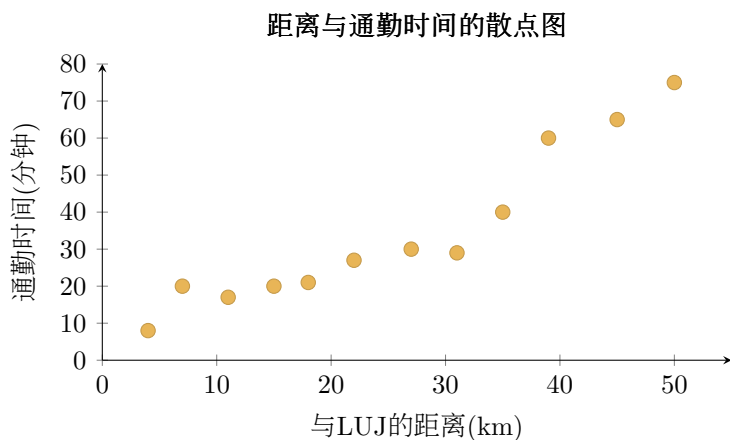
$$r = \frac{\text{cov}(x, y)}{s_x s_y},$$

其中，分子是协方差，它衡量两个变量之间的相互关联程度； s_x 和 s_y 分别是 x 方向和 y 方向上的样本标准差。

1.2 例：距离与通勤时间

在本讲中，我们将持续使用以下数据集作为例子。假设12名学生记录了自己与LUJ 之间的距离以及前往校园的通勤时间。

学生	距离 (km)	时间 (分钟)
1	4	8
2	7	20
3	11	17
4	15	20
5	18	21
6	22	27
7	27	30
8	31	29
9	35	40
10	39	60
11	45	65
12	50	75



对于这些数据，我们在上一讲中已经得到：

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad s_y \approx 21.253, \quad \text{以及} \quad \text{cov}(x, y) \approx 301.606.$$

因此，

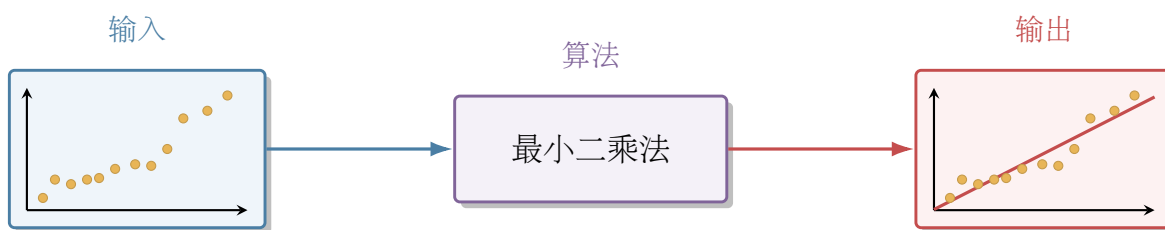
$$r = \frac{301.606}{(15.011)(21.253)} \approx 0.9454,$$

这表明两个变量之间存在很强的正线性关系。

2 最小二乘法

2.1 作为算法的回归

算法是一种固定的过程，它接收输入，并按照可靠的步骤产生输出。在今天的讲座中，我们将介绍一种算法，它以成对数据作为输入，并输出一条“最佳拟合直线”。由于稍后会清楚的原因，我们将这种方法称为最小二乘法。从示意图上看，它可以表示为：



回忆一下，任何非垂直直线都可以用两个数完全描述：

- 斜率，它描述直线的变化率，也就是直线的陡峭程度；
- y 轴截距，即当 $x = 0$ 时直线的值。

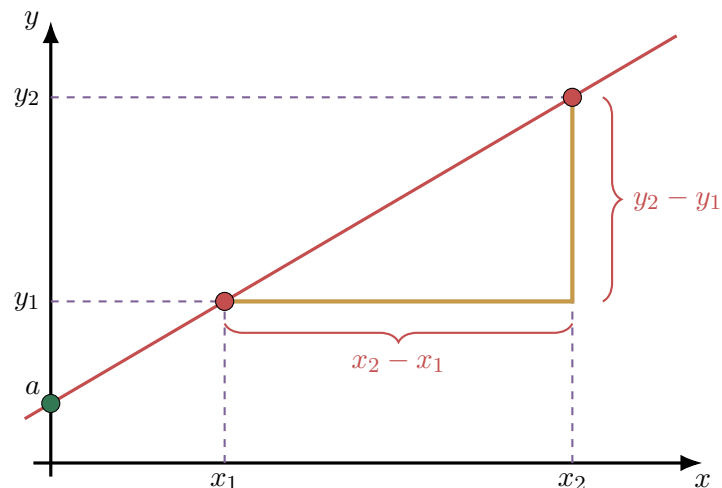
我们使用以下记号表示直线：

$$\hat{y} = a + bx,$$

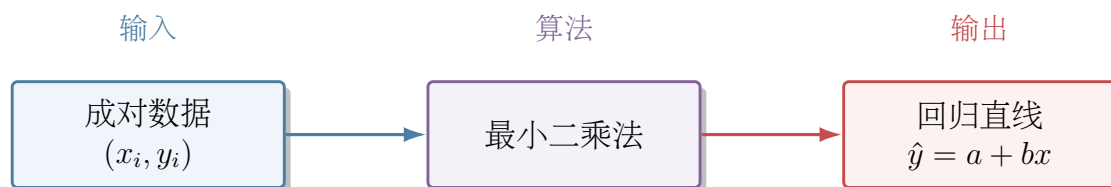
y 上方帽子符号的含义稍后会变得清楚。这里， a 是 y 轴截距， b 是直线的斜率。可以选取直线上的两个点，并使用以下公式计算斜率：

$$b = \frac{y_2 - y_1}{x_2 - x_1},$$

也就是说，用垂直方向的变化量除以水平方向的变化量。

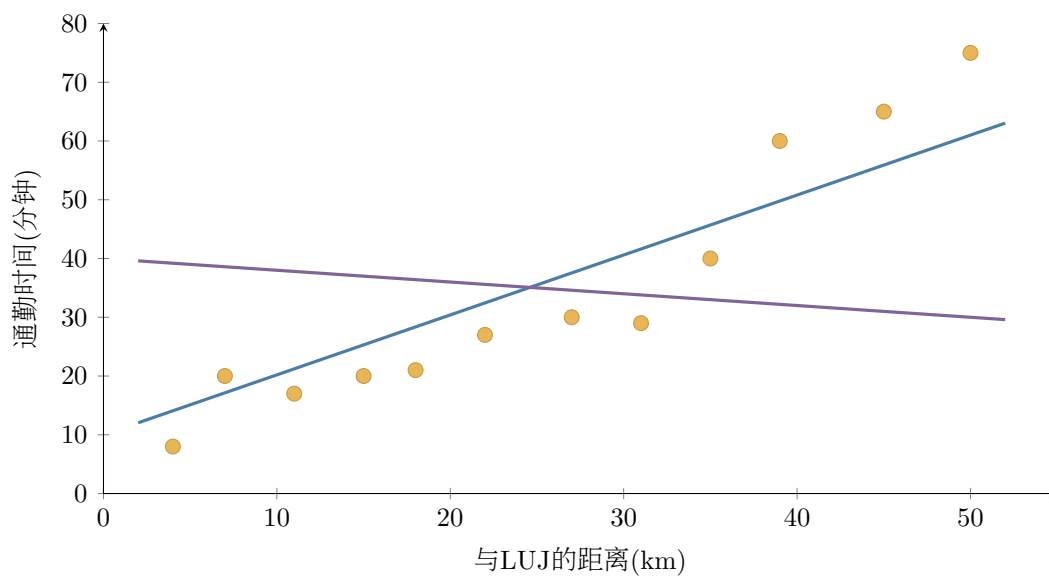


使用这一记号后，最小二乘法的输出实际上是一对数 a 和 b ，使所得直线能够最好地拟合输入数据。这样的直线称为回归直线。因此，更详细的算法示意图可以表示为：



2.2 最小二乘准则

使用第1节中的数据，考虑以下两条直线：

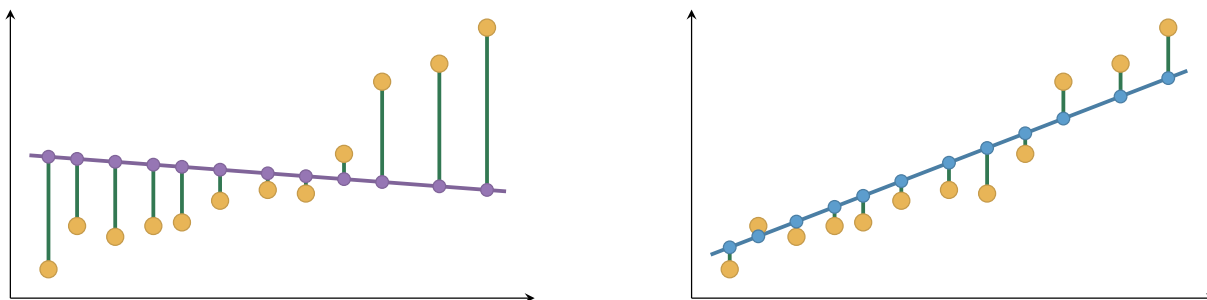


显然，蓝色直线比紫色直线更好地拟合了数据。但是，从数学角度来看，为什么会这样？换句话说，蓝色直线的哪些性质使它优于紫色直线？一种回答方法是观察数据点与回归直线之间的垂直

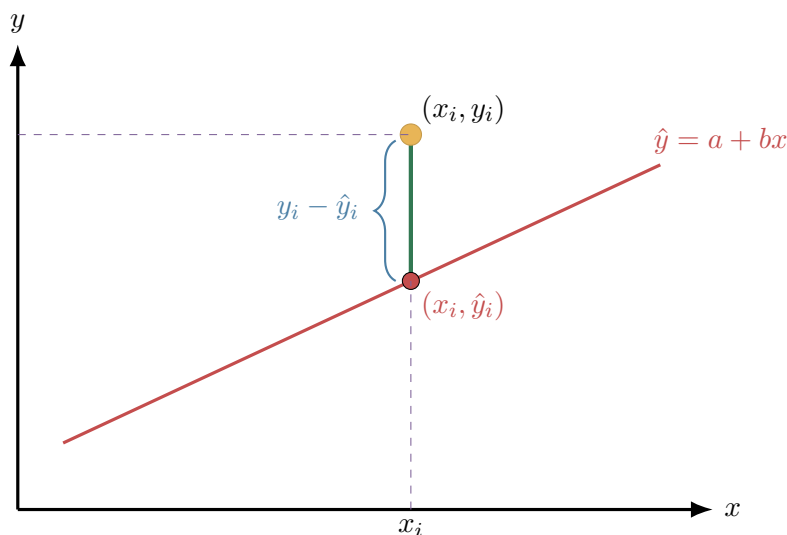
距离。回忆一下，每个数据点都可以写成 (x_i, y_i) 。因此，为了描述每个点到直线的垂直距离，我们可以将 x_i 代入回归直线 $\hat{y}$ 的方程，从而求出直线上的对应输出值。在 x_i 上方的回归直线值为：

$$\hat{y}_i = a + bx_i.$$

将 x_i 和 $\hat{y}_i$ 组合起来，就得到了位于直线上的点 $(x_i, \hat{y}_i)$ ，它位于 x_i 的正上方。理想情况下，这个点应该接近观测数据点 (x_i, y_i) 。从图中可以看出，对于紫色直线，这些到真实数据的垂直距离远大于蓝色直线对应的距离：



正是在这个意义上，我们直观地认为蓝色直线的拟合效果更好：蓝色直线对应的所有垂直线段长度之和小于紫色直线对应的总长度。一般来说，这些带符号的差称为残差，它们就是差值 $y_i - \hat{y}_i$ ：



当 y_i 位于直线上方或下方时， $y_i - \hat{y}_i$ 分别为正值或负值。因此，如果我们直接把所有残差相加，试图构造某种“总误差”，这些正负残差可能会相互抵消，从而破坏我们对“拟合优度”的度量。为了解决这个问题，我们先将所有 $y_i - \hat{y}_i$ 平方，然后再把它们相加。¹因此，正确的“总误差”度量实际上是总平方误差：

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + bx_i))^2.$$

¹这与在样本方差 s_y^2 的公式中对 $y_i - \bar{y}$ 进行平方的原因相同。

不同的 a 和 b 会产生不同的直线，也会产生不同的误差和。在上面的例子中，紫色直线的总平方误差远大于蓝色直线的总平方误差。因此，我们可以将“最佳拟合直线”定义为使总平方误差最小的直线。

最小二乘准则

最小二乘回归直线 $\hat{y} = a + bx$ 的选择方式是，使得

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2$$

在所有直线中尽可能小。

2.3 回归直线的公式

a 和 b 的精确推导需要使用微积分。由于这是一门入门课程，我们不会在这里花时间推导回归直线。² 在这里，我们直接接受 a 和 b 的公式为：

$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{以及} \quad a = \bar{y} - b\bar{x}.$$

利用这两个公式，我们可以写出回归直线 $\hat{y} = a + bx$ 。

计算回归直线

给定成对数据 (x_i, y_i) ，计算

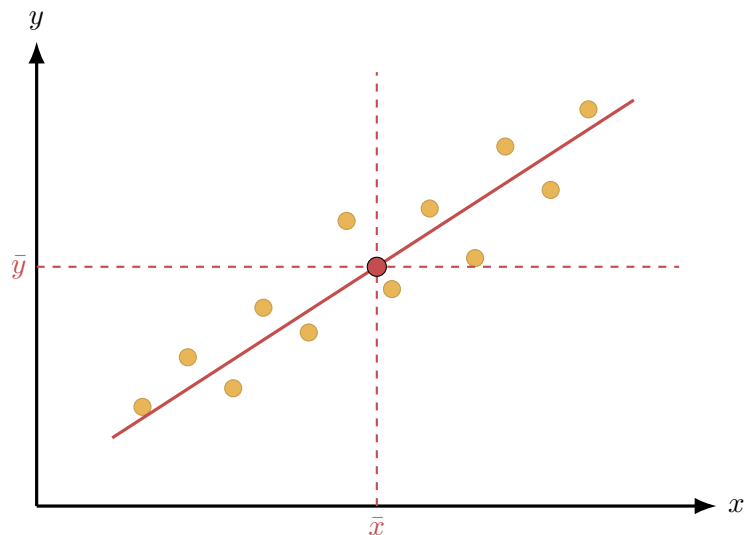
$$b = \frac{\text{cov}(x, y)}{s_x^2} \quad \text{以及} \quad a = \bar{y} - b\bar{x}.$$

然后将最小二乘回归直线写为

$$\hat{y} = a + bx.$$

作为一个简单推论，我们可以看到，点 $(\bar{x}, \bar{y})$ 始终位于最小二乘回归直线上：

²对于感兴趣的同学：我们可以对表达式 $\sum_{i=1}^n (y_i - (a + bx_i))^2$ 分别关于 a 和 b 求导。为了求最小值，令这两个导数等于零，然后求解所得的两个线性方程。



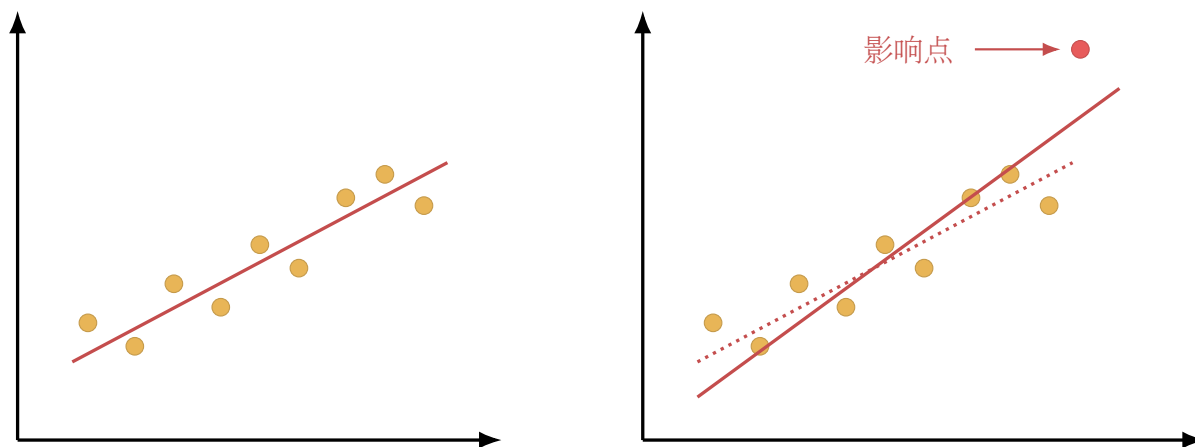
为了验证这一点，可以将 $x = \bar{x}$ 代入方程：

$$\hat{y} = a + b\bar{x} = (\bar{y} - b\bar{x}) + b\bar{x} = \bar{y}.$$

无论从哪一组数据开始，这一结果都不变。它只是回归直线构造方式的一个直接结果。

2.4 影响点

如果某个点远离数据的主要点群，尤其是在水平方向上距离较远，它可能会对回归直线产生很大影响。这样的观测值称为影响点。



影响点本身不一定是一个问题。但是，由于它可能大幅改变回归直线，影响点的存在可能会影响我们使用回归直线进行的预测。因此，在处理这类点时，我们应始终保持谨慎。

2.5 例：计算回归直线

现在计算第1节通勤数据的最小二乘回归直线。从第18讲可知：

$$\bar{x} \approx 25.333, \quad \bar{y} \approx 34.333, \quad s_x \approx 15.011, \quad \text{cov}(x, y) \approx 301.606.$$

首先计算斜率 b :

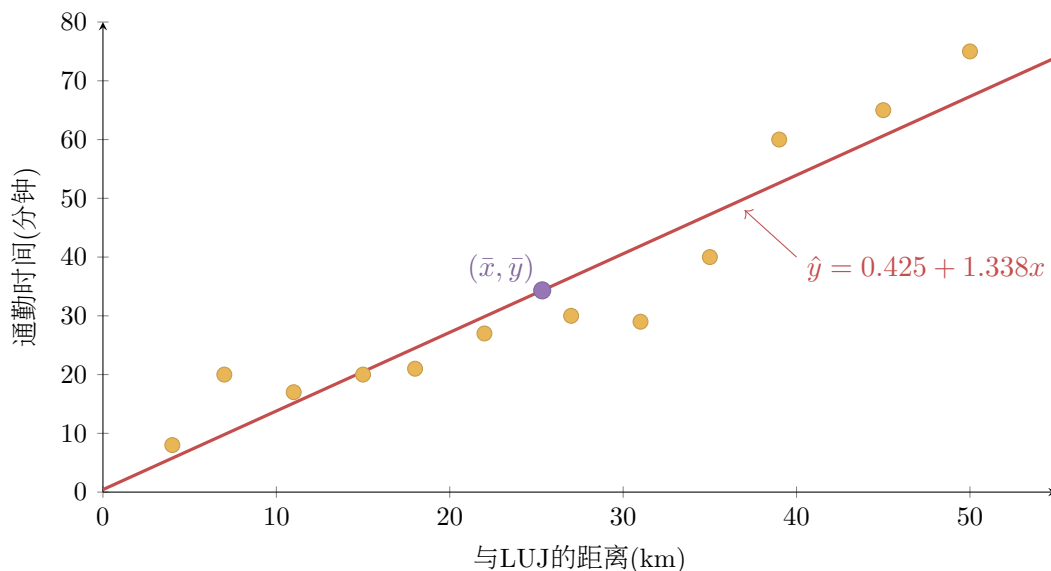
$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{301.606}{(15.011)^2} \approx 1.33849.$$

然后使用这个值计算 y 轴截距 a :

$$a = \bar{y} - b\bar{x} = 34.333 - (1.33849)(25.333) \approx 0.425.$$

将它们组合起来, 回归直线为:

$$\hat{y} = 0.425 + 1.338x.$$



斜率的解释

斜率约为1.338。根据拟合的线性模型, 与LUJ 的距离每增加1公里, 预测通勤时间大约增加1.34 分钟。

截距约为0.425 分钟。在这个例子中, 截距主要是为了将直线放置在正确位置而需要的理论值, 不应自动赋予它现实意义。

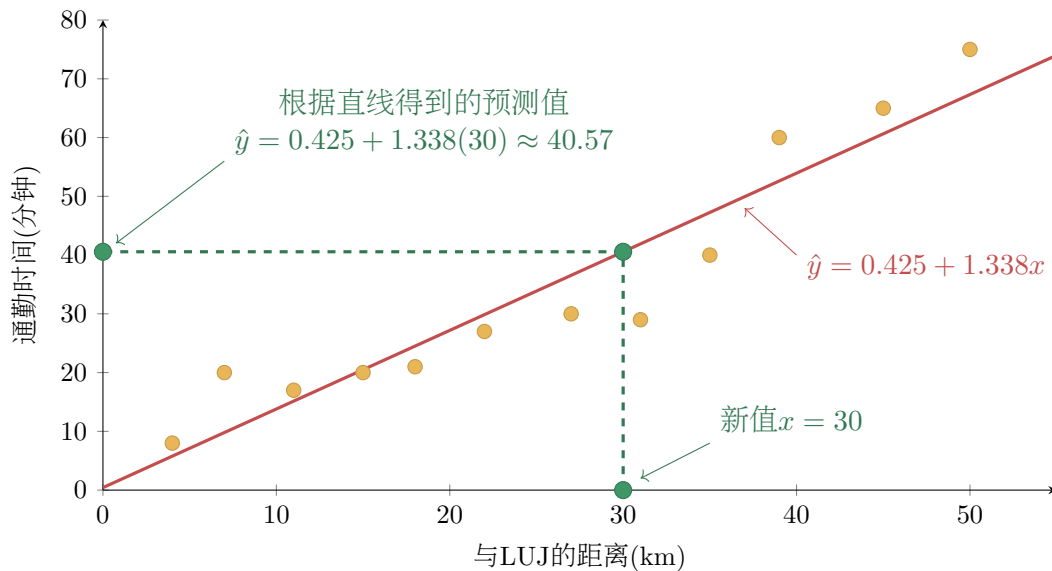
3 使用回归直线进行预测

3.1 进行预测

除了提供直观总结之外, 回归直线的一个重要优点是, 它可以作为模型帮助我们进行预测。继续使用通勤时间和距离的例子, 假设一名学生搬家后, 新公寓距离校园30 km。在这种情况下, 可以将 $x = 30$ 代入回归直线方程, 预测大致的通勤时间:

$$\hat{y} = 0.425 + 1.338(30) \approx 40.57.$$

因此, 该模型预测通勤时间约为41 分钟。从图形上看, 这一预测如下所示:

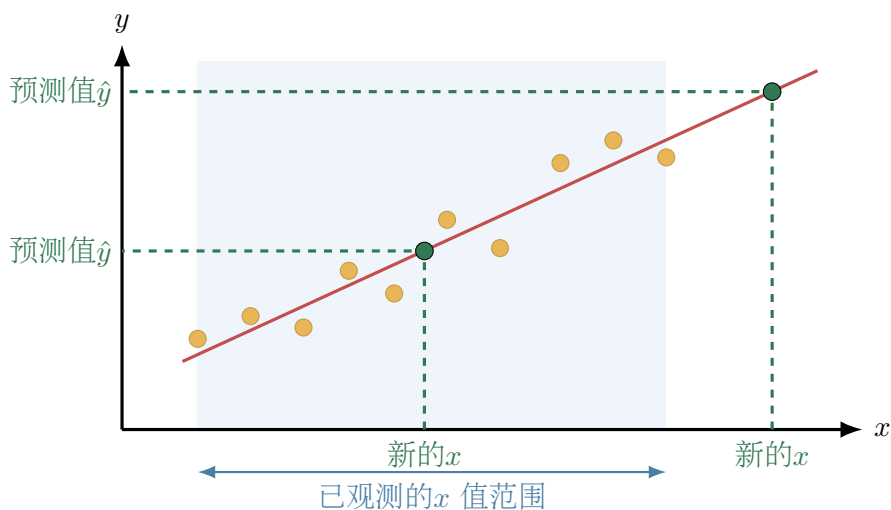


3.2 插值与外推

预测有两种本质上不同的类型，称为插值和外推。

插值与外推

插值是根据位于已观测 x 值范围内的新值进行预测，而**外推**是根据位于已观测 x 值范围之外的新值进行预测。



一般来说，插值通常是使用回归直线时更安全的方法，因为模型被应用在实际观测过数据的区域内。相比之下，外推存在将模型应用于不适当范围之外的风险。当 x 取极端值时，可能有其他变量影响数据，而这些因素并未被回归直线本身考虑。

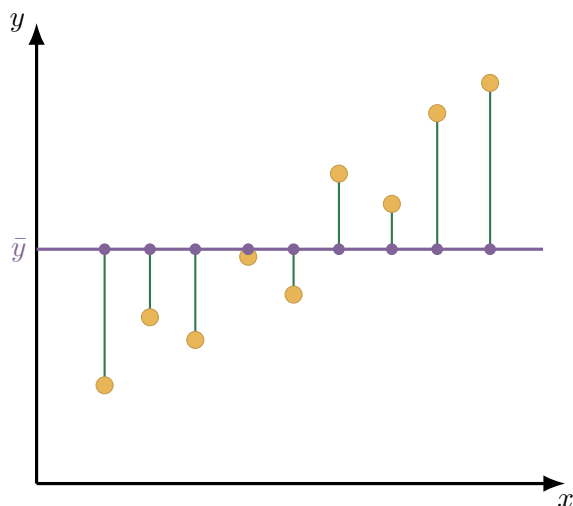
关于外推的警告

在某个 x 值范围内观察到的线性模式，未必会在该范围之外继续成立。因此，即使原始相关性很强，外推预测也可能不可靠。

3.3 决定系数

回归直线概括了散点图，并允许我们根据新的 x 值预测 y 。现在，我们将引入一个数，用于量化这种拟合的好坏。为此，首先需要引入一个基准预测方法。

假设我们完全忽略 x 的数据。在这种情况下，对任意 y 值进行一般预测时，最好的选择是使用样本均值 $\bar{y}$ 。当然，这种预测在大多数时候并不完全正确，但从平均意义上看，这是我们能做到的最佳选择。使用均值 $\bar{y}$ 预测每一个 y 值，相当于选择一条水平直线 $y = \bar{y}$ ：

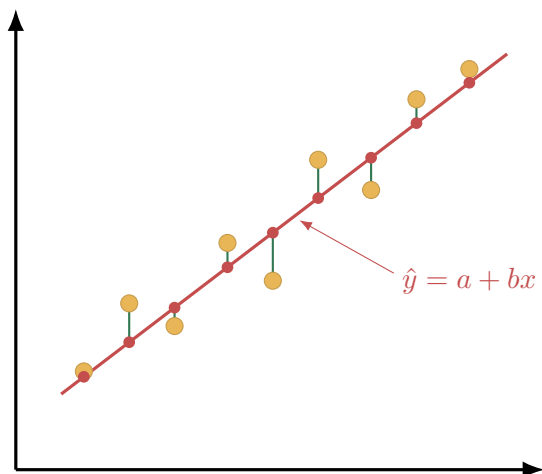


我们可以通过观察直线与观测数据点之间产生的总误差，衡量这种预测方法的好坏。同样，为了避免正负误差意外抵消，我们先将误差平方。这种简单预测产生的总平方误差为：

$$\sum_{i=1}^n (y_i - \bar{y})^2.$$

这个量称为总平方误差，并作为我们的基准。

现在，我们将回归直线的误差与简单预测的误差进行比较，从而衡量回归直线的预测效果。假设我们不再对每一个值都简单地预测为 $\bar{y}$ ，而是将成对输入 x_i 与回归直线一起使用。这样仍然会产生一些误差：



使用回归直线进行预测时得到的总平方误差为：

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

根据最小二乘准则的构造，这正是被最小化的量。因此，始终有

$$\sum_{i=1}^n (y_i - \bar{y})^2 \geq \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

为了比较这两个数的相对大小，可以取它们的比值：

$$\frac{\text{使用回归直线得到的总平方误差}}{\text{使用样本均值得到的总平方误差}} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

如果回归直线是一个好的预测模型，那么与使用样本均值得到的误差相比，它应产生较小的误差。因此，我们期望上述比值较小。相反，如果回归直线不是一个好的预测模型，它产生的误差可能与简单使用均值时一样大，这意味着上述比值接近1。如果我们希望这个数在预测效果好时较大，在预测效果差时较小，就可以用1 减去这个比值：

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

这被称为决定系数。

决定系数

决定系数为

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

它等于1 减去回归直线产生的误差与使用样本均值进行简单预测所产生误差的比值。当该数有定义时，它位于0 和1 之间；数值越大，表示回归模型的拟合效果越好。

3.4 决定系数的其他形式

我们可以通过重新排列决定系数，找到描述这个数的其他方式。一种方法是在表达式 $y_i - \bar{y}$ 中插入 $\hat{y}_i - \hat{y}_i$ ：

$$y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}).$$

回归直线具有以下有用性质：

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n ((y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}))^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2.$$

通常，我们会预期出现形如 $2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$ 的交叉项，但由于构造回归直线时所使用的微积分性质，这些交叉项在求和后恰好为零。通过基本代数运算，可以得到：

$$\begin{aligned} 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \frac{\sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}. \end{aligned}$$

将分子和分母同时除以 $n - 1$ ，得到：

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

这个表达式的分母是 y 的方差，也就是成对数据在垂直方向上的离散程度。分子计算的是预测值与 $\bar{y}$ 之间平方距离的一种平均。这同样是一种方差，但这次计算的是预测值 $\hat{y}_i$ 围绕 $\bar{y}$ 的离散程度。因此，我们可以用这种方式理解决定系数。

决定系数的另一种表达形式

决定系数也可以等价地写成

$$1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

因此，可以将决定系数理解为： y 的方差中有多少比例可以由回归直线 $\hat{y}$ 解释。

3.5 作为 r^2 的决定系数

再次考虑决定系数的另一种形式。为了便于计算，在分子和分母中分别加入 $\frac{1}{n-1}$ ：

$$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

因此，分母变为 y 的样本方差 s_y^2 。由于回归直线为 $\hat{y}_i = a + bx_i$ ，且 $a = \bar{y} - b\bar{x}$ ，所以：

$$\hat{y}_i - \bar{y} = a + bx_i - \bar{y} = (\bar{y} - b\bar{x}) + bx_i - \bar{y} = b(x_i - \bar{x}).$$

因此，决定系数的分子变为：

$$\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 = \frac{1}{n-1} \sum_{i=1}^n (b(x_i - \bar{x}))^2 = b^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 s_x^2.$$

将这些结果组合起来，得到：

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = \frac{b^2 s_x^2}{s_y^2}.$$

使用 b 的定义，有：

$$b = \frac{\text{cov}(x, y)}{s_x^2} = \frac{\text{cov}(x, y)}{s_x^2} \cdot \frac{s_y^2}{s_y^2} = \frac{\text{cov}(x, y)}{s_x s_y} \cdot \frac{s_y^2}{s_x s_y} = r \cdot \frac{s_y}{s_x}.$$

因此，决定系数可以大幅简化为：

$$\frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} = b^2 s_x^2 \div s_y^2 = \frac{\left(r \frac{s_y}{s_x}\right)^2 s_x^2}{s_y^2} = r^2 \frac{s_y^2}{s_x^2} s_x^2 \div s_y^2 = r^2 s_y^2 \div s_y^2 = r^2.$$

事实上，对于只有一个解释变量并包含截距的简单线性回归，决定系数恰好等于样本相关系数的平方：

$$r^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

这会产生以下直接结论：

- 将 r 平方后，它的正负号消失；
- 强度相同的正相关和负相关产生相同的 r^2 ；
- 较强的线性相关产生较大的决定系数；
- 较弱的线性相关产生较小的决定系数。

3.6 例：通勤数据的决定系数

对于通勤数据，我们之前计算得到 $r \approx 0.9454$ 。因此，

$$r^2 \approx (0.9454)^2 \approx 0.894.$$

因此，在该样本中，观测到的通勤时间变异中约有89.4%可以由与LUJ距离的线性回归解释。未被解释的变异可能反映了其他影响通勤时间的变量，例如交通方式、交通状况、换乘、步行时间、测量误差等。