单元 6 统计回归模型

欢迎回到物种分布模型的在线课程。在以前的单元中我们学习了只需要提供物种分布记录来预测物种的分布。在这个单元中，我们重点介绍提供分布有/无数据的统计回归模型中。

我们在单元 3 中介绍，分布无数据可能是真的分布无数据，或者没有分布无数据，你可以自己生成分布无数据，这被称为分布无数据。分布有/无 数据算法比较了物种分布位点和无分布位点的环境条件。

本单元中介绍的统计回归模型使用全部数据来预测变量系数，构建了一个最优函数来描述环境变量对物种分布的影响。一个特定模型的适用性取决于具体的模型假设。在这个单元中，我会讲一下三个统计回归模型的背景：广义线性模型，广义可加模型和多元自适应回归样条法。这些模型是“简单”线性回归模型的扩展。

线性回归模型有几个假设，如我们可以用一条直线来描述响应变量与预测变量之间的关系。这个图表示响应变量随预测变量变化而变化。例如，蟋蟀的鸣声随着温度升高而增加。在一个线性回归模型中假设温度增加 1 个单位会让鸣声增加 1 个单位。这种假设常常与物种分布模型不符，特别是当响应变量是概率时，即物种分布在给定位点的可能性。在这种情况下，可以使用灵活选取响应变量的广义线性模型。

广义线性模型（GLM）

二项分布的广义线性模型，即需要使用物种的分布有/无数据，称为逻辑回归。如果我们在这些数据点上画一条线，就会像一个 S 形，这条线表示物种分布概率。所以，如果你所有的物种位点都无分布，则概率是 0，而如果你所有点位都有分布，则概率为 1。在物种从无分布转换到有分布的环境变量处，对应的概率为 0.5。这意味着在该环境变量值中物种有同样的分布位点和无分布位点。当然，在物种分布模型中，我们并不只看一个环境变量的影响，而要看多个不同的环境变量。在广义线性模型中，综合输入该模型的所有预测因子来考量环境的适宜性，即线性预测。

因为物种分布的概率是基于二项分布数据，而不是正态分布数据，我们需要一个函数来关联响应变量和线性预测变量。这即是所谓的“连接”函数，对于二项分布数据来说是一个对数函数。对数函数中采用比率的对数，这是分布概率和无分布概率的比。

此函数可以转换 y 变量使之能够适于与响应变量和线性变化的直线关系。

这个公式描述比率的对数函数，且有一个不变的基线，它是指所有环境变量和其平均值的比率对数函数，即直线的 y 轴截距。然后，每个环境变量乘以其系数得到每个环境变量的影响值，其会影响线的斜率。

广义加法模型（GAM）

我要解释的下一个统计回归算法是广义可加模型，即 GAM。广义可加模型是广义线性模型的扩展，并和这些类型的模型有相同的特征：线性预测，用输入模型的全部环境变量来进行综合环境适宜性评分，连接函数将响应转换成比率的对数函数。GLMs 建立响应和预测因子的线性函数，GAM 考虑的形式可能不是完全直线，而是更复杂的形式。为了适应这个复杂关系，用平滑函数作为线性预测中预测变量的

系数。对于模型中的每个环境变量，GAM 算法尽可能地计算出最适合的平滑函数。在最后的模型中每个变量的平滑函数进行相加。由于 GAMs 是相加的方法，因此难以包含预测因子之间的相互作用，因此不常用。有许多不同的平滑函数可以用，但在物种分布模型中使用最广泛的是三次方平滑样条法。

多变量自适应回归样条（MARS）

我在本单元中讲的最后一个统计回归算法是多变量自适应回归样条算法。虽然我们将该算法归为统计回归模型，它也有和机器学习模型相似的特征。多变量自适应回归样条算法是线性模型的另一个扩展形式，它非常强大，能够建立响应变量和预测变量之间的复杂关系。

我将用一个物种在一个环境变量中的分布概率分布图作为例子，来解释这个算法是如何运作的。很明显这些点反映出来预测因子与响应变量之间不是线性关系。多变量自适应回归样条算法将预测值的范围分为几组，针对每个组建立单独的线性回归模型，每个模型都有自己的斜率和相应的误差。将单独的线连接起来，这些连接点被称为节点。多变量自适应回归样条算法自动搜索最佳的节点位置。每个节点都有一对基函数。这些基函数描述了环境变量与响应之间的关系。第一个基函数从两个选择中得到最大值：max（0, Env var-Knot）。所以在这个例子中，环境变量的节点是 11。对于低于 11 的值，方程“Env var-Knot”的结果是负数，因此基函数的结果为 0。这意味着对于节点之前的所有环境变量而言，第一个基函数的结果都是 0。而对于节点之后的所有值，第一个基函数的结果是环境变量减去 11。第二基函数则相反，节点之后的环境变量值是 0，节点之前的结果为 11 减去环境变量的值。对于只有 1 个环境变量和 1 个节点的简单模型，最终的模型也包括一个基线，它是响应值的平均值，以及节点两边的两个基函数。然而，大多数模型包括多个节点和多个环境变量，这会是更复杂的模型。

一般来说，这三个统计回归模型是非常有用的，因为它们都可以处理连续和分类类型的预测变量，其设计适合复杂的预测因子与响应之间的关系。广义可加模型和多变量自适应回归样条比 GLM 对异常值更加稳健，但随之而来的问题是它容易过度拟合。广义线性模型和广义可加模型对于较小的数据集效果较好，但要记得，模型中的预测变量越多，物种分布的数据集应该越大。根据经验，预测变量的数量应该小于物种分布数据总数的十分之一。所以，如果你有一个 100 个物种分布位点，你最多有 10 个环境变量可以作为模型的预测因子。多元自适应回归样条可以很好地处理大数据集，尽管算法复杂，运算往往高效快速。虽然这些算法有些难以理解和解释，但当你要运行物种分布模型时，它们是绝对值得考虑的。

感谢你观看物种分布课程的第 6 单元。在第 7 单元中，我们将详细讲解最后一组模型-机器学习模型。