多元回归

概述

“协助”中的多元回归法可使用最小二乘估计将线性和二次模型与多达五个预测变量 (X) 和一个连续反应 (Y) 拟合。用户选择模型类型，“协助”选择模型项。在本文中，我们解释了“协助”用于选择回归模型的标准。

此外，我们研究了对获得有效的回归模型至关重要的几个因素。首先，样本必须足够大，能够提供足够的检验功效并为估计 X 和Y 之间的关系强度提供足够的精度。其次，识别可能影响分析结果的异常数据很重要。我们还考虑误差项遵循正态分布的假设，并评估非正态性对整体模型假设检验的影响。

按照上述因素，“协助”会对您的数据自动执行以下检查并在“报告卡”中报告研究结果：

- 数据量
- 异常数据
- 正态性

在本书中，我们探讨了如何将这些因素与实际应用中的回归分析关联起来，并介绍了如何确立相关准则，以在“协助”中检查这些因素。

回归法

模型选择

“协助”中的回归分析可将一个模型与一个连续反应和二至五个预测变量拟合。有一个可能是类别预测变量。有两种模型可供选择：

- 线性： $F(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
- 二次： $F(x) = \beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j$

“协助”从全线性或二次模型中选择模型项。

目标

我们要研究可用于模型选择的不同方法，以确定“协助”中使用哪一个模型。

方法

我们研究了三种不同类型的模型选择：后退、前进、逐步。我们还研究了这些模型选择类型中包括的以下几个选项：

- 用于从模型中输入或删除项的标准。
- 是否在模型中强加某些项或在初始模型中包含某些项。
- 模型的层次结构。
- 规范模型中的 X 变量。

我们审查了这些选项，查看了它们对过程结果的影响，并考虑了哪些是从业者的首选方法。

结果

我们在“协助”中选择模型项的步骤如下：

- 使用逐步模型选择。一组潜在的 X 变量通常相关，因此一个项的效果将取决于模型中还有其他哪些项。在这种情况下，逐步选择可以说是最好的方法，因为它允许一步输入多个项，不过日后要将某些项删除，这取决于模型中要包括哪些项。
- 模型的层次结构在每个步骤得到维护，多个项可以在同一步骤中输入模型。例如，如果最显著的项是 X_1^2，那么它将与 X_1 一起输入，无论 X_1 是否显著。层次结构是适当的，因为它允许模型从标准化变成非标准化。并且，由于层次结构允许多个项在任何步骤输入模型，从而可以识别重要的平方项或交互项，即使相关的线性项与反应并不密切相关也如此。
- 根据 $\alpha = 0.10$ 从模型中输入或删除项。使用 $\alpha = 0.10$ 的过程将比核心 Minitab 中，采用的 $\alpha = 0.15$ 逐步过程具有更多的选择性。
- 为了选择模型项，将通过减去平均值并除以标准差来规范预测变量。最终模型用非标准化 X 值的单位表示。X 值的标准化将消除线性和平方项之间的大部分相关性，从而减少不必要地增加高阶项的机会。

数据检查

数据量

功效涉及假设检验否定不成立的原假设的可能性。对于回归法，原假设指出，X 和 Y 之间没有关系。如果数据集太小，检验可能无法检测到 X 和 Y 之间实际存在的关系。因此，数据集应足够大，可以高概率地检测到十分重要的关系。

目标

我们想要确定数据量如何影响整体 F 检验的功效，包括 X 和 Y 之间的关系、R^2_{adj} 的精度，X 和 Y 之间关系的强度估计。这些信息对于确定数据集是否足够大至关重要，从而确保在数据中观测到的关系强度成为真正决定性的关系强度的可靠指标。有关 R^2_{adj} 的详细信息，请参见附录 A。

方法

我们采取了类似的方法来确定我们为简单回归法推荐使用的样本量。我们研究了 R^2_{adj} 值中的变化，来确定样本应该多大，R^2_{adj} 才接近 ρ^2_{adj}。我们还证实，即使当 Y 和 X 变量之间的关系强度为中等偏弱时，推荐样本量也提供合理的功效。有关计算的详细信息，请参见附录 B。

结果

对于简单回归法，我们推荐您使用一个足够大的样本，以便拥有 90% 的置信度，让观测到的 R^2_{adj} 值位于 ρ^2_{adj} 的 0.20 的范围内。我们发现，随着您在模型中添加更多的项，所需样本量会增大。因此，我们计算了每个模型大小所需的样本量。推荐大小舍入到最接近 5 的倍数。例如，如果除常数外，模型还具有八个系数，如 4 个线性项，3 个交互项和 1 个平方项，则符合标准所需的最小样本量为 n = 49。“协助”将此值舍入到推荐的样本量 n = 50。有关基于项数的特定样本量推荐的详细信息，请参见附录 B。

我们也验证了推荐样本量提供足够强大的功效。我们发现，对于中等偏弱关系，ρ^2_{adj}= 0.25，功效通常为约 80% 或更多。因此，按照“协助”的样本量建议，可确保您获得相当不错的功效和准确的关系强度估计。

根据上述结果，检查数据量时，“协助报告卡”会显示以下信息：

状态	条件
	**样本量 < 推荐样本量** 样本量不够大，无法提供非常精确的关系强度估计。关系强度测量值（如 R 平方值和调整后的 R 平方值）具有很大的差别。为了获得精确的估计，应为此模型大小使用较大的样本。 **样本量 >= 推荐样本量** 样本足够大，可以获得关系强度的精确估计。

异常数据

在“协助回归”过程中，我们将异常数据界定为包含大标准化残差或大杠杆值的观测值。通常这些测量值用于确定回归分析中的异常数据 (Neter et al., 1996)。由于异常数据会对结果产生重要的影响，您可能需要修正数据，才能使分析有效。然而，过程的自然变化中也会产生异常数据。因此，识别异常行为的原因以确定如何处理此类数据点至关重要。

目标

我们想要确定需要多大的标准化残差和杠杆值，才会发出数据点异常信号。

方法

我们制定了各种准则，用于根据 Minitab 中的标准回归过程（统计 > 回归 > 回归）识别异常观测值。

结果

标准化残差

标准化残差等于残差值 e_i 除以其标准差的估计值。一般来说，如果标准化残差的绝对值大于 2，则观测值被认为异常。但是，这项准则有些保守。您可以期望所有观测值中约 5% 的观测值满足这一标准（如果误差呈正态分布）。因此，一定要研究异常行为的原因，以确定观测值是否确实存在异常。

杠杆值

杠杆值只与观测值的 X 值有关，与 Y 值无关。如果杠杆值是模型系数个数 (p) 除以观测个数 (n) 的 3 倍以上，则观测值被确定为异常。这还是一个常用的临界值，不过一些资料会使用 $\frac{2 \times p}{n}$ (Neter et al., 1996)。

如果您的数据包含任何高杠杆点，考虑它们是否对选定用于拟合数据的模型有不当影响。例如，一个 X 极值可能会导致选择二次模型，而不是线性模型。您应该考虑在二次模型中观测到的曲率是否与您对过程的了解一致。如果不一致，为数据拟合一个简单的模型，或收集更多的数据来深入了解此过程。

检查异常数据时，“协助报告卡”中显示以下状态指标：

状态	条件
	没有异常数据点。
	至少有一个或多个大标准化残差或至少有一个或多个高杠杆点。

正态性

回归法中的典型假设是随机误差（ε）呈正态分布。对系数估计（β）进行假设检验时，正态性假设非常重要。幸运的是，当样本量足够大时，即使随机误差不呈正态分布，检验结果通常也可靠。

目标

我们想要确定需要多大的样本，才能提供基于正态分布的可靠结果。我们想要确定实际的检验结果与检验的目标显著性水平（alpha 值或 I 类误差率）有多接近；也就是说，对于不同的非正态性分布，检验错误地否定原假设的频率与预期相比是高还是低。

方法

为了估计 I 类误差率，我们对显著不同于正态分布的偏态、重尾和轻尾分布进行了多次模拟。我们使用 15 个样本进行了模拟。我们研究了几种模型的整体 F 检验。

对于每种情况，我们进行了 10000 次检验。我们生成了随机数据，使得每次检验的原假设都为真。然后，我们使用 0.10 的目标显著性水平进行了检验。我们计算了这 10000 次检验中实际否定原假设的次数，并将此比率与目标显著性水平进行了比较。如果检验方法很有效，则 I 类误差率应非常接近目标显著性水平。有关模拟的详细信息，请参见附录 C。

结果

无论是哪种整体 F 检验，发现任何非正态分布的统计显著结果的概率差别并不明显。I 类误差率全部介于 0.08820 和 0.11850 之间，相当接近 0.10 的目标显著性水平。

因为相对较小的样本的检验进展顺利，“协助”不检验数据的正态性。相反，“协助”检查样本量，并在样本量小于 15 时指出。“协助”会在“回归报告卡”中显示以下状态指标：

状态	条件
	样本量至少为 15 个，因此正态性不是问题。
	由于样本量小于 15，可能存在正态性问题。解释 p 值时应小心。如果样本数较小，p 值的准确度很容易受非正态残差误差的影响。

参考书

Neter, J., Kutner, M.H., Nachtsheim, C.J., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago: Irwin.

附录 A：模型和统计数据

将预测变量 X 与反应 Y 关联的回归模型的形式如下：

$Y = f(X) + \varepsilon$

其中函数 f(X) 表示为 X（平均值）指定的 Y 预期值。

在“协助”中，函数 f(X) 有两种选择形式：

模型类型	f(X)
线性	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
二次	$\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j$

系数值 β 未知，必须从数据中估计得出。估计方法是最小二乘法，它将最大限度地减少样本中的残差平方和：

$$\min \sum_{i=1}^{n} \left(Y_i - \hat{f}(X_i)\right)^2.$$

根据估计系数法，残差是观测到的反应 Y_i 与拟合值 $\hat{f}(X_i)$ 之间的差。此平方和的最小值是给定模型的 SSE（误差平方和）。

整体 F 检验

这种方法是整体模型（线性或二次）检验。对于选定形式的回归函数 f(X)，它将检验：

$H_0: f(X)$ 是常量

$H_1: f(X)$ 不是常量

已调整 R^2

已调整 R^2（R^2_{adj}）测量有多少反应变化归因于模型的 X 变量。测量 X 与 Y 变量之间观测到的关系强度有两种常用方法：

$$R^2 = 1 - \frac{SSE}{SSTO}$$

而且

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

其中

$$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

SSTO 是平方总和，用于测量有关总平均值 $\bar{Y}$ 的反应变化。SSE 测量有关回归函数 f(X) 的变化。R^2_{adj} 中的调整针对整个模型中的系数个数 (p) 进行，从而保留 n - p 的自由度来估计 ε 的方差。当在模型中添加更多系数时，R^2 从不减小。不过由于调整，当其他系数未改

进模型时，R^2_{adj} 会减小。因此，如果在模型中增加另一个项不能说明反应中发生了任何其他变化，则 R^2_{adj} 会减小，表明附加项没用。因此，应使用已调整的测量值来比较不同大小的模型。

F 检验与 R^2_{adj} 之间的关系

整体模型检验的 F 统计数据可以用 SSE 和 SSTO 来表示，也可用于计算 R^2_{adj}：

$$F = \frac{(SSTO - SSE)/(p-1)}{SSE/(n-p)}$$

$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{adj}}{1-R^2_{adj}}$. 上述公式表明，F 统计数据是 R^2_{adj} 的递增函数。因此，当且仅当 R^2_{adj} 超过检验的显著性水平（α）确定的特定值时，检验才否定 H_0。

附录 B：数据量

在本节中，我们将探讨观测个数 n 如何影响整体模型检验的功效和 R^2_{adj} 的精度、模型的强度估计。

为了量化关系强度，我们将引入一个新的分量 ρ^2_{adj}，作为样本统计数据 R^2_{adj} 的总体分量。回想一下

$$R^2_{adj} = 1 - \frac{SSE/(n-p)}{SSTO/(n-1)}$$

因此，我们定义

$$\rho^2_{adj} = 1 - \frac{E(SSE|X)/(n-p)}{E(SSTO|X)/(n-1)}$$

操作员 E(·|X) 表示预期值，或为 X 值指定的随机变量的平均值。假设正确的模型是 ε 呈独立恒等分布的 $Y = f(X) + \varepsilon$，我们有

$$\frac{E(SSE|X)}{n-p} = \sigma^2 = Var(\varepsilon)$$

$$\frac{E(SSTO|X)}{n-1} = \sum_{i=1}^{n} \frac{(f(X_i) - \bar{f})^2}{(n-1)} + \sigma^2$$

其中 $\bar{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$.

因此，

$$\rho^2_{adj} = \frac{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1)}{\sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/(n-1) + \sigma^2}$$

整体模型显著性

当检验整体模型的统计显著性时，我们假设随机误差 ε 呈独立正态分布。然后，在平均值 Y 为常数 $(f(X) = \beta_0)$ 的原假设下，F 检验统计数据呈 $F(p-1, n-p)$ 分布。在其他假设下，F 统计数据呈非中心 $F(p-1, n-p, \theta)$ 分布，包含非中心参数：

$$\theta = \sum_{i=1}^{n} \left(f(X_i) - \bar{f}\right)^2/\sigma^2$$

$$= \frac{(n-1)\rho^2_{adj}}{1 - \rho^2_{adj}}$$

否定 H_0 的概率增加，n 和 ρ^2_{adj} 中的非中心参数也随之增加。

相关强度

如简单回归法中所示，数据中的统计显著关系并不一定表明 X和 Y 之间存在显著重要的关系。因此许多用户希望通过查看 R^2_{adj} 等指标，来弄清楚这一关系实际上有多么重要。如果我

们考虑 R^2_{adj} 作为 ρ^2_{adj} 的估计值，那么我们有足够的理由相信估计值相当接近于真实的 ρ^2_{adj} 值。

对于每一个可能的模型大小，我们通过确定 n 的最小值（绝对差 $|R^2_{adj} - \rho^2_{adj}|$ 大于 0.20 的机率不超过 10%），来确定可接受的样本量的相应阈值。不管真实值是否为 ρ^2_{adj}。推荐的样本量 n(T) 在下表中总结，其中 T 是模型中除恒定系数以外的系数个数。

T	n(T)
1-3	40
4-6	45
7-8	50
9-11	55
12-14	60
15-18	65
19-21	70
22-24	75
25-27	80
28-31	85
32-34	90
35-38	95
39-41	100
42-45	105
46-48	110
49-52	115
53-56	120
57-59	125
60-63	130
64-67	135
68-70	140
71-73	145

我们为中等偏弱值 $\rho_{adj}^2 = 0.25$ 的模型评估了整体 F 检验的功效，以确认推荐的样本量具有足够的功效。下表中的模型大小显示每个 n(T) 值的最差情况。使用相同的 n(T) 值的模型越小，功效越大。

T	n(T)	功效为 $\rho_{adj}^2 = 0.25$
3	40	0.902791
6	45	0.854611
8	50	0.850675
11	55	0.831818
14	60	0.820592
18	65	0.798003
21	70	0.796425
24	75	0.796911
27	80	0.798856
31	85	0.789861
34	90	0.794367
38	95	0.788625
41	100	0.794511
45	105	0.790864
48	110	0.797487
52	115	0.79525
56	120	0.793698
59	125	0.800982
63	130	0.800230
67	135	0.799906
69	140	0.814664

附录 C：正态性

“协助”中使用的回归模型都是以下形式：

$Y = f(X) + \varepsilon$

有关随机项 ε 的典型假设是，它们是独立恒等分布的正态随机变量，平均值为 0，常见方差为 σ^2。β 参数的最小二乘估计仍然是最佳线性无偏估计，即使我们放弃 ε 是正态分布的假设。正态性假设仅当我们试图为这些估计赋予概率时才至关重要，正如我们在有关 f(X) 的假设检验中所做的那样。

我们想要确定 n 值需要多大，才能信任基于正态假设的回归分析结果。我们进行了各种模拟，探讨了各种非正态误差分布下假设检验的 I 类误差率。

下面的表 1 显示在 10000 次模拟中，对于三种不同模型的各种 ε 分布，有多少次模拟的整体 F 检验在 $\alpha = 0.10$ 时统计显著。在这些模拟中，指出 X 和 Y 变量之间没有任何关系的原假设成立。Minitab 的 RANDOM 命令生成 X 值作为多元正态变量。我们为所有检验使用 n=15 的样本量。所有模型都包括 5 个连续预测变量。第一个模型是包含所有五个 X 变量的线性模型。第二个模型包含所有的线性项和平方项。第三个模型包含所有的线性项和 7 个双向交互项。

表 1 n=15 的非正态分布整体 F 检验的 I 类误差率

分布	线性	线性 + 平方	线性 + 7 个交互
正态	0.09910	0.10270	0.10060
t(3)	0.09840	0.1185	0.118
t(5)	0.09980	0.10010	0.10430
Laplace	0.09260	0.09400	0.09650
均匀	0.10630	0.10080	0.09480
Beta(3, 3)	0.09980	0.10120	0.10020
指数	0.08820	0.09500	0.09960
Chi(3)	0.09890	0.114	0.10970
Chi(5)	0.09730	0.10590	0.10330
Chi(10)	0.10150	0.09930	0.10360
Beta(8, 1)	0.09870	0.10230	0.10490

模拟结果显示，发现统计显著结果的概率与任何误差分布的正态值 0.10 差别并不明显。观测到的 I 类误差率都在 0.08820 和 0.11850 之间。