简单回归

概述

此“协助”中的简单回归过程通过最小二乘估计，使用一个连续预测变量 (X) 和一个连续响应 (Y) 拟合线性和二次模型。用户可以选择模型类型，或让“协助”选择最佳拟合模型。在本白皮书中，我们解释了“协助”用于选择回归模型的标准。

此外，我们还检查了对于获取有效回归模型非常重要的几个因素。首先，样本必须足够大，才能为检验提供足够的功效，才能为估计 X 和 Y 之间的关系强度提供足够的精确度。其次，一定要找出可能会影响分析结果的异常数据。我们还要考虑误差项遵循正态分布的假设，并评估非正态性对整个模型和系数的假设检验的影响。最后，为确保模型有用，选定的模型类型一定要准确反映 X 和 Y 之间的关系。

根据这些因素，“协助”会自动对您的数据执行以下检查，并在“报告卡”中报告发现的结果：

- 数据量
- 异常数据
- 正态性
- 模型拟合

在本白皮书中，我们对这些因素与实际的回归分析的相关性进行了调查，并介绍了如何确定在“协助”中检查这些因素的原则。

回归法

模型选择

“协助”中的回归分析可使用一个连续预测变量和一个连续响应拟合一个模型，并且可以拟合两种类型的模型：

- 线性： $F(x) = \beta_0 + \beta_1 X$
- 二次：$F(x) = \beta_0 + \beta_1 X + \beta_2 X^2$

用户可以在执行分析之前选择模型，或者让“协助”选择模型。可以采用多种方法来确定哪个模型最适合于数据。为确保模型有用，选定的模型类型一定要准确反映 X 和 Y 之间的关系。

目标

我们想要检查可用于模型选择的不同方法，以确定哪种方法可用于“协助”中。

方法

我们检查了通常用于模型选择的三种方法（Neter 等人，1996）。第一种方法确定最高次项显著的模型。第二种方法选择具有最大 R^2_{adj} 值的模型。第三种方法选择整体 F 检验效果显著的模型。有关更多详细信息，请参见附录 A。

为了确定“协助”中使用的方法，我们检查了几种方法，并比较了这些方法之间的计算结果。我们还从质量分析专家那里收集了相关反馈。

结果

根据我们的研究，我们决定使用可根据模型中的最高次项的统计显著性来选择模型的方法。“协助”首先检查二次模型，并检验此模型中的二次项（β_2）的统计意义是否显著。如果该项的统计意义不显著，则它会从模型中删除此二次项，并检验线性项（β_1）。通过此方法选定的模型将出现在“模型选择报告”中。此外，如果用户选择的模型与“协助”选择的模型不同，则我们会在“模型选择报告”和“报告卡”中报告这种情况。

我们之所以选择此方法，部分归因于质量专业人员的反馈，他们通常更喜欢使用排除了不显著项的更简单模型。此外，根据我们对各种方法的比较，使用模型中最高项的统计显著性比根据最高 R^2_{adj} 值选择模型的方法更严格。有关详细信息，请参见附录 A。

虽然我们使用最高模型项的统计显著性来选择模型，但我们还要在“模型选择报告”中提供该模型的 R^2_{adj} 值和整体 F 检验。要查看“报告卡”中显示的状态指示符，请参见下面的“模型拟合数据检查”部分。

数据检查

数据量

功效与假设检验拒绝原假设（原假设为假时）的可能性相关。对于回归，原假设指示 X 和 Y 之间没有任何关系。如果数据集太小，则该检验的功效可能不足以检测到 X 和 Y 之间实际存在的关系。因此，数据集应该足够大，才能有较高的概率检测到实际的重要关系。

目标

我们想要确定数据量如何影响 X 和 Y 之间关系的整体 F 检验的功效、R^2_{adj} 的精确度以及 X 和 Y 之间关系强度的估计值。对于确定数据集是否足够大，以便可以信任在数据中观测到的关系强度是此关系真实基本强度的可靠指标，这些信息至关重要。有关 R^2_{adj} 的详细信息，请参见附录 A。

方法

为了检查整体 F 检验的功效，我们对一系列 R^2_{adj} 值和样本数量进行了功效计算。为了检查 R^2_{adj} 的精确度，我们针对不同的总体调整值 R^2（ρ^2_{adj}）和不同的样本数量进行了 R^2_{adj} 分布模拟。我们检查了 R^2_{adj} 值中的可变性，以确定样本数量应该多大，R^2_{adj} 才能接近于 ρ^2_{adj}。有关计算和模拟的详细信息，请参见附录 B。

结果

我们发现，对于中等大小的样本数量，要检测 X 和 Y 之间的关系，回归的功效比较高，即使这些关系的强度不足以满足实际需求也是如此。更具体地说，我们发现：

- 样本数量为 15，并且 X 和 Y 之间的关系较强（ρ^2_{adj} = 0.65）时，发现统计意义显著的线性关系的概率为 0.9969。因此，如果在数据点为 15 或更多时检验未能找到统计意义显著的关系，则说明真实关系可能不是非常强（ρ^2_{adj} 值小于 0.65）。
- 样本数量为 40，并且 X 和 Y 之间的关系中等偏弱（ρ^2_{adj} = 0.25）时，发现统计意义显著的线性关系的概率为 0.9398。因此，如果有 40 个数据点，则 F 检验可能会找到 X 和 Y 之间的关系，即使在关系为中等偏弱时也是如此。

使用回归可以相当轻松地检测到 X 和 Y 之间的关系。因此，如果您发现统计意义显著的关系，还应该使用 R^2_{adj} 评估此关系的强度。我们发现，如果样本数量不够大，则 R^2_{adj} 不太可靠，并且样本之间的差异可能很大。但是，如果样本数量为 40 或更大，我们发现，R^2_{adj} 值更稳定、更可靠。如果样本数量为 40，则观测的 R^2_{adj} 值在 ρ^2_{adj} 的 0.20 范围内的置信度为 90%，这与实际值和模型类型（线性或二次）无关。有关模拟结果的详细信息，请参见附录 B。

根据这些结果，“协助”将在检查数据量时在“报告卡”中显示以下信息：

状态	条件
	**样本数量小于 40** 您的样本数量不足够大，无法提供非常精确的关系强度估计值。关系强度度量单位(如 R 平方和 R 平方 (调整))可能差异很大。要获得更精确的估计值，应该使用更大的样本数量(通常是 40 或更大)。 **样本数量小于≥ 40** 您的样本足够大，因此可以获得关系强度的准确估计值。

异常数据

在“协助回归”过程中，我们利用较大的标准化残差或较大的杠杆率值将异常数据定义为观测值。这些度量通常用于在回归分析（Neter 等人，1996）中确定异常数据。由于异常数据对结果会产生较大的影响，因此，可能需要更正这些数据以使分析有效。但是，异常数据也可能是由过程中的自然变异导致的。因此，确定异常行为的原因对于确定如何处理此类数据点至关重要。

目标

我们想要确定，要指出某个数据点为异常数据点，需要多大的标准化残差和杠杆率值。

方法

我们根据 Minitab 中的标准回归过程（统计 > 回归 > 回归），制定了用于确定异常观测值的原则。

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。通常，如果标准化残差的绝对值大于 2，则观测值被认为是异常值。但是，此原则有点保守。您应该预计到所有观测值中大约有 5% 的观测值符合此标准（如果误差呈正态分布）。因此，调查异常行为的原因对于确定观测值是否真正异常至关重要。

杠杆率值

杠杆率值仅与观测值的 X 值相关，与 Y 值无关。如果杠杆率值大于模型系数个数 (p) 的 3 倍除以观测值个数 (n)，则观测值将会被确定为异常值。虽然有些教材使用 $\frac{2 \times p}{n}$（Neter 等人，1996），但这仍然是常用的断点值。

如果您的数据包含任何高杠杆率点，请分析这些点是否会对选择用于拟合数据的模型类型产生不利影响。例如，单个 X 极值可能会导致选择二次模型而不是线性模型。您应该分析在二次模型中观测到的弯曲与您对此过程的了解是否保持一致。如果不一致，请将更简单的模型与这些数据拟合，或者收集其他数据，以便更全面地调查此过程。

在检查异常数据时，“协助报告卡”显示以下状态指示符：

状态	条件
	不存在异常数据点。异常数据点会对结果造成极大的影响。
	至少有一个或多个较大的标准化残差或至少有一个或多个高杠杆率值。 您可以将鼠标悬停在点上或使用 Minitab 的笔刷功能识别工作表行。由于异常数据会对结果造成极大影响，因此应尝试找出其异常性质的原因。请更正任何数据输入或测量值错误。可以考虑删除与特殊原因关联的数据，然后重复进行此分析。

正态性

回归中的典型假设是随机误差（ε）呈正态分布。在对系数（β）估计值进行假设检验时，正态性假设非常重要。幸好，在样本数量足够大时，即使随机误差不呈正态分布，检验结果通常也是可靠的。

目标

我们想要确定，根据正态分布，要提供可靠的结果，需要多大的样本数量。我们想要确定实际检验结果与检验的目标显著性水平（alpha 或 I 类错误率）的匹配度；即，对于不同的非正态分布，其检验错误地拒绝原假设这种状况出现的频率比期望的更多还是更少。

方法

为了估计 I 类错误率，我们利用与正态分布有很大偏离的偏斜、重尾和轻尾分布执行了多次模拟。我们使用样本数量 15 进行了线性和二次模型模拟。我们检查了整体 F 检验和模型中的最高次项检验。

对于每种情况，我们执行了 10,000 次检验。我们生成了随机数据，这样，对于每次检验，原假设均为真。然后，我们使用 0.05 的目标显著性水平进行了检验。我们计算了这 10,000 次检验中实际拒绝原假设的次数，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 类错误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

对于整体 F 检验和模型中的最高次项检验，对任何非正态分布能发现结果统计意义显著的概率之差异不是很大。I 类错误率全都介于 0.038 到 0.0529 之间，非常接近目标显著性水平 0.05。

由于数量相对较小的样本只在正态条件下之检验效果才依然较好，而“协助”并未检验数据的正态性，而只是检查样本数量，因而当样本数量小于 15 时予以指出并提醒检验正态性。“协助”会在回归的“报告卡”中显示以下状态指示符：

状态	条件
	样本数量至少为 15，因此，正态性不是问题。

状态	条件
	由于样本数量小于 15，正态性可能是个问题。在解释 p 值时要格外小心。如果样本数较小，p 值的准确度很容易受非正态残差误差影响。

模型拟合

您可以在执行回归分析之前选择线性或二次模型，或者选择让“协助”选择模型。可使用多种方法来选择适当的模型。

目标

我们想要检查用于选择模型类型的不同方法来确定将在“协助”中使用哪种方法。

方法

我们检查了常用于模型选择的三种方法。第一种方法确定最高次项显著的模型。第二种方法选择具有最大 R^2_{adj} 值的模型。第三种方法选择整体 F 检验效果显著的模型。有关更多详细信息，请参见附录 A。

为了确定“协助”中使用的方法，我们检查了几种方法，并比较了这些方法之间的计算结果。我们还从质量分析专家那里收集了相关的反馈。

结果

我们决定使用可根据模型中的最高次项的统计显著性选择模型的方法。“协助”首先检查二次模型，并检验此模型中的二次项（β_3）的统计意义是否显著。如果该项的统计意义不显著，则它会检验线性模型中的线性项（β_1）。通过此方法选定的模型将出现在“模型选择报告”中。此外，如果用户选择的模型与“协助”选择的模型不同，则我们会在“模型选择报告”和“报告卡”中报告这种情况。有关详细信息，请参见上面的回归方法部分。

根据我们的发现，“协助报告卡”显示以下状态指示符：

状态	条件
	**如果用户的模型与“协助”的最佳拟合模型匹配** 您应该根据自己的目标来评估数据和模型拟合状况。查看拟合线图以确保： • 样本充分覆盖 X 值范围。 • 模型与数据中的任何弯曲正确拟合（避免过度拟合）。 • 该线条在任何特定的相关区域中都能进行很好的拟合。 **如果用户的模型与“协助”的最佳拟合模型不匹配** “模型选择报告”将显示可能是更好选择的替代模型。

参考书

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

附录 A：模型选择

预测变量 X 与响应 Y 相关的回归模型形式如下：

$Y = f(X) + \varepsilon$

函数 f(X) 表示给定 X 时 Y 的期望值（均值）。

在“协助”中，针对函数 f(X) 的形式提供了两个选项：

模型类型	f(X)
线性	$\beta_0 + \beta_1 X$
二次	$\beta_0 + \beta_1 X + \beta_2 X^2$

系数 β 的值未知，必须根据数据进行估计。估计方法是最小二乘法，该方法可使样本中的残差平方总和最小化：

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

根据估计的系数，残差是观测响应 Y_i 和拟合值 $\hat{f}(X_i)$ 之间的差值。对于给定模型，此平方和的最小值为 SSE（误差平方和）。

为确定在“协助”中用于选择模型类型的方法，我们评估了以下三个选项：

- 模型中的最高次项的显著性
- 模型的整体 F 检验
- 调整的 R^2 值（R^2_{adj}）

模型中的最高次项的显著性

在此方法中，“协助”从二次模型开始。“协助”将检验二次模型中二次项的假设：

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

如果拒绝此原假设，则“协助”会得出二次项系数非零的结论，并选择二次模型。如果不拒绝原假设，则“协助”将检验线性模型的假设：

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

整体 F 检验

此方法将对整个模型（线性或二次）进行检验。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

调整的 R^2

调整的 R^2（R^2_{adj}）可度量在响应产生的总变异中，此模型归属于 X 的变异有多大。有两种常用方法可度量 X 和 Y 之间观测到的关系强度：

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

和

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

其中

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

SSTO 是总平方和，可度量响应与其总平均值 $\bar{Y}$ 间的变异。SSE 可度量观测值与回归函数 f(X) 间的变异。R^2_{adj} 中的调整针对全模型中的系数个数（p）进行，自由度为 n - p，用于估计 ε 的方差。在向模型添加更多个系数时，R^2 从不减小。但是，由于进行了此调整，如果添加系数不能改进模型效果，R^2_{adj} 就会减小。因此，如果向模型中添加一项并不能增加解释响应中的方差，则 R^2_{adj} 会减小，这表示此添加项没有用。因此，调整的度量将会被用来比较线性和二次模型。

模型选择方法之间的关系

我们想要检查三种模型选择方法之间的关系，如何对其进行计算，以及它们之间如何相互影响。

首先，我们从如何计算整体 F 检验和 R^2_{adj} 查看了他们之间的关系。整个模型的 F 检验的统计量可按 SSE 和 SSTO（也用于 R^2_{adj} 计算中）形式来表示：

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

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R^2_{adj}}{1-R^2_{adj}}.$$

上述公式表明，F 统计量是 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，此检验才会拒绝 H_0。为了说明这点，针对下表 1 中所示的不同样本数量，我们采用 $\alpha = 0.05$ 计算了获取二次模型的统计显著性所需的最小值 R^2_{adj}。例如，如果 n = 15，则模型的 R^2_{adj} 值必须至少为 0.291877，整体 F 检验才具有显著的统计意义。

表 1　针对不同样本数量，采用 $\alpha = 0.05$ 时效果显著的二次模型的整体 F 检验的最小值 R^2_{adj}

样本数量	最小值 R^2_{adj}
4	0.9925
5	0.90
6	0.773799
7	0.66459

样本数量	最小值 R^2_{adj}
8	0.577608
9	0.508796
10	0.453712
11	0.408911
12	0.371895
13	0.340864
14	0.314512
15	0.291877
16	0.272238
17	0.255044
18	0.239872
19	0.226387
20	0.214326
21	0.203476
22	0.193666
23	0.184752
24	0.176619
25	0.169168
26	0.162318
27	0.155999
28	0.150152
29	0.144726
30	0.139677
31	0.134967
32	0.130564
33	0.126439
34	0.122565

样本数量	最小值 R^2_{adj}
35	0.118922
36	0.115488
37	0.112246
38	0.109182
39	0.106280
40	0.103528
41	0.100914
42	0.098429
43	0.096064
44	0.093809
45	0.091658
46	0.089603
47	0.087637
48	0.085757
49	0.083955
50	0.082227

随后，我们检查了模型中最高次项的假设检验和 R^2_{adj} 之间的关系。最高次项的检验（如二次模型中的二次项）可以按整个模型（例如，二次模型）的平方和或 R^2_{adj} 形式表示，也可以按简化模型（例如，线性）的 R^2_{adj} 形式表示：

$$F = \frac{SSE(\text{简化}) - SSE(\text{完整})}{SSE(\text{完整})/(n-p)}$$

$$= 1 + \frac{(n-p+1)\left(R^2_{adj}(\text{完整}) - R^2_{adj}(\text{简化})\right)}{1 - R^2_{adj}(\text{完整})}.$$

这些公式表明，对于 $R^2_{adj}(\text{简化})$ 的固定值，F 统计量是 $R^2_{adj}(\text{完整})$. 的递增函数。这些公式还表明检验统计量与两个 R^2_{adj} 值之间的差值的依赖程度。尤其是，整个模型的值必须大于简化模型的值，才能获得足以使统计意义显著的较大 F 值。因此，使用最高次项显著性来选择最佳模型的方法比选择具有最高 R^2_{adj} 的模型的方法更严格。最高次项方法也符合许多用户偏爱选择更简单的模型的要求。因此，我们决定在“协助”中使用最高次项的统计意义显著性来选择模型。

有些用户更倾向于选择与数据拟合度最佳的模型，即具有最高 R^2_{adj} 的模型。“协助”会在“模型选择报告”和“报告卡”中提供这些值。

附录 B：数据量

在此部分中，我们将分析观测值个数 n 如何影响整个模型检验的功效、R^2_{adj} 的精确度以及估计模型的强度。

为了量化关系强度，我们引入了一个新的量 ρ^2_{adj}，作为样本统计量 R^2_{adj} 的总体相应项。请回想一下

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

因此，我们定义

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

运算符 E(·|X) 表示期望值，或给定 X 值时随机变量的均值。假设正确的模型中是 $Y = f(X) + \varepsilon$ 带有独立同分布的 ε，则我们可以得到

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

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

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

因此，

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

整个模型显著性

在检验整个模型的统计意义显著性时，我们假设随机误差 ε 呈独立正态分布。然后，在 Y 的均值为常量 $(f(X) = \beta_0)$ 的原假设下，F 检验统计量具有 $F(p-1, n-p)$ 分布。在备择假设下，F 统计量具有非中心 $F(p-1, n-p, \theta)$ 分布（带有非中心参数）：

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

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

当 n 和 ρ^2_{adj} 都递增时，非中心参数也递增，则拒绝 H_0 的概率会增加。

对于线性和二次模型，当 n = 15 时，我们使用上面的公式，计算了一系列 ρ^2_{adj} 值的整体 F 检验的功效。要获得相关结果，请参见表 2。

表 2　在 n=15 时，具有不同 ρ^2_{adj} 值的线性和二次模型的功效

ρ^2_{adj}	θ	F 的功效 线性	F 的功效 二次
0.05	0.737	0.12523	0.09615
0.10	1.556	0.21175	0.15239
0.15	2.471	0.30766	0.21896
0.20	3.50	0.41024	0.2956
0.25	4.667	0.5159	0.38139
0.30	6.00	0.62033	0.47448
0.35	7.538	0.71868	0.57196
0.40	9.333	0.80606	0.66973
0.45	11.455	0.87819	0.76259
0.50	14.00	0.93237	0.84476
0.55	17.111	0.96823	0.91084
0.60	21.00	0.9882	0.95737
0.65	26.00	0.99688	0.98443
0.70	32.667	0.99951	0.99625
0.75	42.00	0.99997	0.99954
0.80	56.00	1.00	0.99998
0.85	79.333	1.00	1.00
0.90	126.00	1.00	1.00
0.95	266.00	1.00	1.00

总之，我们发现，在 X 和 Y 之间的关系紧密，并且样本数量至少为 15 时，此检验已具有较高的功效。例如，在 ρ^2_{adj} = 0.65 时，表 2 显示，在 $\alpha = 0.05$ 时发现统计意义显著的线性关系的概率为 0.99688。通过 F 检验检测不到这一紧密关系的情况仅出现在不到 0.5% 的样本中。即使对于二次模型，通过 F 检验检测不到这一关系的情况仅出现在不到 2% 的样本中。因此，在此检验采用 15 或更多观测值还找不到统计意义显著的关系时，这就表明，真正的关系只能是其 ρ^2_{adj} 值小于 0.65。请注意，即使 ρ^2_{adj} 未达到 0.65 这样大，实际工作中也可能是有意义的。

我们也想在样本数量更大（n=40）时检查整体 F 检验的功效。我们确定，样本数量 n = 40 是 R^2_{adj} 的精确度的重要阈值（请参见以下关系强度），并想要评估样本数量与功效值。对于

线性和二次模型，当 n = 40 时，我们计算了一系列 ρ_{adj}^2 值的整体 F 检验的功效。要获得相关结果，请参见表 3。

表 3　n = 40 时，具有不同 ρ_{adj}^2 值的线性和二次模型的功效

ρ_{adj}^2	θ	F 的功效 线性	F 的功效 二次
0.05	2.0526	0.28698	0.21541
0.10	4.3333	0.52752	0.41502
0.15	6.8824	0.72464	0.60957
0.20	9.75	0.86053	0.76981
0.25	13.00	0.9398	0.88237
0.30	16.7143	0.97846	0.94925
0.35	21.00	0.99386	0.98217
0.40	26.00	0.99868	0.99515
0.45	31.9091	0.9998	0.99905
0.50	39.00	0.99998	0.99988
0.55	47.6667	1.00	0.99999
0.60	58.50	1.00	1.00
0.65	72.4286	1.00	1.00

我们发现，即使在 X 和 Y 之间的关系中等偏弱时，功效仍较高。例如，即使在 ρ_{adj}^2 = 0.25 时，表 3 仍显示在 $\alpha = 0.05$ 时发现统计意义显著的线性关系的概率为 0.93980。如果有 40 个观测值，则 F 检验不大可能未检测到 X 和 Y 之间的关系，即使此关系中等偏弱也是如此。

关系强度

正前所述，数据中统计意义显著的关系不一定指示 X 和 Y 之间较强的潜在关系。这就是为什么许多用户需要查看指标（如 R_{adj}^2）来了解关系的实际强度。如果我们将 R_{adj}^2 作为 ρ_{adj}^2 的估计值，则我们想要有置信区间，使得我们能估计出它和其 ρ_{adj}^2 真值的合理接近程度。

为说明 R_{adj}^2 和 ρ_{adj}^2 之间的关系，我们针对不同的 ρ_{adj}^2 值模拟了 R_{adj}^2 的分布，以查看 R_{adj}^2 随着不同 n 的变化情况。下面的图 1-4 中的图形显示 R_{adj}^2 的 10,000 个模拟值的直方图。在每对直方图中，ρ_{adj}^2 的值相同，因此，我们可以比较样本数量为 15 至 40 时 R_{adj}^2 的变化情况。我们检验了 0.0、0.30、0.60 和 0.90 的 ρ_{adj}^2 值。所有模拟都采用线性模型进行。

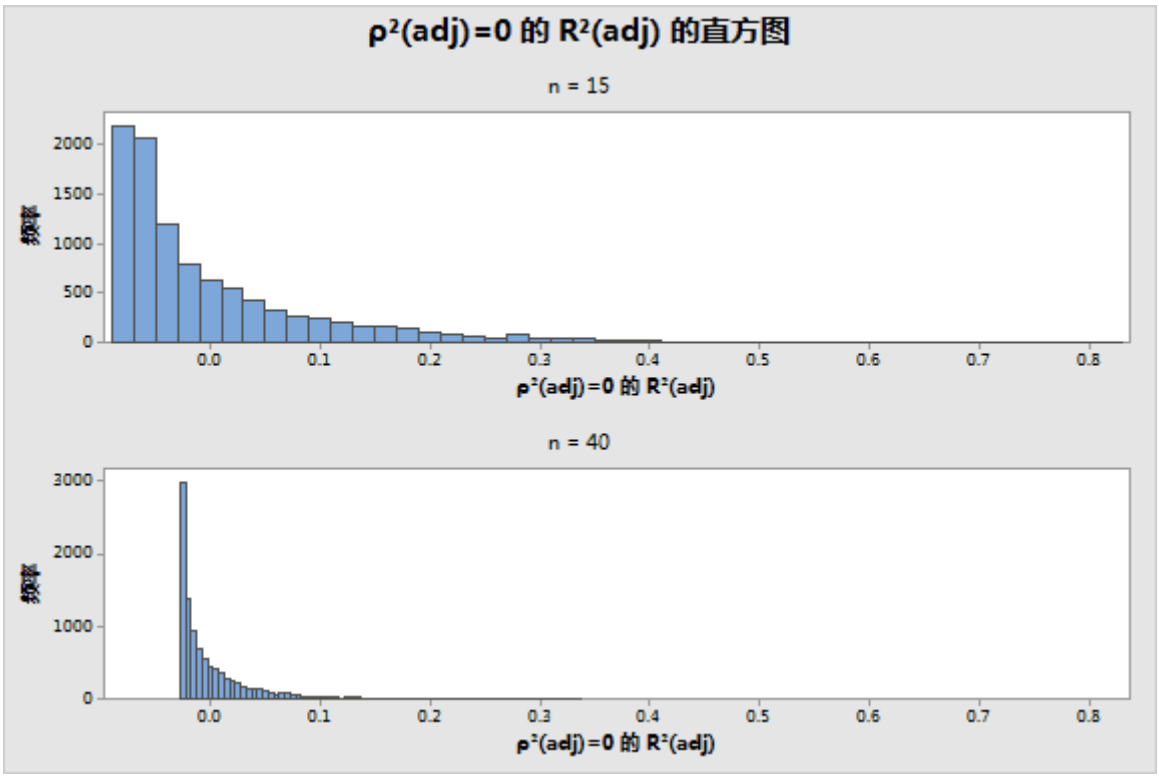

图 1　对于 n=15 和 n=40，ρ^2_{adj} = 0.0 的模拟 R^2_{adj} 值

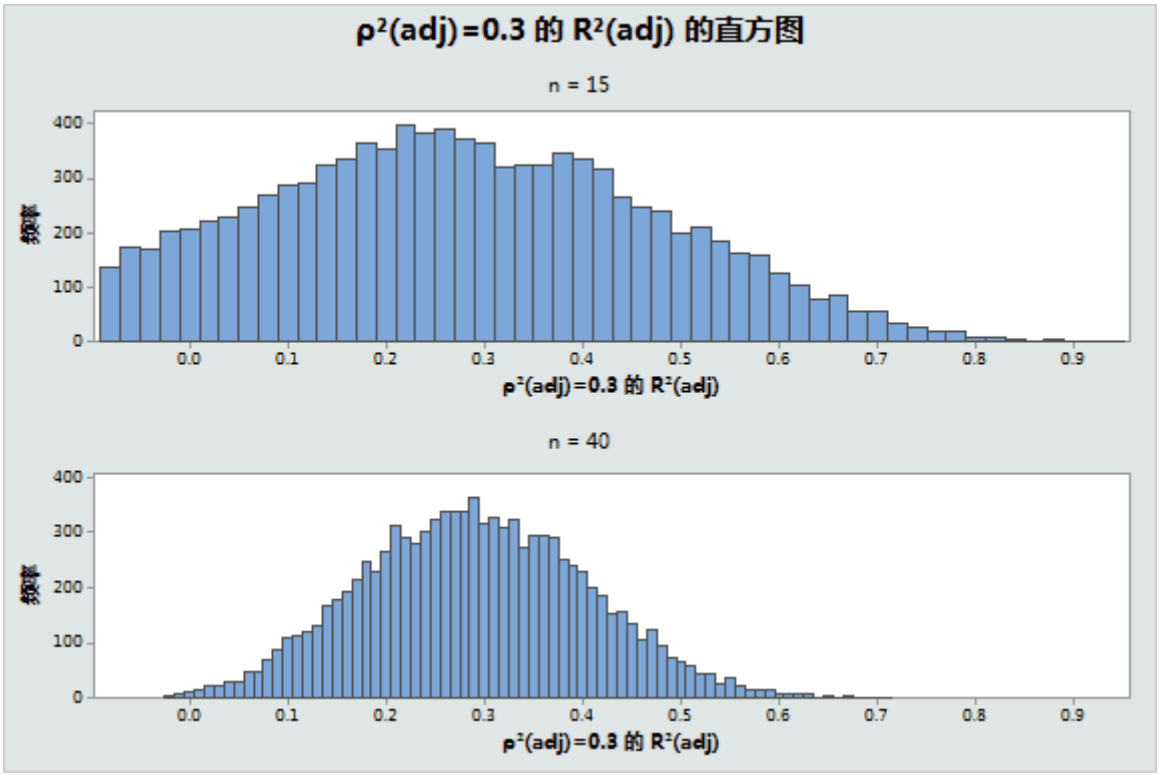

图 2　对于 n=15 和 n=40，ρ^2_{adj} = 0.30 的模拟 R^2_{adj} 值

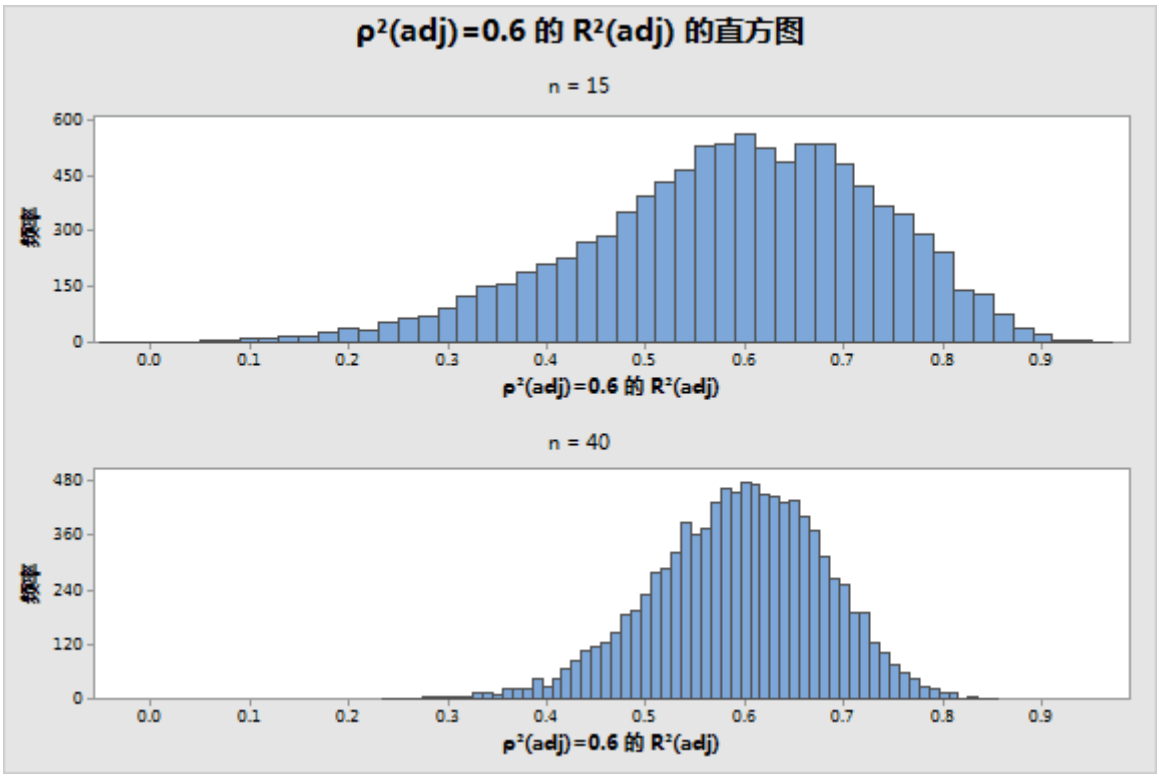

图 3　对于 n=15 和 n=40，ρ^2_{adj} = 0.60 的模拟 R^2_{adj} 值

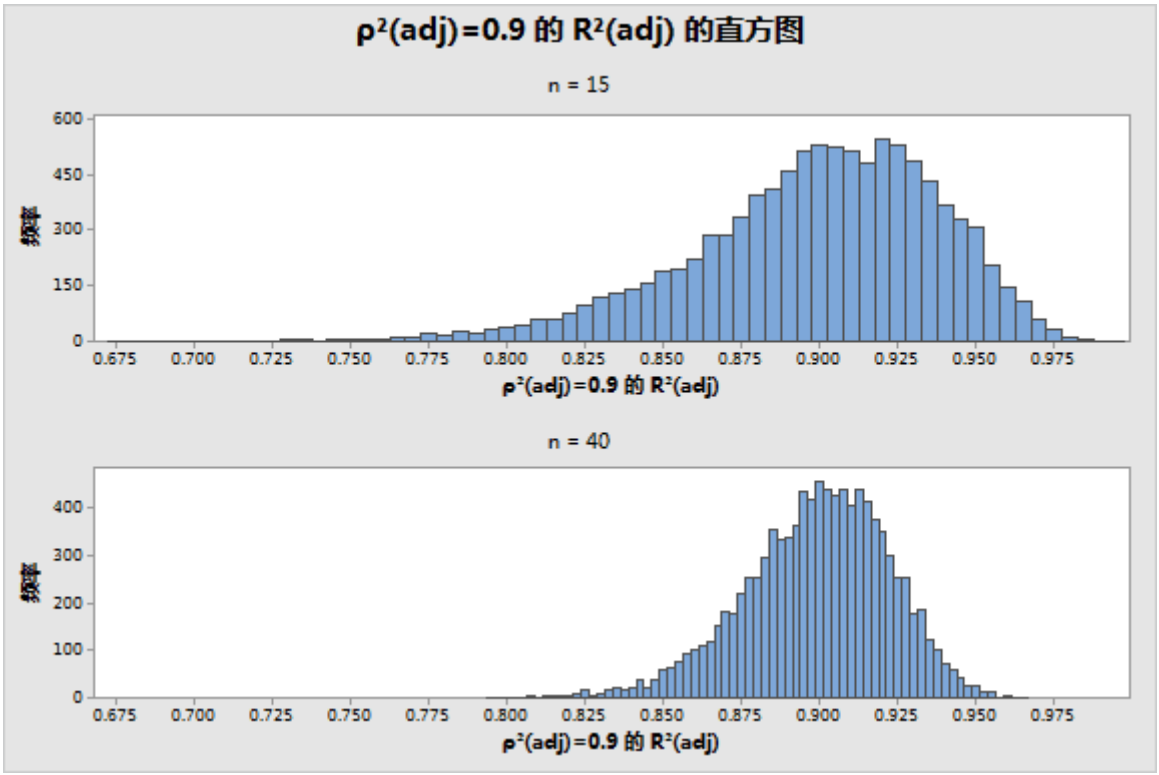

图 4　对于 n=15 和 n=40，ρ^2_{adj} = 0.90 的模拟 R^2_{adj} 值

总之，模拟表明，此关系的实际强度（ρ^2_{adj}）与在数据中观测到的关系强度（R^2_{adj}）之间存在相当大的差异。将样本数量从 15 增加到 40 可以大大降低差值可能的量值。我们通过找出大于 0.20 的绝对差值 $|R^2_{adj} - \rho^2_{adj}|$ 出现的概率不超过 10% 的最小值 n，确定 40 个观测值是合适的阈值。在任何分析的模型中，这与 ρ^2_{adj} 的真值无关。对于线性模型，最难的情况是 ρ^2_{adj} = 0.31，这种情况需要 n = 36。对于二次模型，最难的情况是 ρ^2_{adj} = 0.30，这种情况需要 n = 38。如果采用 40 个观测值，则 R^2_{adj} 的观测值在 ρ^2_{adj} 的 0.20 范围内的置信度为 90%（与此值的大小以及使用线性模型还是二次模型无关）。

附录 C：正态性

“协助”中使用的回归模型都采用此形式：

$$Y = f(X) + \varepsilon$$

有关随机项 ε 的典型假设是这样的：这些随机项都是带有均值零和公共方差 σ^2 的独立同分布的正态随机变量。β 参数的最小二乘估计量仍是最佳线性无偏估计量，即使我们放弃 ε 呈正态分布的假设也是如此。正态性假设仅在我们讨论这些估计值的概率分布时才变得重要，因为我们在进行关于 f(X) 的假设检验时必须要用到的。

我们想要根据正态性假设确定需要多大的 n 我们才能相信回归分析的结果。我们进行了模拟，以研究各种非正态错误分布情况下假设检验的 I 类错误率。

下面的表 4 显示了针对线性和二次模型的各种分布 ε，我们进行了 10,000 次模拟，在 $\alpha = 0.05$ 下整体 F 检验能得到效果显著结果的所占比率。在这些模拟中，X 和 Y 之间不存在关系的原假设是真的。X 在某个区间内是等间隔取值的。针对所有检验，我们使用了样本数量 n=15。

表 4 针对非正态分布，n=15 的线性模型和二次模型的整体 F 检验的 I 类错误率

分布	线性显著	二次显著
正态	0.04770	0.05060
t(3)	0.04670	0.05150
t(5)	0.04980	0.04540
Laplace	0.04800	0.04720
均匀	0.05140	0.04450
Beta(3, 3)	0.05100	0.05090
指数	0.04380	0.04880
Chi(3)	0.04860	0.05210
Chi(5)	0.04900	0.05260
Chi(10)	0.04970	0.05000
Beta(8, 1)	0.04780	0.04710

随后，我们检查了用于选择最佳模型的最高次项的检验。对于每次模拟，我们分析了二次项是否显著。对于二次项不显著的情况，我们分析了线性项是否显著。在这些模拟中，原假设为真，目标为 $\alpha = 0.05$ 且 n=15。

表 5　针对非正态分布，n=15 的线性或二次模型的最高次项检验的 I 类错误率

分布	平方	线性
正态	0.05050	0.04630
t(3)	0.05120	0.04300
t(5)	0.04710	0.04820
Laplace	0.04770	0.04660
均匀	0.04670	0.04900
Beta(3, 3)	0.05000	0.04860
指数	0.04600	0.03800
Chi(3)	0.05110	0.04290
Chi(5)	0.05290	0.04490
Chi(10)	0.04970	0.04610
Beta(8, 1)	0.04770	0.04380

模拟结果显示，对于整体 F 检验和模型中的最高次项检验，发现出现统计意义显著结果之概率对于任何错误分布差异不是很大。I 类错误率全部介于 0.038 和 0.0529 之间。