Regressão simples

Visão geral

O procedimento de regressão simples no Assistente ajusta modelos lineares e quadráticos com uma preditora contínua (X) e uma resposta contínua (Y), usando estimação por mínimos quadrados. O usuário pode selecionar o tipo de modelo ou permitir que o Assistente selecione o modelo de melhor ajuste. Neste artigo, explicamos os critérios que o Assistente usa para selecionar o modelo de regressão.

Além disso, nós examinamos vários fatores que são importantes para a obtenção de um modelo de regressão válido. Primeiramente, a amostra dever ser grande o suficiente para fornecer poder suficiente para o teste e para fornecer precisão para a estimativa de força do relacionamento entre X e Y. A seguir, é importante identificar dados incomuns que possam afetar os resultados da análise. Nós também consideramos a suposição de que o termo de erro siga uma distribuição normal e avaliamos o impacto da não normalidade nos testes de hipóteses do modelo geral e dos coeficientes. Finalmente, para garantir a utilidade do modelo, é importante que o tipo de modelo selecionado reflita com precisão a relação entre X e Y.

Com base nesses fatores, o assistente realiza automaticamente as seguintes verificações em seus dados e relata os resultados no Cartão de relatório:

- Quantidade de dados
- Dados atípicos
- Normalidade
- Ajuste do modelo

Neste artigo, investigamos como esses fatores se relacionam com a análise de regressão na prática e descrevemos como estabelecemos as orientações para verificar estes fatores no Assistente.

Métodos de regressão

Seleção do modelo

A análise de regressão no Assistente ajusta um modelo com uma preditora contínua e uma resposta contínua e pode ajustar dois tipos de modelos:

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

O usuário pode selecionar o modelo antes de realizar a análise ou pode permitir que o Assistente selecione o modelo. Existem vários métodos que podem ser utilizados para determinar qual o modelo mais adequado aos dados. Para garantir a utilidade do modelo, é importante que o tipo de modelo selecionado reflita com precisão a relação entre X e Y.

Objetivo

Desejávamos examinar os diferentes métodos que podem ser usados para a seleção do modelo a fim de determinar qual deles usar no Assistente.

Método

Examinamos três métodos que são normalmente utilizados para a seleção do modelo (Neter et al., 1996). O primeiro método identifica o modelo em que o termo da ordem mais alta seja significativo. O segundo método seleciona o modelo com o valor R^2_{aj} mais alto. O terceiro método seleciona o modelo em que o teste F global seja significativo. Para obter mais detalhes, consulte o Anexo A.

Para determinar a abordagem no Assistente, examinamos os métodos e comparamos os cálculos uns com os outros. Também coletamos os pareceres de especialistas em análise de qualidade.

Resultados

Com base em nossa pesquisa, decidimos usar o método que seleciona o modelo com base na significância estatística do termo de ordem mais alta no modelo. O Assistente primeiro examina o modelo quadrático e testa se o termo quadrado (β_2) no modelo é estatisticamente significativo. Se esse termo não for significativo, ele descarta o termo quadrático do modelo e testa o termo linear (β_1). O modelo selecionado por meio desta abordagem é apresentado no Relatório de Seleção de Modelos. Além disso, se o usuário selecionar um modelo diferente do selecionado pelo Assistente, "reportamos isso no Relatório de Seleção de Modelo e no Cartão de Relatório.

Em parte, escolhemos este método por causa dos pareceres de profissionais da qualidade que disseram que preferem modelos mais simples, que excluam termos que não sejam significativos. Além disso, com base em nossa comparação dos métodos, a utilização da significância estatística do termo mais alto no modelo é mais rigorosa do que o método que seleciona o modelo com base no valor mais alto de R^2_{aj}. Para obter mais detalhes, consulte o Anexo A.

Apesar de usarmos a significância estatística do termo mais alto no modelo para selecionar o modelo, apresentamos também o valor e o teste F geral para o modelo no Relatório de Seleção de Modelos. Para ver os indicadores de status apresentados no Cartão de Relatório, veja a seção Dados de verificação do ajuste do modelo abaixo.

Verificações dos dados

Quantidade de dados

O poder se relaciona ao quão provável um teste de hipóteses é de rejeitar a hipótese nula, quando esta é falsa. Para a regressão, a hipótese nula declara que não existe relacionamento entre X e Y. Se o conjunto de dados for pequeno demais, o poder do teste pode não ser adequado para detectar um relacionamento entre X e Y que realmente existe. Portanto, o conjunto de dados deve ser grande o suficiente para detectar, com alta probabilidade, um relacionamento importante em termos práticos.

Objetivo

Desejávamos determinar como a quantidade de dados que afeta o poder do teste F geral da relação entre X e Y e a precisão de R^2_{aj}, a estimativa da força da relação entre X e Y. Esta informação é crucial para determinar se o conjunto de dados é grande o suficiente para garantir que a força da relação observada nos dados seja um indicador confiável da verdadeira força subjacente da relação. Para mais informações sobre R^2_{aj}, consulte o Anexo A.

Método

Para examinar o poder do teste F geral, realizamos cálculos de poder para uma série de valores de R^2_{aj} e tamanhos de amostra. Para analisar a precisão de R^2_{aj}, simulamos a distribuição de R^2_{aj} para valores diferentes da população ajustado R^2 (ρ^2_{aj}) e diferentes tamanhos de amostra. Examinamos a variabilidade nos valores de R^2_{aj} para de terminar o tamanho que a amostra deve ter para que R^2_{aj} esteja perto de ρ^2_{aj}. Para mais informações sobre os cálculos e as simulações, consulte o Anexo B.

Resultados

Descobrimos que, para amostras moderadamente grandes, a regressão apresenta bom poder para detectar relações entre X e Y, mesmo que as relações não sejam fortes o suficiente para ser de interesse prático. Mais especificamente, verificou-se que:

- Com um tamanho de amostra de 15 e uma relação forte entre X e Y (ρ^2_{aj} = 0,65), a probabilidade de encontrar uma relação linear estatisticamente significativa é de 0,9969. Portanto, quando o teste não consegue encontrar uma relação estatisticamente significativa com 15 ou mais pontos de dados, é provável que o relacionamento verdadeiro não seja muito forte (ρ^2_{aj} valor <0,65).
- Com um tamanho de amostra de 40 e uma relação moderadamente fraca entre X e Y (ρ^2_{aj} = 0,25), a probabilidade de encontrar uma relação linear estatisticamente significativa é 0,9398. Portanto, com 40 pontos de dados, é provável que o teste F encontre relacionamentos entre X e Y, mesmo quando a relação é moderadamente fraca.

A regressão pode detectar relações entre X e Y com bastante facilidade. Portanto, se você encontrar uma relação estatisticamente significativa, deve também avaliar a força da relação usando R^2_{aj}. Descobrimos que, se o tamanho da amostra não for grande o suficiente, R^2_{aj} não é muito confiável e pode variar muito de uma amostra para outra. No entanto, com um tamanho de amostra de 40 ou mais, descobrimos que os valores de R^2_{aj} são mais estáveis e confiáveis. Com um tamanho de amostra de 40, você pode ter 90% de confiança de que o valor observado de R^2_{aj} ficará dentro de 0,20 de ρ^2_{aj}, independentemente do valor real e do tipo de modelo (linear ou quadrático). Para obter mais detalhes sobre os resultados das simulações, consulte o Anexo B.

Com base nestes resultados, o Assistente mostra as informações a seguir no Relatório de cartão:

Status	Condição
	**Tamanho amostral < 40** O tamanho de sua amostra não é grande o suficiente para fornecer uma estimativa precisa da força da relação. As medições da força da relação, como um R-Quadrado e R-Quadrado (ajustado), podem variar muito. Para obter uma estimativa mais precisa, devem ser usadas grandes quantidades de amostras (normalmente 40 ou mais). **Tamanho amostral ≥ 40** A amostra é grande o suficiente para obter uma estimativa precisa da força da relação.

Dados atípicos

No procedimento de Regressão do Assistente, nós definimos dados incomuns como observações com grandes resíduos padronizados ou valores altos de leverage. Estas medições normalmente são usadas para identificar dados atípicos na análise de regressão (Neter et al., 1996). Como os dados atípicos podem ter forte influência sobre os resultados, talvez seja necessário corrigir os dados para tornar a análise válida. Entretanto, os dados atípicos também resultam da variação natural no processo. Portanto, é importante identificar a causa do comportamento incomum para determinar como lidar com tais pontos de dados.

Objetivo

Desejamos determinar o quão grandes os resídulos padronizados e os valores de leverage devem ser para sinalizar que um ponto dos dados é atípico.

Método

Nós desenvolvemos nossas orientações para identificar observações atípicas com base no procedimento de Regressão padrão no Minitab (**Estat** > **Regressão** > **Regressão**).

Resultados

RESÍDUO PADRONIZADO

O resíduo padronizado é igual ao valor de um resíduo, e_i, dividido por uma estimativa de seu desvio padrão. Em geral, uma observação é considerada atípica se o valor absoluto do resíduo padronizado for maior do que 2. Entretanto, esta orientação é um tanto

conservadora. Espera-se que aproximadamente 5% de todas as observações atendam a este critério por acaso (se os erros forem normalmente distribuídos). Portanto, é importante investigar a causa do comportamento atípico para determinar se uma observação realmente é atípica.

VALOR DE LEVERAGE

Os valores de leverage estão relacionados somente ao valor de X de uma observação e não dependem do valor de Y. Uma observação é determinada como incomum se o valor de leverage for maior do que 3 vezes o número de coeficientes do modelo (p) dividido pelo número de observações (n). Novamente, este é um valor de corte comumente usado, embora alguns dos livros didáticos usem $\frac{2 \times p}{n}$ (Neter et al., 1996).

Se seus dados incluírem algum ponto de leverage alto, pense se eles exercem influência indevida sobre o tipo de modelo selecionado para o ajuste dos dados. Por exemplo um único valor extremo de X poderia resultar na seleção de um modelo quadrático em vez de um modelo linear. Você deve considerar se a curvatura observada no modelo quadrático é consistente com sua compreensão do processo. Em caso negativo, ajuste um modelo mais simples aos dados ou reúna dados adicionais para realizar uma investigação mais aprofundada no processo.

Quando procura por dados incomuns, o Cartão de Relatório do Assistente exibe os indicadores de status a seguir:

Status	Condição
	Não há pontos de dados atípicos. Os pontos de dados atípicos podem ter uma forte influência nos resultados.
	Existem pelo menos um ou mais resíduos padronizados ou pelo menos um ou mais valores de leverage altos. Você pode passar o cursor sobre um ponto ou usar o recurso da Função Brush do Minitab para identificar as linhas da worksheet. Como dados atípicos podem ter uma forte influência nos resultados, tente identificar a causa de sua natureza atípica. Corrija quaisquer erros de entrada de dados ou medições. Considere remover os dados que estão associados a causas especiais e refazer a análise.

Normalidade

Uma suposição típica na regressão é que erros aleatórios (ε) são normalmente distribuídos. A suposição de normalidade é importante quando são conduzidos testes de hipótese das estimativas dos coeficientes (β). Felizmente, mesmo quando erros aleatórios não são distribuídos normalmente, os resultados de teste normalmente são confiáveis quando a amostra é grande o suficiente.

Objetivo

Desejamos determinar o tamanho amostral necessário para o fornecimento de resultados confiáveis com base na distribuição normal. Desejamos determinar o quão próximo os resultados de teste reais corresponderam ao nível alvo de significância (alfa ou taxa de erro

tipo I) para o teste, ou seja, se o teste rejeitou incorretamente a hipótese nula com mais ou menos frequência do que era esperado para diferentes distribuições não normais.

Método

Para estimar a taxa de erro tipo I, realizamos várias simulações com distribuições assimétricas, caudas pesadas e caudas leves que se desviam substancialmente da distribuição normal. Conduzimos simulações para modelos lineares e quadráticos usando um tamnho de amostra de 15. Analisamos o teste F geral e o teste do termo de ordem mais alta no modelo.

Para cada condição, realizamos 10.000 testes. Geramos dados aleatórios de forma que, para cada teste, a hipótese nula seja verdadeira. Depois disso, realizamos os testes usando um nível de significância alvo de 0,05. Contamos o número de vezes entre os 10.000 que os testes realmente rejeitaram a hipótese nula e comparamos essa proporção com o nível de significância alvo. Se o teste se desempenha bem, as taxas de erro tipo I deveriam estar muito próximas do nível de significância de destino. Para obter mais informações sobre as simulações, consulte o Anexo C.

Resultados

Tanto para o teste F geral quanto para o teste do termo de ordem mais alta no modelo, a probabilidade de encontrar resultados estatisticamente significativos não difere substancialmente para nenhuma das distribuições não normais. As taxas de erro do tipo I estão todas entre 0,038 e 0,0529, muito próximas do nível de significância alvo de 0,05.

Como os testes se desempenham bem com amostras relativamente pequenas, o Assistente não testa os dados quanto à normalidade. Em vez disso, o Assistente verifica o tamanho da amostra e indica quando a amostra é menor do que 15. O Assistente exibe os indicadores de status a seguir no Cartão do relatório para regressão:

Status	Condição
	Os tamanhos amostrais são de pelo menos 15, logo a normalidade não é uma preocupação.
	Como o tamanho da amostra é inferior a 15, a normalidade pode ser um problema. Deve-se tomar cuidado ao interpretar o valor-p. Com amostras pequenas, a precisão do valor de p é sensível a erros residuais não normais.

Ajuste do modelo

É possível selecionar o modelo linear ou quadrático antes de realizar a análise de regressão ou você pode deixar que o Assistente selecione o modelo. Vários métodos podem ser utilizados para selecionar um modelo adequado.

Objetivo

Desejávamos examinar os diferentes métodos utilizados para selecionar um tipo de modelo para determinar qual abordagem deve ser utilizada no Assistente.

Método

Examinamos três métodos que são normalmente utilizados para a seleção do modelo. O primeiro método identifica o modelo em que o termo da ordem mais alta seja significativo. O segundo método seleciona o modelo com o valor R^2_{aj} mais alto. O terceiro método seleciona o modelo em que o teste F global seja significativo. Para obter mais detalhes, consulte o Anexo A.

Para determinar a abordagem usada no Assistente, examinamos os métodos e como os seus cálculos se comparam. Também coletamos os pareceres de especialistas em análise de qualidade.

Resultados

Decidimos usar o método que seleciona o modelo com base na significância estatística do termo de ordem mais alta no modelo. O Assistente primeiro examina o modelo quadrático e testa se o termo quadrado no modelo (β_3) é estatisticamente significativo. Se o termo não for significativo, então ele testa o termo linear (β_1) no modelo linear. O modelo selecionado por meio desta abordagem é apresentado no Relatório de Seleção de Modelos. Além disso, se o usuário selecionar um modelo que diferente do selecionado pelo Assistente, informamos esse fato no Relatório de Seleção de Modelo e no Cartão de Relatório. Para obter mais informações, consulte a seção Método de regressão acima.

Com base em nossos resultados, o Assistente do Cartão de Relatório exibe o seguinte indicador de status:

Status	Condição
	**Se o modelo do usuário corresponde ao modelo de melhor ajuste do Assistente** Você deve avaliar os dados e ajuste do modelo em termos de seus objetivos. Observe os gráficos de linha ajustada para ter certeza de que: • A amostra cobre adequadamente o intervalo de valores de X. • O modelo ajusta adequadamente qualquer curvatura nos dados (evite o super-ajuste) • A linha se ajusta bem a qualquer área de interesse especial **Se o modelo do usuário não corresponde ao modelo de melhor ajuste do Assistente** O Relatório de Seleção de Modelo exibe um modelo alternativo que pode ser uma escolha melhor.

Referências

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

Anexo A: Seleção do modelo

Um modelo de regressão relacionando uma preditora X a uma resposta Y é da forma:

$Y = f(X) + \varepsilon$

em que a função f(X) representa o valor esperado (média) de Y dado X..

No Assistente, há duas escolhas para a forma da função f(X):

Tipo do modelo	f(X)
Linear	$\beta_0 + \beta_1 X$
Quadrático	$\beta_0 + \beta_1 X + \beta_2 X^2$

Os valores dos coeficientes β são desconhecidos e devem ser estimados a partir dos dados. O método de estimação é o de mínimos quadrados, que minimiza a soma dos resíduos quadráticos na amostra:

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

Um resíduo é a diferença entre a resposta observada Y_i e o valor ajustado $\hat{f}(X_i)$ com base nos coeficientes estimados. O valor minimizado desta soma de quadrados é o SEQ (soma de quadrados do erro) para um determinado modelo.

Para determinar o método usado no Assistente para selecionar o tipo de modelo, avaliamos três opções:

- Significância do termo de ordem mais alta no modelo
- O teste F geral do modelo
- Valor de R^2 ajustado (R^2_{aj})

Significância do termo de ordem mais alta no modelo

Nesta abordagem, o Assistente começa com o modelo quadrático. O Assistente testa as hipóteses para o termo quadrado no modelo quadrático:

$H_0: \beta_2 = 0$

$H_1: \beta_2 \neq 0$

Se esta hipótese nula for rejeitada, o Assistente conclui que o coeficiente do termo quadrado é diferente de zero e seleciona o modelo quadrático. Caso contrário, o assistente testa as hipóteses para o modelo linear:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

Teste F geral

Este método é um teste do modelo geral (linear ou quadrático). Para a forma selecionada da função de regressão f(X), ele testa:

$H_0: f(X)$ é constante

$H_1: f(X)$ não é constante

Ajustado R^2

Ajustado R^2 (R^2_{aj}) mede o quando da variabilidade na resposta é atribuída a X pelo modelo. Existem duas maneiras comuns de medir a força do relacionamento observado entre X e Y:

$$R^2 = 1 - \frac{\text{SEQ}}{STQ}$$

E

$$R^2_{aj} = 1 - \frac{\text{SEQ}/(n-p)}{STQ/(n-1)}$$

Em que

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

O STQ é a soma de quadrados total, que mede a variação das respostas sobre sua média geral $\bar{Y}$ O SEQ mede sua variação sobre a função de regressão f(X). O ajuste em R^2_{aj} é para o número de coeficientes (p) no modelo completo, o que deixa n – p graus de liberdade para estimar a variância de ε. R^2 nunca diminui quando são adicionais mais coeficientes ao modelo; entretanto, devido ao ajuste, R^2_{aj} pode diminuir quando coeficientes adicionais não melhoram o modelo. Portanto, se adicionar outro termo ao modelo não explicar nenhuma variância adicional na resposta, R^2_{aj} diminui, indicando que o termo adicional não é util. Portanto, a medida ajustada deve ser usada para comparar o linear e o quadrático.

Relacionamento entre os métodos de seleção do modelo

Desejávamos examinar a relação entre os três métodos de seleção de modelo, como eles são calculados e como eles afetam um ao outro.

Primeiro, observamos a relação entre a forma como o teste F geral e R^2_{aj} são calculados. A estatística F do teste do modelo geral pode ser expressa em termos de SEQ e STQ, que também são usados no cálculo de R^2_{aj}:

$$F = \frac{(STQ - SEQ)/(p-1)}{SEQ/(n-p)}$$

$$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R_{aj}^2}{1-R_{aj}^2}.$$

As fórmulas acima mostram que a estatística F é uma função crescente de R_{aj}^2. Portanto, o teste rejeita H_0 se, e somente se, R_{aj}^2 exceder a um valor específico determinado pelo nível de significância (α) do teste. Para ilustrar isso, calculamos o mínimo necessário para obter a significância estatística do modelo quadrático em $\alpha = 0{,}05$ para os tamanhos de amostras diferentes apresentados na Tabela 1 abaixo. Por exemplo, com n = 15, o valor de R_{aj}^2 para o modelo deve ser de pelo menos 0,291877 para o teste F geral ser estatisticamente significativo.

Tabela 1 O mínimo de R_{aj}^2 para um para um teste F global para o modelo quadrático em $\alpha = 0{,}05$ em vários tamanhos de amostra

Tamanho amostral	**Mínimo de R_{aj}^2**
4	0,9925
5	0,90
6	0,773799
7	0,66459
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

Tamanho amostral	Mínimo de R^2_{aj}
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
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

Tamanho amostral	Mínimo de R^2_{aj}
49	0,083955
50	0,082227

Em seguida, examinamos a relação entre o teste de hipótese do termo de ordem mais alta em um modelo, e R^2_{aj}. O teste para o termo de ordem mais alta, como o termo quadrado em um modelo quadrático, pode ser expresso em termos das somas dos quadrados ou o R^2_{aj} do modelo completo (por exemplo, quadrático) e do R^2_{aj} do modelo reduzido (por exemplo, linear):

$$F = \frac{\text{SEQ}(Reduzido) - \text{SEQ}(Completo)}{\text{SEQ}(Completo)/(n-p)}$$

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

As fórmulas mostram que, para um valor fixo de $R^2_{aj}(Reduzido)$, a estatística F é uma função crescente da$R^2_{aj}(Completo)$. Elas também mostram como a estatística de teste depende da diferença entre os dois valores de R^2_{aj}. Em particular, o valor do modelo completo deve ser maior do que o valor para o modelo reduzido para que se possa obter um valor F grande o suficiente para ser estatisticamente significativo. Assim, o método que utiliza a significância do termo de ordem mais alta para selecionar o melhor modelo é mais rigoroso do que o método que escolhe o modelo com o R^2_{aj} mais elevado. O método de termo ordem mais alta também é compatível com a preferência de muitos usuários por um modelo mais simples. Dessa forma, decidimos usar a significância estatística do termo de ordem mais alta para selecionar o modelo no Assistente.

Alguns usuários estão mais inclinados a escolher o modelo que melhor ajusta os dados; ou seja, o modelo com R^2_{aj} mais elevado. O Assistente fornece estes valores no Relatório de Seleção de Modelos e no Cartão de Relatório.

Anexo B: Quantidade de dados

Nesta seção, consideramos como n, o número de observações, afeta o poder do teste geral do modelo e a precisão de R^2_{aj}, a estimativa de força do modelo.

Para quantificar a força do relacionamento, nós introduzimos uma nova quantidade, ρ^2_{aj}, como a contrapartida populacional da estatística amostral R^2_{aj}. Lembre-se de que

$$R^2_{aj} = 1 - \frac{\text{SEQ}/(n-p)}{STQ/(n-1)}$$

Portanto, definimos

$$\rho^2_{aj} = 1 - \frac{E(\text{SEQ}|X)/(n-p)}{E(STQ|X)/(n-1)}$$

O operador E(·|X) denota o valor esperado (ou a média) de uma variável aleatória, determinado o valor de X. Supondo-se que o modelo correto seja $Y = f(X) + \varepsilon$ com ε independente identicamente distribuído, nós temos

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

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

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

Por isso,

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

Significância do modelo geral

Quando o testamos a significância estatística do modelo geral, assumimos que os erros aleatórios ε são independentes e normalmente distribuídos. Então, mediante a hipótese nula que a média de Y é constante ($f(X) = \beta_0$), a a estatística de teste F tem uma distribuição $F(p-1, n-p)$. Sob a hipótese alternativa, a estatística F tem uma distribuição $F(p-1, n-p, \theta)$ não central com parâmetro de não centralidade:

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

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

A probabilidade de rejeitar H_0 aumenta com o parâmetro de não centralidade, que está aumentando em n e ρ^2_{aj}.

Com a fórmula acima, calculamos o poder dos testes F gerais para uma série de valores de ρ_{aj}^2 quando n = 15 para os modelos linear e quadrático. Consulte a Tabela 2 para ver os resultados.

Tabela 2 Poder para os modelos linear e quadrático com valores de ρ_{aj}^2 diferentes com n=15

ρ_{aj}^2	θ	**Poder de F Linear**	**Poder de F Quadrático**
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

Em geral, descobrimos que o teste tem alto poder quando a relação entre X e Y é forte e o tamanho da amostra é de pelo menos 15. Por exemplo, quando ρ_{aj}^2 = 0,65, a Tabela 2 mostra que a probabilidade de encontrar uma relação linear estatisticamente significativa em $\alpha = 0{,}05$ é de 0,99688 . A falha na detecção de uma relação tão forte com o teste F ocorreria em menos de 0,5% das amostras. Mesmo para um modelo quadrático, a incapacidade de detectar a relação com o teste F ocorreria em menos de 2% das amostras. Portanto, quando

o teste não consegue encontrar uma relação estatisticamente significativa com 15 observações ou mais, é uma boa indicação de que a verdadeira relação, se realmente existir uma, tem um valor de ρ_{aj}^2 inferior a 0,65. Observe que ρ_{aj}^2 não precisa ser tão grande como 0,65 para apresentar interesse prático.

Também desejávamos analisar o poder do teste F global quando o tamanho da amostra fosse maior (n = 40). Nós determinamos que o tamanho da amostra n = 40 é um limite importante para a precisão do R_{aj}^2 (consulte Força da relação abaixo) e pretendíamos avaliar os valores de poder para o tamanho da amostra. Calculamos o poder dos testes F gerais para uma série de valores quando n = 40 para os modelos linear e quadrático. Consulte a Tabela 3 para ver os resultados.

Tabela 3 Poder para os modelos linear e quadrático com valores de ρ_{aj}^2 diferentes com n=40

ρ_{aj}^2	θ	**Poder de F Linear**	**Poder de F Quadrático**
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

Descobrimos que o poder foi elevado, mesmo quando a relação entre X e Y era moderadamente fraca. Por exemplo, mesmo quando ρ_{aj}^2 = 0,25, a Tabela 3 mostra que a probabilidade de encontrar uma relação linear estatisticamente significativa na $\alpha = 0{,}05$ é de 0,93980. Com 40 observações, é improvável que o teste F não consiga detectar uma relação entre X e Y, mesmo que essa relação seja moderadamente fraca.

Força da relação

Como já mostramos, um relacionamento estatisticamente significativo nos dados não indica necessariamente um relacionamento subjacente forte entre X e Y. Esta é a razão pela qual muitos usuários olham indicadores como R^2_{aj} para dizer o quanto o relacionamento é realmente forte. Se considerarmos R^2_{aj} como uma estimativa de ρ^2_{aj}, desejamos ter a confiança de que a estimativa está razoavelmente próxima do valor verdadeiro de ρ^2_{aj}.

Para ilustrar a relação entre R^2_{aj} e ρ^2_{aj}, simulamos a distribuição de R^2_{aj} para valores diferentes de ρ^2_{aj} para observar como a R^2_{aj} variável se comporta para valores de n diferentes. Os gráficos das Figuras 1-4 abaixo mostram histogramas de 10.000 valores simulados de R^2_{aj}. Em cada par de histogramas, o valor de ρ^2_{aj} é o mesmo, de modo que podemos comparar a variabilidade de R^2_{aj} para tamanhos de amostras de 15 com amostras de tamanho 40. Testamos os valores de ρ^2_{aj} de 0,0, 0,30, 0,60 e 0,90. Todas as simulações foram feitas com o modelo linear.

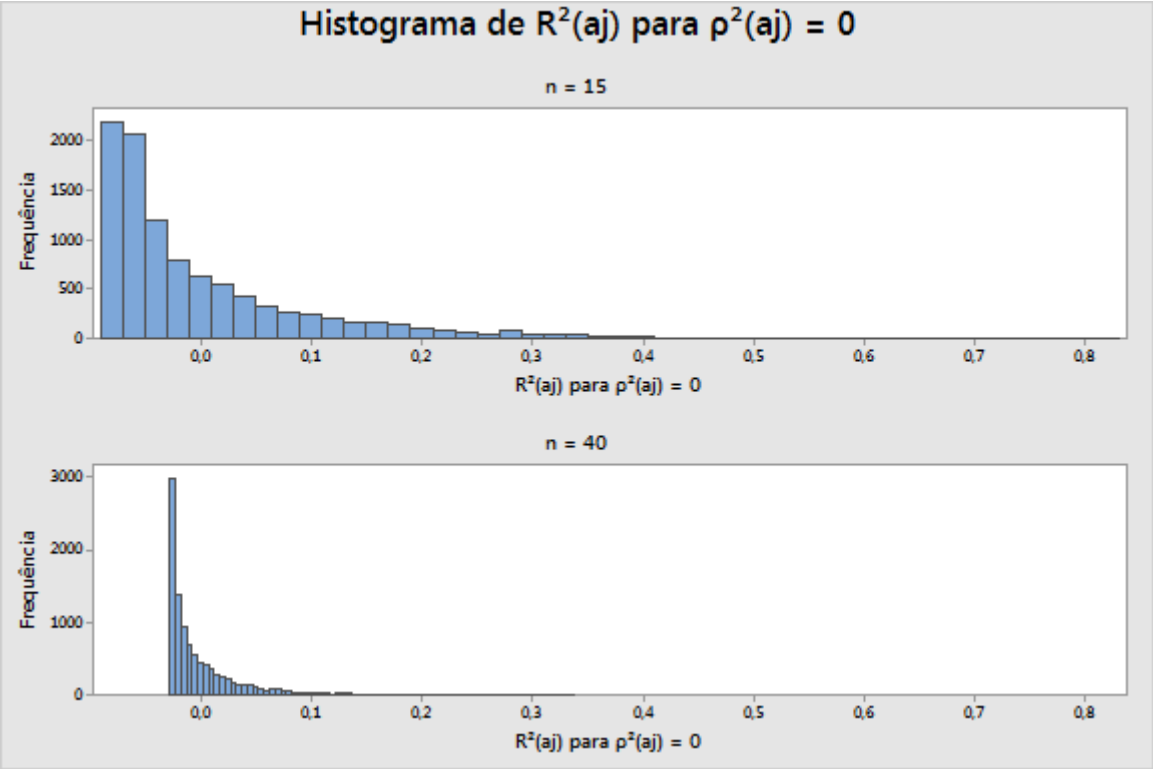

Figura 1 Valores de R^2_{aj} simulados para ρ^2_{aj} = 0,0 para n=15 e n=40

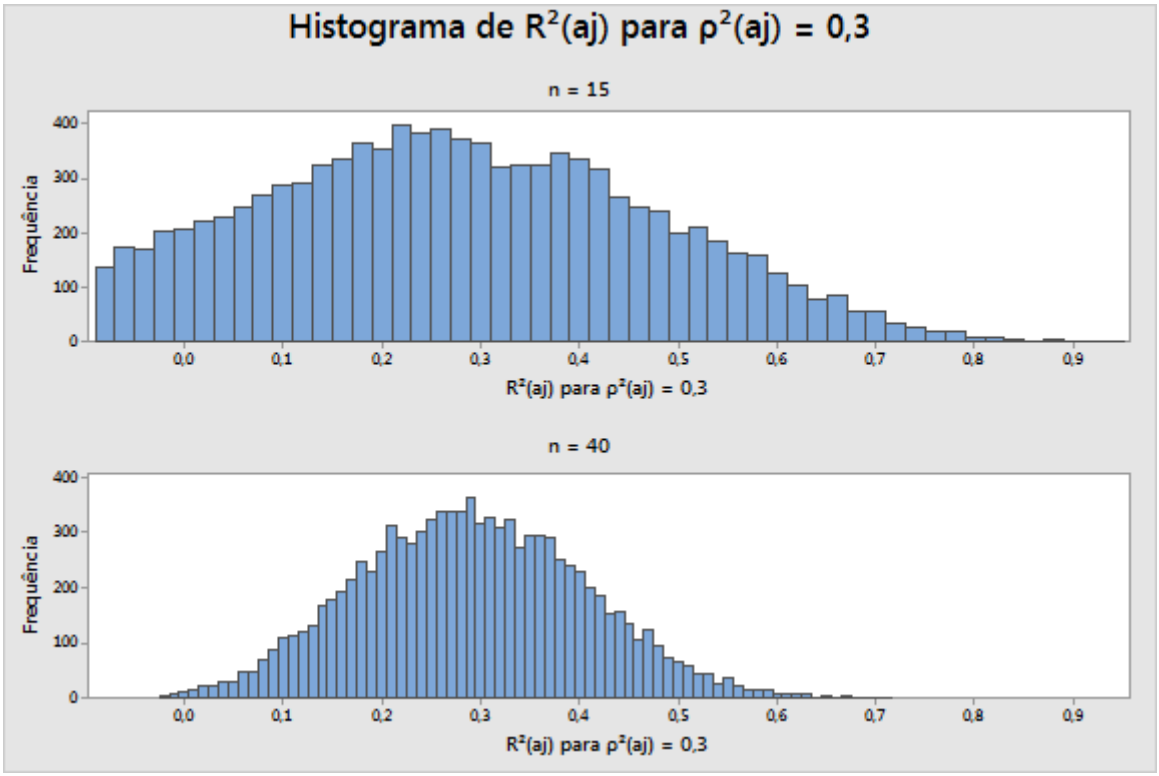

Figura 2 Valores de R^2_{aj} simulados para ρ^2_{aj} = 0,30 para n=15 e n=40

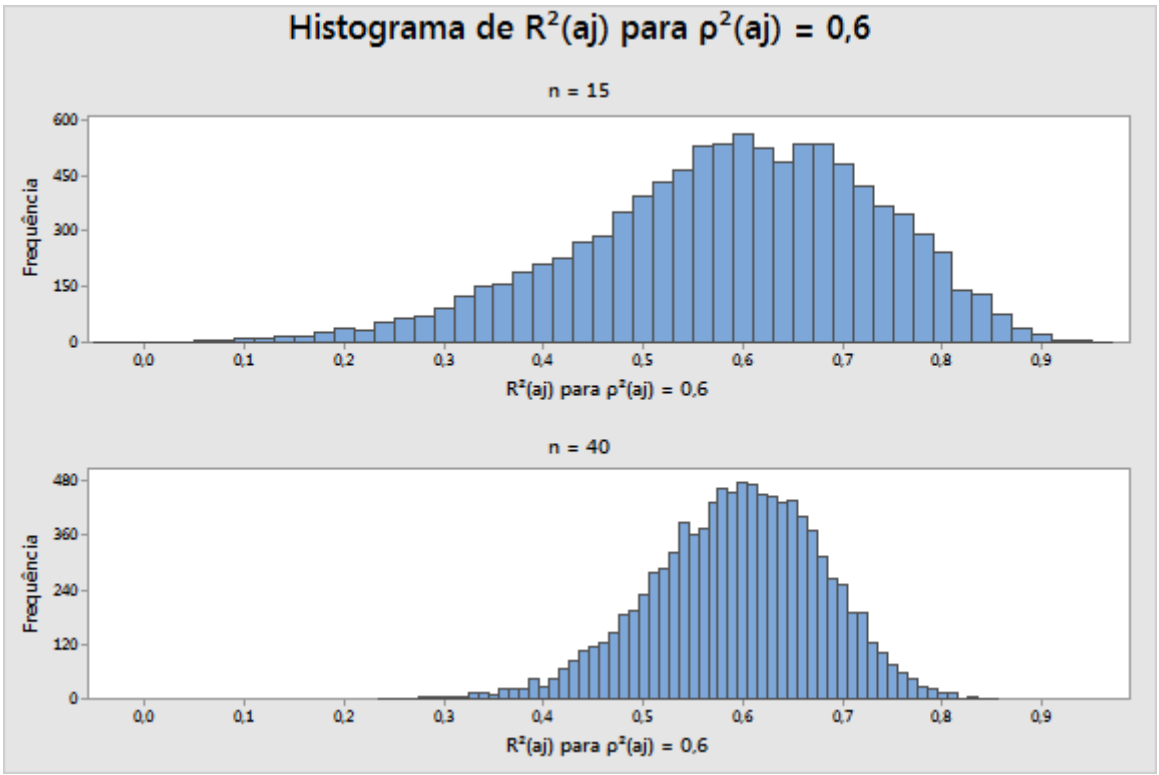

Figura 3 Valores de R^2_{aj} simulados para ρ^2_{aj} = 0,60 para n=15 e n=40

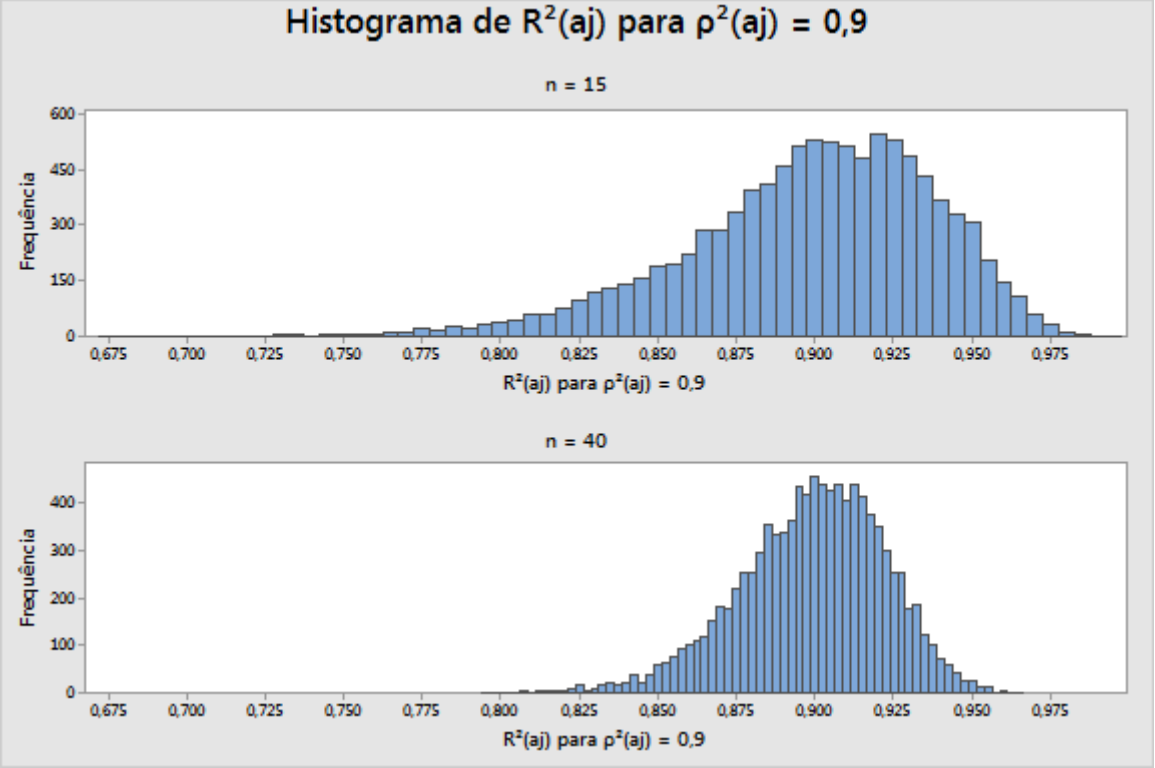

Figura 4 Valores de R^2_{aj} simulados para ρ^2_{aj} = 0,90 para n=15 e n=40

Em geral, as simulações mostram que pode haver uma diferença considerável entre a força real da relação (ρ^2_{aj}) e a relação observada nos dados (R^2_{aj}). Aumentar o tamanho de amostra de 15 a 40 reduz grandemente a magnitude provável da diferença. Determinamos que 40 observações é um limite apropriado por meio da identificação do valor mínimo de n para o qual as diferenças absolutas $|R^2_{aj} - \rho^2_{aj}|$ maior do que 0,20 ocorrem com uma probabilidade não maior do que 10%. Isso acontece independentemente do valor real de ρ^2_{aj} em qualquer um dos modelos considerados. Para o modelo linear, o caso mais difícil foi ρ^2_{aj} = 0,31, que exigiu n = 36. Para o modelo quadrático, o caso mais difícil foi ρ^2_{aj} = 0,30, o que exigiu n = 38. Com 40 observações, é possível ter 90% de confiança de que o valor observado estará dentro de 0,20, independentemente de o que esse valor seja e de qual modelo está sendo utilizado (linear ou quadrático).

Anexo C: Normalidade

Os modelos de regressão usados no Assistente são todos da forma:

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

A suposição típica em relação aos termos aleatórios ε é que eles são variáveis aleatórias normais distribuídas de maneira independente e idêntica com média zero e σ^2 de variância comum. As estimativas de mínimos quadrados dos parâmetros de β ainda são as melhores estimativas não viciadas, mesmo se deixarmos de lado a suposição de que ε sejam normalmente distribuídos. A suposição de normalidade somente se torna importante quando tentamos fixar probabilidades a estas estimativas, como fazemos nos testes de hipótese sobre f(X).

Desejamos determinar o tamanho que n precisa ter para que possamos confiar nos resultados de uma análise de regressão com base na suposição de normalidade. Realizamos simulações para explorar as taxas de erro tipo I dos testes de hipótese mediante uma variedade de distribuições de erro não normais.

A Tabela 4 a seguir mostra a porcentagem de 10.000 simulações em que o teste F geral foi significativo em $\alpha = 0{,}05$ para várias distribuições de ε para os modelos linear e quadrático. Nestas simulações, a hipótese nula, que declara que não existe relacionamento entre X e Y, foi verdadeira. Os valores de X foram espaçados uniformemente ao longo de um intervalo. Usamos um tamanho amostral de n=15 para todos os testes.

Tabela 1 As taxas de erro tipo I para testes F gerais para modelos linear e quadrático com n=15 para distribuições não normais

Distribuição	Linear significativo	Quadrático significativo
Normal	0,04770	0,05060
t(3)	0,04670	0,05150
t(5)	0,04980	0,04540
Laplace	0,04800	0,04720
Uniforme	0,05140	0,04450
Beta(3, 3)	0,05100	0,05090
Exponencial	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

Em seguida, analisamos o teste do termo de ordem mais alta usado para selecionar o melhor modelo. Para cada simulação, consideramos se o termo quadrado era significativo. Para os casos em que o termo quadrado não era significativo, consideramos se o termo linear era significativo. Nestas simulações, a hipótese nula era verdadeira, $\alpha = 0{,}05$ do alvo e n = 15.

Tabela 5 As taxas de erro do tipo I para os testes de termo de ordem mais alta para os modelos linear e quadrático com n=15 para distribuições não normais

Distribuição	Quadrado	Linear
Normal	0,05050	0,04630
t(3)	0,05120	0,04300
t(5)	0,04710	0,04820
Laplace	0,04770	0,04660
Uniforme	0,04670	0,04900
Beta(3, 3)	0,05000	0,04860
Exponencial	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

Os resultados da simulação mostram que, para o teste F geral e o teste do termo de ordem mais alta no modelo, a probabilidade de chegar a resultados estatisticamente significativos não difere substancialmente de nenhuma distribuição de erro. As taxas de erro tipo I estão todas entre 0,038 e 0,0529.