Este artigo é parte de uma série de artigos que explicam a pesquisa conduzida pelos estatísticos do Minitab para desenvolver os métodos e verificações de dados usados no Assistente no Minitab Statistical Software.

Regressão múltipla

Visão geral

O procedimento de regressão múltipla nos modelos lineares e quadráticos de ajuste do Assistente com até cinco preditoras (X) e uma resposta contínua (Y) usando a estimativa dos mínimos quadrados. O usuário seleciona o tipo de modelo e o Assistente seleciona os termos do modelo. 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, o exemplo mostra extensão suficiente para fornecer potência 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 hipótese do modelo geral.

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 incomuns
- Normalidade

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 se ajusta a uma resposta contínua e de duas a cinco preditoras. Uma das preditoras pode ser categórica. Existem dois tipos de modelos para escolher:

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

O Assistente seleciona os termos do modelo no modelo linear ou quadrático completo.

Objetivo

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

Método

Examinamos os três tipos de seleção de modelo: para trás, para frente e stepwise. Estes tipos de seleção de modelo incluem várias opções que também examinamos, incluindo:

- Os critérios usados para inserir ou remover termos do modelo.
- Seja para forçar determinados termos no modelo ou para incluir determinados termos no modelo inicial.
- A hierarquia dos modelos.
- Padronização das variáveis X no modelo.

Nós revisamos estas opções, observamos seu efeitos no resultado do procedimento e consideramos quais seriam os métodos preferíveis pelos profissionais.

Resultados

O procedimento que usamos para selecionar os termos do modelo no Assistente é o seguinte:

- É usada a seleção de modelo stepwise. Com frequência, um conjunto de variáveis X potenciais são correlacionadas, de forma que o efeito de um termo dependerá de quais outros termos também estão no modelos. A seleção stepwise é, possivelmente, a melhor abordagem sob esta condição, porque ela permite que os termos sejam inseridos em uma única etapa, mas que sejam removidos posteriormente, dependendo de quais termos sejam incluídos no modelo.
- A hierarquia do modelo é mantida em cada etapa e vários termos podem entrar no modelo na mesma etapa. Por exemplo, se o termo mais significativo for X_1^2, então, ele é inserido junto com X_1, independentemente de X_1 ser significativo. A hierarquia é desejável porque permite que o modelos seja traduzido de unidades padronizadas para não padronizadas. E, como a hierarquia permite que vários termos entrem no

modelo em qualquer etapa, é possível identificar um quadrado ou um termo de interação importante, mesmo se o termo linear associado não esteja fortemente relacionado à resposta.

- Os termos são inseridos ou removidos do modelo com base em $\alpha = 0{,}10$. A utilização de $\alpha = 0{,}10$ torna o procedimento mais seletivo do que o procedimento stepwise no núcleo do Minitab, que usa $\alpha = 0{,}15$.
- Para fins de seleção dos termos do modelo, as preditoras são padronizadas subtraindo-se a média e dividindo pelo desvio padrão. O modelo final é exibido em unidades dos Xs não padronizados. A padronização de X remove a maioria da correlação entre os termos lineares e quadrados, o que reduz as chances de adicionar termos de ordem mais alta desnecessariamente.

Verificações dos dados

Quantidade de dados

O poder está preocupado com o grau de probabilidade apresentado por um teste de hipótese para rejeitar uma hipótese nula, quando for falso. 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 reacionamento importante em termos práticos.

Objetivo

Desejamos determinar como a quantidade de dados afeta o poder do teste F geral do relacionamento entre X e Y e a precisão de R^2_{aj}, a estimativa de força do relacionamento entre X e Y. Esta informação é essencial para determinar se o conjunto de dados é grande o suficiente para oferecer confiança na força de que o relacionamento observado nos dados seja um indicador confiável da verdadeira força subjacente do relacionamento. Para obter mais informações sobre R^2_{aj}, Consulte o Anexo A.

Método

Usamos uma abordagem semelhante para determinar o tamanho amostral recomendado que usamos para uma regressão simples. 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}. Também confirmamos que o tamanho amostral recomendado forneceu o poder razoável mesmo quando a força do relacionamento entre as variáveis X e Y é moderadamente fraco. Para obter mais informações sobre os cálculos, consulte o Anexo B.

Resultados

Como com a regressão simples, nós recomendamos uma amostra grande o suficiente para que você possa ter 90% de que o valor observado de R^2_{aj} estará dentro de 0,20 de ρ^2_{aj}. Descobrimos que o tamanho amostral necessário aumenta conforme são adicionados mais termos ao modelo. Portanto, calculamos o tamanho amostral necessário para cada tamanho de modelo. O tamanho recomendado é arredondado para o múltiplo de 5 mais próximo. Por exemplo, se o modelo tem oito coeficientes além da constando, como quatro termos lineares, três termos de interação e um termo quadrado, o tamanho amostral mínimo necessário para atender a esses critérios é n = 49. O Assistente arredonda esse valor para um tamanho amostral recomendado de n = 50. Para obter mais informações sobre as recomendações de tamanho amostral específico com base no número de termos, consulte o Anexo B.

Também nos certificamos de que os tamanhos amostrais recomendados ofereciam poder bom o suficiente. Descobrimos que, para relacionamentos moderadamente fracos, ρ^2_{aj}= 0,25, o poder normalmente fica em torno de 80% ou mais. Portanto, seguir as recomendações do

Assistente para tamanho amostral garante que você tem poder razoavelmente bom e boa precisão na estimativa de força do relacionamento.

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

Status	Condição
	**Tamanho amostral < recomendado** O tamanho amostral não é grande o suficiente para fornecer uma estimativa muito precisa da força do relacionamento. As medições de força do relacionamento, como Raiz quadrada e Raiz quadrada (ajustada), podem variar enormemente. Para obter uma estimativa precisa, amostras maiores devem ser usadas por um modelo deste tamanho. **Tamanho amostral >= recomendado** O tamanho amostral é grande o suficiente para obter uma estimativa precisa da força do relacionamento.

Dados incomuns

No procedimento de Regressão do Assistente, nós definimos dados incomum como observações com grandes resíduos de padronização ou valores de leverage maiores. Estas medições normalmente são usadas para identificar dados incomuns na análise de regressão (Neter et al., 1996). Como os dados incomuns podem ter forte influência sobre os resultados, talvez seja necessário corrigir os dados para tornar a análise válida. Entretanto, os dados incomuns 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 tamanho necessário para que os resíduos padronizados e os valores de leverage indiquem que um ponto de dados é incomum.

Método

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

Resultados

RESÍDUO PADRONIZADO

O resíduo padronizado é equalizado ao valor de um resíduo, e_i, dividido por uma estimativa de seu desvio padrão. Em geral, uma observação é considerada incomum 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 incomum para determinar se uma observação realmente é incomum.

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 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 incomuns..
	Existem pelo menos um ou mais resíduos padronizados ou pelo menos um ou mais pontos de leverage altos.

Normalidade

Uma suposição típica na regressão é que erros aleatórios (ε) são distribuídos normalmente. A suposição de anormalidade 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 de destino 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 distribuições não normais diferentes não normais.

Método

Para estimar a taxa de erro tipo I, realizamos várias simulações com distribuições assimétricas, caudas pesadas e caudas leves que partiram substancialmente da distribuição

normal. Conduzimos simulações usando um tamanho amostral de 15. Examinamos os teste F geral para vários modelos.

Para cada condição, nós realizamos 10.000. Geramos dados aleatórios de forma que, para cada teste, a hipótese nula seja verdadeira. Depois disso, realizamos os testes usando um nível de significância de destino de 0,10. Contamos o número de vezes entre os 10.000 que os testes realmente rejeitaram a hipótese nula e comparamos essa proporção como o nível de significância de destino. Se o teste corretamente realizado, 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

Para ambos os testes F gerais, a probabilidade de chegar a resultados estatisticamente significativos não difere substancialmente de nenhuma distribuição não normal. As taxas de erro do tipo I estão todas entre 0,08820 e 0,11850, razoavelmente próximas do nível de significância de destino de 0,10.

Como os testes foram corretamente realizados 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, de forma que a normalidade não é uma preocupação.
	Como o tamanho de amostra é mentor do que 15, a normalidade pode ser uma preocupação. Deve-se ter cautela ao interpretar o valor p. Com amostras pequenas, a precisão do valor p é sensível a erros de resíduo não normais.

Referências

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

Anexo A: Modelo e estatística

Um modelo de regressão relacionado a uma preditora X para uma resposta Y está na forma:

$Y = f(X) + \varepsilon$

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

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

Tipo do modelo	f(X)
Linear	$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$
Quadrático	$\beta_0 + \sum_i \beta_i X_i + \sum_i \beta_{ii} X_i^2 + \sum_{i<j} \beta_{ij} X_i X_j$

Os valores dos coeficientes β são conhecidos e devem ser estimados a partir dos dados. O método de estimativa é dos quadrados mínimos, o que minimiza a soma dos resíduos dos quadrados 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 dos erros dos quadrados) para um determinado modelo.

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

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

O STQ é a soma total dos quadrados, 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_{aj}^2 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 mais coeficientes são adicionados ao modelo. Entretanto, devido ao ajuste, R_{aj}^2 pode diminuir quando coeficientes adicionais não melhoram o modelo. Portanto, se adicionar outro termo ao modelo não explicar nenhuma variância adicional na resposta, R_{aj}^2 diminui, indicando que o termo adicional não é útil. Portanto, a medição ajustada deve ser usada para comparar modelos de tamanhos diferentes.

Relacionamento entre o teste F e $\boldsymbol{R_{aj}^2}$

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_{aj}^2:

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

$= 1 + \left(\frac{n-1}{p-1}\right)\frac{R_{aj}^2}{1-R_{aj}^2}$. As fórmulas acima mostram que a estatística de F são uma função aumentada 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.

Anexo B: Quantidade de dados

Nesta seção, consideramos como n, o número de observações, afeta o poder de todo o teste do modelo geral 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 da população da estatística de amostra R^2_{aj}. Lembre-se de que

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

Portanto, definimos

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

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

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

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

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 ε dos erros aleatórios são independentes e normalmente distribuídos. Então, mediante a hipótese nula que a média de Y é constante ($f(X) = \beta_0$), a estatística do teste F tem uma distribuição $F(p-1, n-p)$. Dante da 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 do H_0 de rejeição aumenta com o parâmetro de não centralidade, que está aumentando em n e ρ^2_{aj}.

Força da relação

Como mostramos para uma regressão simples, 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 no valor verdadeiro de ρ^2_{aj}.

Para cada tamanho de modelo possível, nós determinamos um limite apropriado para o tamanho amostral aceitável por meio da identificação do valor mínimo de n para o qual as diferenças $|R^2_{aj} - \rho^2_{aj}|$ maiores do que 0,20 ocorrem em uma probabilidade não maior do que 10%. Isso é independente do valor verdadeiro de ρ^2_{aj}. Os tamanhos amostrais recomendados n(T) são resumidos na tabela abaixo, em que T é o número de coeficientes no modelo em vez do coeficiente da constante.

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

T	n(T)
60-63	130
64-67	135
68-70	140
71-73	145

Nós avaliamos o poder do teste F geral do modelo para um valor moderadamente fraco de $\rho_{aj}^2 = 0{,}25$ a fim de confirmar que existe poder suficiente nos tamanhos amostrais recomendados. Os tamanhos do modelo na tabela abaixo representam o pior caso para cada valor de n(T). Modelos menores com o mesmo n(T) terão poder maior.

T	n(T)	Poder em $\boldsymbol{\rho_{aj}^2 = 0{,}25}$
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

T	n(T)	Poder em $\rho_{aj}^2 = 0,25$
67	135	0,799906
69	140	0,814664

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 dos 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 de tipo I dos testes de hipótese mediante uma variedade de distribuições de erro não normais.

A tabela 1 abaixo mostra a proporção de 10.000 simulações nas quais o teste F geral foi significativo a $\alpha = 0{,}10$ para várias distribuições de ε para três modelos diferentes. Nestas simulações, a hipótese nula, que declara que não existe relacionamento entre X e Y, foi verdadeira. Os valores de X foram gerados como variáveis normais multivariadas pelo comando RANDOM do Minitab. Usamos um tamanho amostral de n=15 para todos os testes. Todos os modelos envolviam preditoras contínuas. O primeiro modelo era o modelo linear com cinco variáveis X. O segundo modelo tinha todos os termos lineares e quadrados. O terceiro modelo tinha todos os termos lineares e sete das interações de 2 fatores.

Tabela 1 As taxas de erro do tipo I para testes F gerais com n=15 para distribuições não normais

Distribuição	Linear	Linear + quadrado	Linear + 7 interações
Normal	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
Uniforme	0,10630	0,10080	0,09480
Beta(3, 3)	0,09980	0,10120	0,10020
Exponencial	0,08820	0,09500	0,09960
Chi(3)	0,09890	0,114	0,10970
Chi(5)	0,09730	0,10590	0,10330

Distribuição	Linear	Linear + quadrado	Linear + 7 interações
Chi(10)	0,10150	0,09930	0,10360
Beta(8, 1)	0,09870	0,10230	0,10490

Os resultados de simulação mostram que a probabilidade de se obter resultados estatisticamente significativos não difere substancialmente do valor nominal de 0,10 para nenhuma das distribuições de erro. As taxas de erro tipo I observadas estão todas entre 0,08820 e 0,11850.