PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Supervised Machine Learning: Análise de Regressão Simples e Múltipla III
Data: 10/08/2021

Marcos Antonio Lopes Freixo Filho
Professor, em termos de procedimento, sei que são métodos diferentes, mas em essência, qual a diferença crucial entre padronização e normalização?
A padronização serve para colocar numa mesma escala as variáveis do modelo, enquanto a normalização serve para ajustar os dados a uma curva com distribuição normal.

Priscila Schall
gente, a linha 902 aqui deu erro fatal e encerrou a sessão do R hahah. preciso mudar algo por aqui?
Só confere se seu R está na versão 4.1.0 () e dá uma atualizada nos pacotes

Murilo Marin Pechoto
No fim o r^2 ajustado do modelo nao linear ficou pior do que o linear
Pode acontecer, dependendo do Dataset.

Rodrigo Dias Botelho Pellicano
O R^2 ajustado do segundo modelo é menor q do primeiro modelo, não deveria ser maior?
Nem sempre. Depende do dataset.

Jailson de Oliveira Arieira
Prof. Favero, boa noite, você disse que podemos usar qualquer ferramenta (o R é uma) para fazer as análises dos nossos dados. Qual sua opinião sobre o uso do SPSS como alternativa ao R?
Oi Jailson, o professor Fávero já publicou alguns livros onde aborda análise de dados e estatística utilizando SPSS, Stata e outros. Clique aqui e veja os livros.

Mariana Rillo Otero
Pode retomar o passo a passo de forma rápida do processo todo?
Mariana, não entendi muito bem. Mas na gravação, no momento 03:04:30 o professor Fávero faz um compilado do procedimento de forma macro.

Yuri José de Santana Furtado
Não aparece mensagem de erro. O gráfico fica carregando e não aparece. Acaba meio que travando.
Yuri, tente reinstalar todos os pacotes. Caso o erro persista, acho de bom tom reinstalar o R para uma versão mais nova.

Ricardo José Pfitscher
Professor, em termos de interpretação, está correto estimar que se ativo for 0 e liquidez for 0 o retorno de investimento será negativo? Mas faz sentido um retorno de investimento negativo?
A interpretação do modelo é algo muito pessoal do analista a partir das suas experiências práticas.

Leandro Goulart
Boa noite! O Pareto não seria uma forma de ver quais betas são estatisticamente insignificantes?
Não conheço a utilização do modelo de Pareto para estimar regressões.

Alex De Lima Bassi
Professor pq no teste qchisq você usou df = 1
Um grau é para avaliar a significância de cada parâmetro individualmente, visto que o processo de stepwise insere(forward) ou remove(backward) cada parâmetro individualmente para identificar os melhores preditores.
Lembramos que para a abscissa de 3.84, para 1 grau de liberdade, na distribuição qui quadrado resulta em um p-value de 0.05 (5%).

Bruno Daniel Lanza dos Santos
por que o professor usou somente 1 grau de liberdade no proxedimento step, para encontrar o valor crítico de k
Um grau é para avaliar a significância de cada parâmetro individualmente, visto que o processo de stepwise insere(forward) ou remove(backward) cada parâmetro individualmente para identificar os melhores preditores.
Lembramos que para a abscissa de 3.84, para 1 grau de liberdade, na distribuição qui quadrado resulta em um p-value de 0.05 (5%).

Mariana Guimarães Castelo Borges Maié
Não entendi porque usar o quiquadrado para definir o K. Poderia explicar?
Olá Mariana. É uma forma que o professor utiliza para estimar um modelo

Matheus Garcia
Professor, como é sabido o número de graus de liberdade na função qchisq()?
Você determina os graus de liberdade a partir dos parâmetros do modelo.

André Araújo
Por que com apenas 1 grau de liberdade?
O grau de liberdade é calculado a partir dos parâmetros do modelo.

Rafael Viegas De Carvalho Carlos Gomes
Porque apenas 1 grau de liberdade? A base de amostras tem bem mais empresas
O grau de liberdade é calculado a partir dos parâmetros do modelo.

Israel Luiz Harmendani Diniz
Esse algoritmo step funciona como uma espécie de Solver?
Oi Israel, o stepwise é um procedimento para seleção de variáveis. Ele pode ser forward, que é quando começa zerado e as variáveis são inseridas uma a uma ou pode ser backward, que é quando começa com todas as variáveis e elas são removidas uma até que se identifiquem os melhores preditores.

Felipe Francisco Nusda
no chrome ao baixar os arquivos, altera a extensão para txt
Felipe, infelizmente não conseguimos replicar este erro e nunca fomos acionados neste sentido. Tente (1) abrir o site através de uma janela anônima, (2) outra hipótese é que alguma extensão esteja manipulando o comportamento do seu browser e (3) tente acessar por outro navegador de internet (Edge/IE/Firefox...)

Reinaldo Roberto Aranha
Professor, porque o disclosure passou sendo que o p-value dele é menor que o p-value do endividamento e o endividamento não passou?
Compare com valor de 0,05. Quanto menor, melhor.

Teresa Arlinda De Souza Campos
no primeiro gráfico de correlação, "pearson" só aparecem as esferas e mais nada :(
Douglas, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages… Atualização pelo Console não tem funcionado para alguns alunos.

Douglas Pasquali Pedroso
no primeiro gráfico de correlação, "pearson" só aparecem as esferas e mais nada :(
Douglas, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages… Atualização pelo Console não tem funcionado para alguns alunos.

Luiz Novi
meu gráfico não aparece as linhas verdes que ligam os pontos, por que?
Luiz, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages… Atualização pelo Console não tem funcionado para alguns alunos.

Alexandre Barros dos Santos
Para mim o gráfico não apareceu corretamente
Alexandre, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages… Atualização pelo Console não tem funcionado para alguns alunos.

Mariana Rillo Otero
No meu gráfico de correlation, só aparecem as bolinhas e não os traçados. Podem me ajudar?
Mariana, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages... Atualização pelo Console não tem funcionado para alguns alunos.

Israel Luiz Harmendani Diniz
No meu gráfico da 681 ao executar a linha aparecem só as bolinhas sem as linhas. È erro?
Israel, infelizmente pode ser um erro do pacote Rcpp. Tente (1) conferir se o seu R está na versão 4.1, caso não esteja por favor atualize e (2) conferir se o pacote Rcpp está na versão 1.0.7, caso não esteja, atualize ele pela aba Packages... Atualização pelo Console não tem funcionado para alguns alunos.

Reinaldo Roberto Aranha
Boa noite. Sugestão o professor poderia trazer algum exemplo voltado para a empresas do ramo de bioenergia/Agricola.
Reinaldo, vamos repassar para o professor. Obrigado

Ivan Cesar Desuo
Falando em padronização, eu também passo padronizar variaveis X, no caso de estarem em escalas diferentes, em modelos glm ou devo usar rodar os modelo com as brutas mesmo?
Nada impede de realizar a padronização das variáveis.

Raphael Fidelis Valadares
Boa noite! Tudo bem? Gostaria de fazer a seguinte pergunta: o resultado da função export_summs mostra um valor entre parênteses logo abaixo dos valores dos parâmetros. O que eles significam?
São os erros associados a cada parâmetro do modelo.

Daniel Valentins de Lima
Boa noite professor, uma dúvida. Pensei que por estarmos comparando dois modelos (linear e com transformação box cox), teríamos que comparar usando o r quadrado. Por que isso não ocorreu? Obrigado!
Olá Daniel. Você pode comparar o R2 dos dois modelos e verificar se houve mudança.

Mariela Rocha Braga
eu não entendi bem, eu posso testar a normalidade usando SF e depois uso o Boxcox para chegar ao melhor aderência?
Box-Cox serve para diminuir o problema da multicolinearidade, enquanto o teste SF mede o grau de normalidade dos resíduos. Ambos têm objetivos diferentes.

Danilo Steckelberg

É recomendável usar a transformação Box-Cox para usar em problemas de clustering quando temos variáveis que seguem uma distribuição de cauda-longa, como a power-law, ou exponencial, por exemplo?

O objetivo primordial do modelo de box-cox é a redução do problema de multicolinearidade. A clusterização não utiliza a correlação para verificar as distâncias entre as observações. Portanto, a princípio não considero que a utilização do referido modelo seria útil.

Rafael Viegas De Carvalho Carlos Gomes

Ele fez um histograma de residuos? Porque o eixo das ordenadas nao tem valores inteiros, mas sem decimais? Nao entendi.

Olá Rafael, poderia ser mais específico com a indicação da linha do código do modelo? De toda forma, caso a dúvida persista pode enviar um email para monitoria.

Priscila Schall

Boa noite! :) O professor pode revisar o cálculo do R^2 que vimos nas últimas aulas? Estudando para resolver as questões da prova tive dificuldade com isso.

Olá Priscila. No livro do Professor Fávero na biblioteca do Pecege você pode verificar na pág. 522. Tente ver a aula novamente e ler esse capítulo do livro. Mas, caso a dúvida persista, pode mandar um email para monitoria.

Aline Gobbi

O kolmogorov smirnov é no mesmo pacote (nortest) que o shapiro?

Olá Aline. Nesse link verificamos que o teste K-S decorre tem no pacote “dgof”, assim como no pacote “norest”: https://www.geeksforgeeks.org/kolmogorov-smirnov-test-in-r-programming/

Carlos Eduardo Almeida Gomes

boa noite professor, fiquei confuso, a respeito do R^2 , foi dita no começo da aula que a variância respondida entre y em função do X seria 1- R^2 e agora antes do intervalo, era R^2 qual está correta?

Segundo Fávero et ai. (2009), a capacidade explicativa do modelo é analisada pelo R 2 da regressão, conheci- do também por coeficiente de ajuste ou de explicação. Para um modelo de regressão simples, esta medida mostra quanto do comportamento da variável Y é explicado pelo comportamento de variação da variável X, sempre lembrando que não existe, necessariamente, uma relação de causa e efeito entre as variáveis X e Y. Para um modelo de regressão múltipla, esta medida mostra quanto do comportamento da variável Y é explicado pela variação conjunta das variáveis X consideradas no modelo.

Matheus Garcia

Prezados, nesse código da linha 506, o que diferencia as linhas (linear não) é o parâmetro method='lm', correto?

Matheus, é a função “geom_smooth”. Tente isolar cada função para ver como o gráfico se comporta.

Gabriel Campos Vieira
Resposta não me pareceu coerente com a pergunta. Professor, na última aula não ficou muito clara a construção do gráfico com as variáveis dummy. Poderia explicar novamente, por gentileza?
Olá Gabriel, na gravação da aula a explicação do professor começou em 03:19:00.

Caroline Mendes da Silva
Nos vamos aprender modelos com regularizacao? exemplo: elastic net, lasso e etc...
Olá Caroline, veja aqui a programação do curso.

Thiago da Silva Lima
é possível estimar parâmetros considerando apenas strings ?
Olá Thiago. Strings são uma sequência de caracteres, portanto não acho que possam ser consideradas variáveis métricas.

Flávia Ruiz Leão
Fiquei com dificuldade de leitura no gráfico do Plot Sums....Conseguem passar uma explicação dos eixos X e Y dessa plotagem ?
Oi Flávia, tudo bom? Este gráfico é bem legal para comparar parâmetros. No eixo Y temos os parâmetros e no X temos o valor médio e o quanto o valor varia. Se quiser rever este trecho, acesse a aula gravada e avance para 02:55:00.

Ronei Gomes de Almeida
Professor, por gentileza, poderia explicar melhor por que no teste Shapiro-Francia, o p-value é aderente quando maior que 0,05?
É o padrão do teste. A hipótese nula é a existência de normalidade, portanto você não quer rejeita-lá nesse caso.

Pedro Henrique Esteves Trindade
Poderia comentar as abordagens forward e backward do stepwise? Thanks
Oi Pedro, forward é quando a equação começa sem nada e cada preditor entra, um por um, na equação. backward é contrário, todos preditores são incluídos no inicio, e depois são retirados, um a um, até identificar os melhores.

João Batista Duarte
Reitero a pergunta: Por que "k = 3.84" (um grau de liberdade para a distrib. qui.quadrado)?
Oi João, este valor de k foi definido baseado no qui quadrado. Se precisar voltar na explicação, acesse a aula gravada e avance para 02:43:00.

Vanessa Hoffmann de Quadros
Professor, nós vamos ver modelos de séries temporais?
Teremos dados em painel do tipo multinível com medidas repetidas.

EMANUEL RODRIGUES DE VARGAS

Professor, nós teremos modelos com dados em painel (panel data) nesse curso? Se sim, teremos também dados em painel espaciais?

Emanuel, tudo bom? Primeiramente desculpe a resposta errada durante a aula ao vivo, são muitas perguntas e acabei confundindo panel data com visualizações em dashboards. Abordaremos dados em paineis nos modelos multinível,mas paineis espaciais não.

João Batista Duarte

Prof., Por que "um" grau de liberdade?

Um grau é para avaliar a significância de cada parâmetro individualmente, visto que o processo de stepwise insere(forward) ou remove(backward) cada parâmetro individualmente para identificar os melhores preditores.

Lembramos que para a abscissa de 3.84, para 1 grau de liberdade, na distribuição qui quadrado resulta em um p-value de 0.05 (5%).

Gustavo Jorge Braghetto

Ao executar a linha de comando 682 para gerar o gráfico de correlações apareceram apenas os círculos com os nomes das variáveis, sem as linhas ligando a cada uma delas

Olá Gustavo. Tente atualizar os pacotes utilizados na sala e rodar o código novamente. Caso não seja suficiente, tenter reinstalar o R.

EMANUEL RODRIGUES DE VARGAS

Professor, no caso do uso de variáveis financeira é possível que haja endogeneidade. O que você sugere para tratar esse problema?

Olá Emanuel, a multicolinearidade representa um dos problemas mais difíceis de serem tratados em modelagem de dados (Fávero, 2017). Algumas pessoas aplicam o procedimento Stepwise, o que de fato vai eliminar as variáveis explicativas correlacionadas e corrigir essa multicolinearidade, mas também pode omitir uma variável relevante. Sugiro leitura do livro do professor Fávero e deste paper.

Lucas Alves Dias Cardoso

Prof, o k é o número de parâmetros contando ou não contando com alfa? Minha referência são o número de betas (vide slides 9 e 21), mas isto difere do k usado no cálculo de R^2 e do teste F.

Conforme Fávero (2017) o k representa o número de parâmetros do modelo estimado (inclusive o intercepto).

André Zacharias

Prof. Neste slide 27, esses R2 sao ajustados?

André, estes R2 não são ajustados.

Paulo Araujo

Poderia por gentileza nos passar o/um código confiável para realização do teste Anderson-Darling no R??? Os que achei na NET apresentaram resultados discrepantes...

Olá, Paulo. Consultando a documentação do R verifiquei que o único código que roda esse modelo é o “ad.test ()”. Não aparece outros algoritmos.

Simone Mazer

Quais premissas devo considerar para rodar um modelo linear ou não linear?

1) Os resíduos devem apresentar distribuição normal. 2) Não existem correlações elevadas entre as variáveis explicativas e existem mais observações do que variáveis explicativas. 3) Os resíduos não apresentam correlação com qualquer variável X. 4) Os resíduos são aleatórios e independentes, conforme Fávero (2017).

Carla Porto Veiga

professor, modelos lineares, o lm(), são sempre modelos OLS? Ou existem outros "métodos"?

Oi Carla, tudo bem? Existem outros modelos de regressão além do OLS.

EMANUEL RODRIGUES DE VARGAS

Professor, como interpretar a significância do intercepto? Devo considerar que há informações relevantes ou uma forma diferentes que deveria considerar quando o beta0 é significante?

Olá Emanuel, a referida pergunta foi abordada em aula.