PERGUNTAS E RESPOSTAS – MBA EM DATA SCIENCE E ANALYTICS

Disciplina: Supervised Machine Learning: Análise de Regressão Simples e Múltipla IV
Data: 17/08/2021

Bruno Speria
Quando eu crio dummies e elas não são significativas no modelo, excluimos? Se eu excluo não estou deixando de olhar para uma informação? ou esta categoria tem a mesma predição da cat referencia?
Bruno, você pode retirar as variáveis não significantes para a explicação do fenômeno em estudo. Utilizando o procedimento stepwise (abordado pelo professor Fávero na aula do dia 10/agosto) você automatiza o trabalho e o resultado já será a melhor combinação.

Bruno Speria
Se utilzo variaveis Dummies mas ao fazer isso elas estoram o numero de observações?
Bruno, a resposta da pergunta acima acaba complementando esta, o stepwise vai reduzir a quantidade de variáveis. Mais variáveis do que observações pode implicar em o modelo não aprender bem com a análise.

Bruno Speria
Porque não posso ter mais variaveis que observações. Em uma das aulas foi comentado mas não achei uma explicação matematica.
Bruno, porque para a estimação dos parâmetros beta é dada pelo seguinte cálculo matricial:
beta = (X'X)^-1 X'Y
Esse é um pressuposto matemático básico. Como o algoritmo computacional utilizado para o cálculo inverte a matriz para encontrar os coeficientes adequados, caso o número de parâmetros a serem estimados supere a quantidade de observações, a estimação torna-se impossível matematicamente.

Wiliam Rocha
Professor, você poderia repetir novamente a parte de como interpretar os valores do ols_test_breusch_pagan?
William, a hipótese nula do teste de Breusch-Pagan é ausência de heterocedasticidade. Caso a estatística de teste seja maior que o p-valor associado, não se rejeita a hipótese nula, ou seja, atende-se ao pressuposto.

Murilo Marin Pechoto
Professor, poderia fazer exemplo de predict quando utilizamos variaveis dummies. obrigado
O professor Fávero irá compartilhar no script auxiliar, Murilo.

Lucas de Oliveira Lima
Professor, se eu dummizar mais de uma variável, o alfa vai absorver o comportamento de todas? Caso sim, como discernir?
Exatamente, Lucas. Neste caso as duas (ou quantas forem as variáveis dummy) irão para o parâmetro alfa. Neste caso a análise é feita dentro de cada variável, pois estamos testando a categoria de referência de cada variável contra as outras categorias daquela mesma variável.

Gustavo de Holanda Campos
Utilizamos BOXCOX na variável Y para maximizar a aderência à normalidade. Como saber o momento de aplicar BOXCOX em variáveis X? Obrigado.
Gustavo, para cada caso é necessário que se siga o passo a passo indicado pelas boas práticas. Cada base de dados deve ser analisada para verificar a questão da linearidade, caso não haja linearidade pode ser considerada a transformação box-cox ou mesmo a adoção de uma outra especificação.

Gabriela Alves De Almeida
prof e se uma das dummys de uma var quali não for significativa, o que podemos fazer?
Gabriela, você pode retirar as variáveis não significantes para a explicação do fenômeno em estudo. Utilize o procedimento stepwise (abordado pelo professor Fávero na aula do dia 10/agosto) para automatizar esse trabalho.

Danilo Steckelberg
A heterocedasticidade pode ser reduzida por alguma transformação direta na variável x? Por exemplo, log ou box-cox?
Danilo, existe sim essa alternativa. Para maiores explicações, verificar a função BOX-COX do pacote MASS.

Mariana Guimarães Castelo Borges Maié
Você recomendaria usar um modelo de regressão para prever o custo de planos de saúde tendo uma das variáveis categorica indicando a presença ou não da COVID? Se não, algum outro método?
Mariana, é possível sim. Esta variável dicotômica pode ser uma variável explicativa de um modelo de Regressão Linear Múltipla; pode estar ainda como a variável dependente de um modelo de Regressão Logística (que aprenderemos nas próximas aulas), dentre inúmeras outras possibilidades entre os modelos que ainda aprenderemos ao longo do curso.

Ricardo Rocha Pavan da Silva
qual a diferença entre (residuals)^2 e (residuals^2)
Ricardo, é uma diferenciação matemática. O somatório dos resíduos ao quadrado é numericamente diferente do somatório dos quadrados dos resíduos.

Gabriel Campos Vieira
As perguntas sumiram para mim. Pergunto novamente se a base de dados saeb pode ser considerada censitária.
Gabriel, bases censitárias são bases que contém toda e qualquer observação referente a determinado fenômeno. Se considerarmos que a SAEB tem os dados de todas as escolas públicas do país, ela é uma base censitária em relação às escolas públicas.

Vitor Bruno da Silveira Guimarães
No caso da heterocedasticidade, o ideal é identificar essas variáveis que foram omitidas, usando stepwise ou existem outras formas. E quanto é assim, porque a maioria teve um comportamento padrão?
Vitor, com o procedimento stepwise estaremos retirando as variáveis não significantes para a explicação do fenômeno em estudo. No caso do teste de heterocedasticidade estaremos

verificando se há variáveis relevantes para a análise e que não constam do modelo. Possíveis caminhos para suprir essa omissão são: verificar a literatura de referência, encontrar outros modelos semelhantes, buscar na prática anterior ou sua expertise, dentre outras técnicas.

Gabriel Campos Vieira
Professor, essa base de dados de escolas pode ser considerada censitária? Trabalho com uma base parecida com essa, com dados sobre abastecimento de água nos municípios do Brasil e tenho essa dúvida tb
Gabriel, bases censitárias são bases que contém toda e qualquer observação referente a determinado fenômeno. Se considerarmos que a SAEB tem os dados de todas as escolas públicas do país, ela é uma base censitária em relação às escolas públicas.

Rodrigo Alves Pereira Gitirana
Como testar de uam maneira não manual, a multicolinearidade de um dataset com muitas variaveis ? Sem ter que testar uma contra a outra ? Existe alguma forma de automatizar ?
Rodrigo, você pode considerar o uso do procedimento Stepwise abordado pelo professor Fávero na aula do dia 10/agosto.

Gabriel Rodrigues Coutinho Pereira
No caso em que uma dummy for pro alfa, os valores de p-value do teste t para o alfa, precisa passar também para validar o modelo? Ou o alfa pode nesse caso também ser não significante?
Gabriel, nesse caso o alfa é somente a categoria de referência, uma espécie de parâmetro a ser testado contra as outras categorias em cada variável dummy.

André Araújo
Voltando a pergunta do Thiago Colette Vegi, e as análises de demanda onde é necessário prever as vendas de uma empresa que está numa crescente?
André, para cada caso é necessário que se siga o passo a passo indicado pelas boas práticas. Cada base de dados deve ser analisada para verificar a questão da linearidade, caso não haja linearidade pode ser considerada a transformação box-cox ou mesmo a adoção de uma outra especificação.

Juliano Sartorelli Dias
Professor, quando se tem mais de uma variável categórica, ao torna-las dummy duas categorias vão ir para o parâmetro alfa?
Exatamente, Juliano. Neste caso as duas (ou quantas forem as variáveis dummy) irão para o parâmetro alfa.

LUCAS FERRER FRANCO
Professor vi que a UF_RO nao tinha passado no teste, tudo bem mesmo assim?
Lucas, significa que essa categoria não é significante para explicar o fenômeno em comento. A presença ou ausência desta característica não influencia no resultado final.

Vanessa Hoffmann de Quadros
Professor, vamos ver também o problema de autocorrelação dos resíduos?
Sim, Vanessa, visto durante a aula. Caso não tenha ficado claro, nos encaminhe um e-mail que ficaremos felizes em auxiliar.

Juliano Sartorelli Dias
Professor e se excluir os NA, o $ funciona?
Juliano, sim. O comando para a exclusão dos NA's somente retira linhas da base, as colunas permanecem inalteradas.

Leandro Modesto Prates Beltrão
O modelo com heterocedasticidade deve ser considerado inválido para uso?
Leandro, os modelos de MQO com distribuição heterocedástica do erro, isto é, que não atendem ao pressuposto em comento, perdem a propriedade de encontrar a melhor estimativa dos parâmetros populacionais.

Leandro Modesto Prates Beltrão
Professor, quais seriam os problemas associados à utilização de um modelo com heterocedasticidade?
Leandro, esse é um pressuposto matemático básico. Como o algoritmo computacional utilizado para o cálculo inverte a matriz para encontrar os coeficientes adequados, caso o número de parâmetros a serem estimados supere a quantidade de observações, a estimação torna-se impossível matematicamente.

Igor Cabral Corrêa
A heterocedasticidade pode não ter um padrão cônico e ser somente um reflexo de ruído?
Igor, a heterocedasticidade se apresenta de diversas formas. A dispersão dos resíduos pode se apresentar de forma linear, constante, quadrática, etc.