Análise de regressão múltipla e multivariada de variância e discriminante

Você vai explorar técnicas estatísticas para analisar, prever e comparar variáveis, por meio da regressão múltipla, regressão logística, análise discriminante e análise de variância multivariada (MANOVA).

Profa. Manoela Gonçalves Cabo

Objetivos

- Investigar relações entre variáveis quantitativas por meio de modelos de regressão múltipla.
- Identificar padrões de classificação utilizando análise discriminante e regressão logística com variáveis dependentes categóricas.
- Comparar vetores de médias de grupos utilizando a análise de variância multivariada (MANOVA) em contextos experimentais ou observacionais.

Introdução

Neste conteúdo, serão apresentadas três técnicas estatísticas multivariadas amplamente utilizadas na modelagem de dados quantitativos e categóricos: a análise de regressão múltipla, a análise discriminante e logística, e a análise de variância multivariada (MANOVA). Essas técnicas fazem parte da análise multivariada com foco em dependência entre variáveis, e permitem desenvolver modelos preditivos e de classificação a partir de dados complexos.

Na primeira etapa do conteúdo, será abordada a regressão múltipla, uma técnica de dependência que permite prever uma variável dependente métrica com base em várias variáveis independentes. Serão discutidos conceitos fundamentais como o modelo de regressão linear, a interpretação dos coeficientes, as suposições estatísticas, o processo de decisão e a construção de modelos utilizando ferramentas como Excel e R.

Em seguida, o foco será a análise discriminante e a regressão logística, técnicas indicadas quando a variável dependente é categórica. A análise discriminante é útil para classificar elementos em grupos predefinidos com base em variáveis métricas, enquanto a regressão logística é ideal para variáveis dependentes dicotômicas. O conteúdo inclui também o processo de decisão, critérios de avaliação de modelos e exemplos práticos.

Por fim, a MANOVA será apresentada como uma extensão da ANOVA, capaz de lidar com múltiplas variáveis dependentes simultaneamente. Essa técnica permite comparar grupos com base em vetores de médias, controlando taxas de erro e aumentando o poder estatístico dos testes. O conteúdo também abordará as suposições do modelo, os testes estatísticos aplicáveis e a interpretação dos resultados em contextos reais.

Análise de regressão múltipla

Para começar, assista ao vídeo e conheça análise de regressão múltipla.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A análise de regressão é uma técnica de dependência usada de forma versátil, aplicável na tomada de decisões em economia e negócios. Não é a técnica mais usada em análise multivariada; no entanto, seus usos variam desde os problemas mais gerais atéf os mais específicos, sendo que em cada caso relaciona um fator (ou fatores) a um resultado específico.

Ela é o fundamento para realizar previsões dos problemas estudados, podendo ser utilizada para modelos econômicos e modelos de desempenho de negócios em diversos ramos. A análise de regressão múltipla pode ser compreendida como uma coleção de técnicas estatísticas para desenvolver e analisar modelos que descrevem de maneira plausível as relações entre a variável dependente e as variáveis independentes de um determinado processo. A forma básica de apresentação é a associação da variável dependente (Y) às variáveis independentes X_1 , conforme a seguir:

$$Y_i = X_1 + X_2 + X_3 + \ldots + X_n$$

A especificação funcional com o termo aleatório:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e_i$$

Em que,

β_0 é uma constante ou um termo intercepto;

X_i as variáveis independentes;

β_i os coeficientes da regressão ou pesos;

e_i os erros do modelo.

Saiba mais

O objetivo principal da análise de regressão múltipla é usar as variáveis independentes (Xi), que são os critérios, cujos valores são conhecidos para prever os valores da variável dependente (Y), ou preditora, selecionada pelo pesquisador. Cada variável independente, ou critério, será ponderada pelos pesos (βi) na análise de regressão, com objetivo de garantir máxima previsão.

Esses pesos significam a contribuição das variáveis independentes para o modelo geral e facilitam a interpretação sobre a influência de cada variável em fazer a previsão, apesar de a correlação entre as variáveis independentes complicar a aplicação da técnica, diferentemente da análise fatorial.

> O conjunto de variáveis independentes ponderadas forma a variável estatística de regressão ou equação de regressão ou modelo de regressão, sendo uma das técnicas mais utilizadas na estatística e econometria.

Os coeficientes de regressão (β_i) representam a variação estimada na variável dependente por variação unitária da variável independente. Se o coeficiente de regressão é percebido como significativo, o valor do coeficiente de regressão indica a extensão na qual a variável independente se associa com a dependente.

Um outro ponto importante é que, ao fazer a previsão da variável dependente, podemos melhorar nossa precisão usando uma constante no modo de regressão (β_0). Conhecida como intercepto, ela representa o valor da variável dependente quando todas as variáveis independentes têm um valor nulo. Existem dois tipos de dados com que iremos trabalhar:
Variáveis Quantitativas e Variáveis Qualitativas.

Variáveis Quantitativas	Variáveis Qualitativas
Implicam em relações de mensuração, medida, contagem (por exemplo, preços, quantidades, taxas, retornos).	Expressam atributos, qualidades do indivíduo pesquisado (por exemplo, classificação de crédito, setor, profissão, escolaridade, satisfação).

Atenção

É possível incluir dados qualitativos ou não métricos como variáveis independentes (transformando dados ordinais ou nominais com codificação dicotômica), ou como a variável dependente (pelo uso de uma medida binária na técnica especializada de regressão logística). Para identificarmos o melhor modelo a ser aplicado, temos que obter a menor soma possível de quadrados dos erros (chamados de erros quadrados) como nossa medida de precisão de previsão. O objetivo é minimizar a soma dos quadrados dos erros ou mínimos quadrados ordinários (MQO).

Um outro ponto importante a se observar na análise de regressão é o **estabelecimento de um intervalo de confiança para os coeficientes de regressão e o valor previsto**. Como usamos apenas uma amostra de observações para estimar uma equação de regressão, podemos esperar que os coeficientes de regressão variem se selecionarmos outra amostra de observações e estimarmos outra equação de regressão.

Para esse tipo de verificação, existe o teste empírico para ver se o coeficiente de regressão é diferente de zero ou se iguala a zero em outra amostra. No teste estatístico do intercepto e dos coeficientes de regressão para determinar se eles são significativamente diferentes de zero, considera-se o erro padrão da estimativa (SE_E) e acrescenta-se mais ou menos um desvio padrão.

Exemplo

O acréscimo de ±1,96 desvio padrão à média define um intervalo para grandes amostras que inclui 95% dos valores de uma variável: Além de verificar o intervalo de confiança para os coeficientes de regressão, temos que analisar a avaliação da precisão de previsão. A soma de quadrados dos erros (SSE) representa uma medida do erro de previsão.

Também devemos conseguir determinar uma medida de nosso sucesso de previsão; chamamos de soma de quadrados da regressão (SSR) a soma dessas duas medidas, que nos dará a soma total de quadrados (SST):

$$SST = SSE + SSR$$

Para calcularmos a soma de quadrados dos erros (SSE), a soma de quadrados da regressão (SSR) e a soma total de quadrados (SST):

$$SST = \sum_{i=1}^{n} (y_{i-} - \bar{y})^2$$

$$SSE = \sum_{i=1}^{n} (y_{i-} - \hat{y})^2$$

$$SSR = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2$$

$$\sum_{i=1}^{n} (y_{i-} - \bar{y})^2 = \sum_{i=1}^{n} (y_{i-} - \hat{y})^2 + \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 \quad SST = SSE + SSR$$

Em que,

$y =$ valor da observação i

$\bar{y} =$ média de todas as observações

$\hat{y} =$ valor previsto da observação i

No vídeo a seguir, conheça mais detalhes sobre a análise de regressão múltipla.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

> Coeficiente de determinação $(R^2) = \frac{\text{soma de quadrados da regressão}}{\text{soma dos quadrados totais}}$. Quanto mais próxima de 1, melhor é a precisão da previsão.

Processo de decisão para a análise de regressão múltipla

O processo de decisão para análise de regressão múltipla constitui-se na construção de modelo em três estágios:

I - Seleção das variáveis preditoras

Em que verifica-se a hipótese a ser testada, a especificação dos objetivos da análise de regressão e selecionam-se as variáveis que farão parte do modelo, incluindo as variáveis dependente e independentes, considerando os fatores como o tamanho da amostra e a necessidade de transformações de variáveis.

II - Escolha do modelo de regressão e estimativa dos coeficientes

Com o modelo de regressão formulado, as suposições inerentes à análise de regressão são primeiramente testadas para as variáveis individuais; se todas as suposições forem atendidas, então o modelo será estimado.

III - Fase da abrangência do modelo

Quando já temos os resultados. São feitas as análises diagnósticas para garantir que o modelo geral atenda às suposições de regressão e que nenhuma observação tenha influência indevida sobre os resultados. O próximo estágio é a interpretação da variável estatística de regressão, quando se examina o papel desempenhado por cada variável independente na previsão da medida dependente.

> Onde, no primeiro estágio, o problema é escolher um conjunto de variáveis que podem ou devem ser incluídas no modelo. No segundo, é possível usar um modelo teórico e aproximações por modelos polinomiais. E, no terceiro, geralmente, é necessário restringir a abrangência do modelo para alguns valores ou região da(s) variável(is) preditora(s).

Na etapa de abrangência do modelo devemos fazer várias suposições sobre as relações entre as variáveis dependentes e independentes que afetam o procedimento estatístico (mínimos quadrados) usado para regressão múltipla. Devemos realizar os testes para as suposições e ações corretivas no caso de ocorrerem violações. A questão básica é se, no curso do cálculo dos coeficientes de regressão e de previsão da variável dependente, as suposições da análise de regressão são atendidas. Os erros na previsão são resultado de uma ausência real de uma relação entre as variáveis, ou são causados por algumas características dos dados não acomodadas pelo modelo de regressão.

1. Linearidade do fenômeno medido.

2. Variância constante dos termos de erro.

3. Independência dos termos de erro.

4. Normalidade da distribuição dos termos de erro.

> Para a avaliação das suposições estatísticas, os testes de suposições devem ser feitos não apenas para a variável dependente e cada variável independente, mas também para a variável estatística.

> As análises gráficas, como gráficos de regressão parcial, de resíduos e de probabilidade normal, são os métodos mais amplamente usados de avaliação de suposições para a variável estatística. As ações corretivas para problemas encontrados na variável estatística devem ser realizadas pela modificação de uma ou mais variáveis independentes.

Aplicação da análise de regressão

No vídeo a seguir, veja na prática uma aplicação de regressão múltipla.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Podemos resolver a Análise de Regressão utilizando o Excel ou o software R. Abordaremos as duas formas a seguir; no entanto, o processo de decisão para a Análise de Regressão Múltipla constitui-se na construção de modelo em três estágios.

Estágio I - Seleção das variáveis preditoras

As variáveis podem ser selecionadas verificando a hipótese que se deseja testar. Como exemplo, segue a análise do PIB:

Exemplo

Uma equipe de pesquisadores deseja determinar o crescimento econômico do Brasil no ano de 2019; para essa análise, os economistas utilizaram o PIB (Produto Interno Bruto) como proxy do crescimento econômico. Ao iniciar a pesquisa, selecionaram algumas variáveis para formular a hipótese a ser estudada; a variável dependente selecionada foi o PIB a preço de mercado real, e as variáveis independentes foram o Consumo de Energia Elétrica do Comércio (Energia) e a Taxa de desemprego (Desemprego) nos anos compreendidos entre 1995 e 2018. As informações selecionadas foram dispostas em forma de uma série temporal anual conforme a Tabela 1 a seguir. Ao analisarem as informações, os pesquisadores verificaram que o Consumo de Energia Elétrica e a Taxa de desemprego seriam variáveis que poderiam explicar muito bem o comportamento do PIB. Para atingir o objetivo da pesquisa, que é determinar o crescimento econômico do Brasil, esses economistas utilizaram um modelo de regressão linear múltipla para testar a teoria de que o PIB do ano de 2019 pode ser previsto utilizando-se o Consumo de Energia Elétrica do Comércio e a Taxa de desemprego dos anos entre 1995 e 2018. As informações para o trabalho foram obtidas pelos pesquisadores no site do IPEADATA.

Ano	PIB	Energia	Desemprego
1995	2.434.107	32.276	158
1996	2.487.873	34.338	180
1997	2.572.333	38.198	189
1998	2.581.030	41.544	218
1999	2.593.107	43.588	231
2000	2.706.892	47.626	212

Ano	PIB	Energia	Desemprego
2001	2.744.515	44.434	210
2002	2.828.317	45.222	228
2003	2.860.584	47.522	239
2004	3.025.352	49.609	226
2005	3.122.228	52.985	204
2006	3.245.930	55.475	191
2007	3.442.954	58.744	180
2008	3.618.345	61.949	162
2009	3.613.793	65.379	166
2010	3.885.847	69.170	145
2011	4.040.287	73.481	127
2012	4.117.908	79.231	130
2013	4.241.644	83.696	125
2014	4.263.020	89.816	129
2015	4.111.863	90.543	154
2016	3.977.162	87.896	199
2017	4.029.775	88.130	216
2018	4.082.856	88.797	198
2019	4.239.223	92.119	184

Tabela 1 - PIB a preço de mercado real, o Consumo de Energia Elétrica do Comércio (Energia) e a Taxa de desemprego (Desemprego) dos anos entre 1995 e 2018.

Estágio II - Escolha do modelo de regressão

Em seguida, formula-se o modelo a ser escolhido. Nesse caso, o modelo será o seguinte

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + e_i$$

Em que a variável Y é o PIB, a variável X1 é a Energia e a variável X2 é a taxa de desemprego:

$$\text{PIB} = \beta_0 + \beta_1 * \text{Energia} + \beta_2 * \text{Desemprego} + e_i$$

Depois de analisarem as informações, utilizamos a ferramenta de Análise de Dados do software Excel e fizemos uma regressão linear múltipla (Análise de Dados/Regressão):

ANOVA

	gl	*SQ*	*MQ*	*F*	*F de significação*
Regressão	2	9.610.053.931.678	4.805.026.965.839	331,94	0,00
Resíduo	21	303.984.505.991	14.475.452.666		
Total	23	9.914.038.437.668			

	Coeficientes	*Erro padrão*	*Stat t*	*valor-P*
Interseção	2.220.236	205816,1479	10,78747376	0,00000
Energia	29	1,459055544	20,04842081	0,00000
Desemprego	-3.546	791,7147257	-4,478458971	0,00021

Os parâmetros do modelo estimado são:

$$\beta_0 = 2.220.236$$

$$\beta_1 = 2.29$$

$$\beta_2 = -3.546$$

O modelo final estimado: PIB = 2.220.236 + 29*Energia – 3,546*Desemprego + ei.

> No R, a função para realizar o ajuste do modelo é:
> modelo ← lm(PIB ~ Energia + Desemprego, data=dados) summary (modelo)

Estágio III - Abrangência do modelo

Para encontrarmos o coeficiente de determinação R_2 temos que verificar a SST (soma dos quadrados totais), SSE (soma dos quadrados erros) e SSR (soma dos quadrados regressão):

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 9.914.038.437.668$$

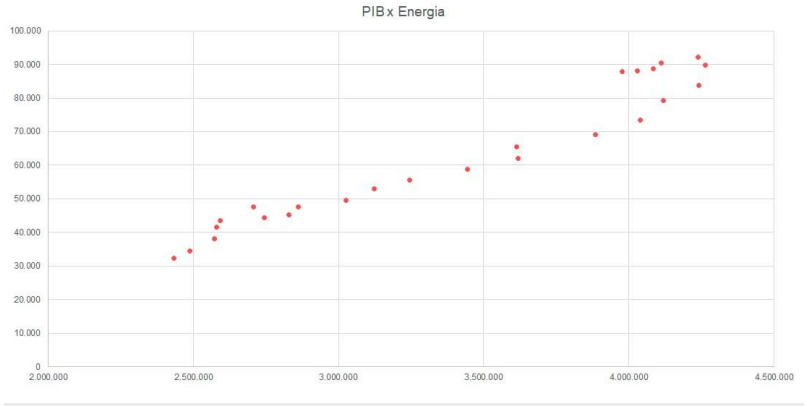
$$SSE = \sum_{i=1}^{n} (y_{i-} - \hat{y})^2 = 303.984.505.910$$

$$SSR = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = 9.610.053.931.678$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - 303.984.505.910/9.914.038.437.668 = 0,969338$$

O coeficiente de determinação é definido pela razão entre a variação total de Y em relação a todas as variáveis X. indicando se a curva testada permite explicar o comportamento da variável Y a partir do comportamento da variável X. Se o valor R_2 for baixo, isto quer dizer apenas que a curva testada não o faz, mas nada impede que outra curva se ajuste melhor. Nesse caso, o teste de Eficácia do Ajustamento chega a explicar 97% do PIB, sendo considerado um bom ajuste.

Para verificar a linearidade do fenômeno medido, podemos utilizar o gráfico de dispersão da variável dependente com a variável independente. No Excel, podemos inserir o gráfico de dispersão pelo comando: Inserir Gráfico/ Gráfico de dispersão.



Verificando o aprendizado

Questão 1

Qual a forma básica de apresentação do modelo de regressão para análise de regressão múltipla?

Chave de resposta

É uma técnica importante para relacionar matematicamente duas ou mais variáveis, com uma função importante para entender os fenômenos físicos, químicos, biológicos, sociais, médicos, entre outros. A forma básica de apresentação é a associação da variável dependente (Y) às variáveis independentes (Xi), conforme a seguir:

$$Y_i = X_1 + X_2 + X_3 + \ldots + X_n$$

A especificação funcional com o termo aleatório:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + e_i$$

Em que,

β_0 é uma constante ou um termo intercepto;

X_i as variáveis independentes;

β_i os coeficientes da regressão ou pesos;

e_i os erros do modelo.

Questão 2

Descreva as etapas do processo de decisão para a análise de regressão múltipla:

Chave de resposta

A primeira é a seleção das variáveis preditoras, em que verifica-se a hipótese a ser testada, a especificação dos objetivos da análise de regressão e selecionam-se as variáveis que farão parte do modelo, incluindo as dependentes e independentes, considerando os fatores como o tamanho da amostra e a necessidade de transformações de variáveis.

A segunda é a escolha do modelo de regressão, e estimam-se os coeficientes do modelo. Com o modelo de regressão formulado, as suposições inerentes à análise de regressão são primeiramente testadas para as variáveis individuais; se todas as suposições forem atendidas, então o modelo será estimado.

A terceira é a fase da abrangência do modelo, quando já temos os resultados e são feitas as análises diagnósticas para garantir que o modelo geral atenda às suposições de regressão e que nenhuma observação tenha influência indevida sobre os resultados.

O próximo estágio é a interpretação da variável estatística de regressão, quando se examina o papel desempenhado por cada variável independente na previsão da medida dependente.

Questão 3

Encontre o coeficiente de determinação R_2 , sendo que a SST (soma dos quadrados totais) = 200, SSE (soma dos quadrados erros) = 80 e SSR (soma dos quadrados regressão) = 160.

Chave de resposta

Gabarito

$$SST = \sum_{i=1}^{n} (y_{i-} - \bar{y})^2 = 200$$

$$SSE = \sum_{i=1}^{n} (y_{i-} - \hat{y})^2 = 80$$

$$SSR = \sum_{i=1}^{n} (\hat{y} - \bar{y})^2 = 160$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{80}{200} = 0,6$$

Nesse caso, o teste de Eficácia do Ajustamento chega a explicar 60%.

Análise discriminante e regressão logística

Para começarmos, conheça a análise discriminante e regressão logística.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

A regressão múltipla é sem dúvida a técnica de dependência multivariada mais amplamente empregada. A base para a popularidade da regressão tem sido sua habilidade de prever e explicar variáveis métricas.

> Mas o que acontece quando variáveis não métricas ou dicotômicas tornam a regressão múltipla inadequada?

Confira as técnicas que podem ser usadas para classificação dos elementos, a análise discriminante e a regressão logística, que tratam da situação de uma variável dependente não métrica.

Análise discriminante

A análise discriminante é uma técnica que pode ser utilizada para classificação de elementos de uma amostra ou população, mas que difere dos métodos de análise de conglomerados ou *clusters*. Para a sua aplicação, é necessário que os grupos para os quais cada elemento amostral pode ser classificado sejam predefinidos, ou seja, conhecidos a *priori* considerando-se suas características gerais.

Regressão logística

A regressão logística consiste em uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Com relação às variáveis independentes, elas podem ser categóricas ou não. A regressão logística nos permite estimar as probabilidades associadas à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

A característica dessa regressão é que ela busca estimar a probabilidade de a variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis. O resultado da análise fica entre zero e um.

O objetivo da análise discriminante e regressão logística é estimar a relação entre uma variável dependente não métrica (dicotômica ou categórica) e um conjunto de variáveis independentes métricas.

Saiba mais

Elas podem ser aplicadas em diversas áreas, como gestão de risco, classificação de crédito, Balanced Scorecard (BSC), o sucesso ou fracasso de um novo produto, decidir se um estudante deve ser aceito em uma faculdade, classificar estudantes quanto a interesses vocacionais, entre outras.

São representadas nesta forma geral:

$$Y = X_1 + X_2 + \ldots + X_n$$

(Não métrica) (Métricas ou não métricas)

Com relação à análise discriminante, sua função discriminante é uma variável estatística das variáveis independentes selecionadas por seu poder discriminatório usado na previsão de pertinência ao grupo. O valor previsto da função discriminante é o escore Z discriminante, o qual é calculado para cada objeto (pessoa, empresa ou produto) na análise.

Ele toma a forma da equação linear:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \ldots + W_n X_{nk}$$

Em que,

$Z_{jk} =$ escore Z discriminante da função discriminante j para o objeto k

$a =$ intercepto

$W_i =$ peso discriminante para a variável independente i

$X_{ik} =$ variável independente i para o objeto k

De acordo com a função, o escore discriminante para cada objeto na análise pode ser uma empresa, escola, pessoa, firma etc., como sendo uma soma dos valores obtidos pela multiplicação de cada variável independente por seu peso discriminante.

Atenção

Essa é uma técnica para testar a hipótese de que as médias de grupo de um conjunto de variáveis independentes para dois ou mais grupos são iguais, fazemos isso calculando a média dos escores discriminante para todos os indivíduos do grupo e essa média do grupo é chamada de centroide.

Analogia entre regressão e MANOVA

A análise discriminante é quase a mesma da análise de regressão, ou seja, a função discriminante é uma combinação linear (variável estatística) de medidas métricas para duas ou mais variáveis independentes, e é usada para descrever ou prever uma única variável dependente.

Atenção

A diferença chave é que a análise discriminante é adequada a problemas de pesquisa nos quais a variável dependente é categórica (nominal ou não métrica), ao passo que a regressão é usada quando a variável dependente é métrica. A regressão logística é uma variante da regressão, tendo assim muitas semelhanças, exceto pelo tipo de variável dependente, que seria o tipo de regressão da análise discriminante.

A análise de variância multivariada (MANOVA - *Multiple Analysis Of Variance*) é uma extensão ou forma generalizada da análise de variância (ANOVA). Ela é utilizada em casos com duas ou mais variáveis dependentes, e é um procedimento para comparação de médias amostrais multivariadas. Analisa simultaneamente múltiplas medidas de cada indivíduo ou objeto sob investigação.

> Como um procedimento multivariado, é usada quando há duas ou mais variáveis dependentes, e é tipicamente seguida por testes de significância envolvendo variáveis dependentes individuais separadamente.

As hipóteses a serem testadas é se a média das amostras são iguais; para isso, são utilizados quatro testes na análise MANOVA:

1. Traço de Pillai;

2. Traço de Hoetelling;

3. Lambda de Wilks;

4. Maior raiz de Roy.

A análise discriminante também é comparável à análise multivariada de variância (MANOVA) “reversa”. Na análise discriminante, a variável dependente é categórica e as independentes são métricas.

O oposto é verdadeiro em MANOVA, que envolve variáveis dependentes métricas e variáveis independentes categóricas. As duas técnicas usam as mesmas medidas estatísticas de ajuste geral do modelo.

Processo de decisão para análise discriminante e regressão logística

A seguir, conheça sobre o processo de decisão usada para análise discriminante e regressão logística.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

O processo de decisão para análise discriminante é realizado através do método de seis estágios, conforme a seguir:

Problema de pesquisa

Selecionamos o objetivo; calculamos as diferenças de grupo em um perfil multivariado; classificamos as observações em grupos; identificamos as dimensões de discriminação entre grupos.

Questões de planejamento de pesquisa

Selecionamos as variáveis independentes; fazemos as considerações sobre tamanho de amostra; realizamos a criação de amostras de análise e teste.

Suposições

Normalidade de variáveis independentes; linearidade de relações; falta de multicolinearidade entre variáveis independentes; matrizes de dispersão iguais.

Estimação do modelo discriminante e avaliação do ajuste geral

Fazer a estimação simultânea ou *stepwise*; verificar a significância de funções discriminantes. Avaliação de precisão preditiva com matrizes de classificação: determinar os escores de corte ótimo; especificar critérios para avaliação de razão de sucesso; verificar a significância estatística de precisão preditiva.

Interpretação das funções discriminantes

Quantas funções serão interpretadas?; Avaliação da função e da função separada (Pesos discriminantes, Cargas discriminantes, Valores F parciais); Avaliação de funções combinadas (Rotação das funções, índice de potência, representação gráfica dos centroides de grupos e de cargas).

Validação dos resultados discriminantes

Amostras particionadas ou validação cruzada; Perfil de diferença de grupos.

A análise discriminante pode abordar qualquer um dos seguintes objetivos de pesquisa:

- Determinar se existem diferenças estatisticamente significantes entre os perfis de escore médio em um conjunto de variáveis para dois (ou mais) grupos definidos *a priori*.
- Determinar quais das variáveis independentes explicam o máximo de diferenças nos perfis de escore médio dos dois ou mais grupos.
- Estabelecer o número e a composição das dimensões de discriminação entre grupos formados a partir do conjunto de variáveis independentes.
- Estabelecer procedimentos para classificar objetos (indivíduos, firmas, produtos e assim por diante) em grupos, com base em seus escores em um conjunto de variáveis independente.

No projeto de pesquisa para análise discriminante, a aplicação bem-sucedida requer a consideração de várias questões. Tais questões incluem a seleção da variável dependente e das variáveis independentes, o tamanho necessário da amostra para a estimação das funções discriminantes, e a divisão da amostra para fins de validação.

Já no planejamento de análise discriminante, a variável dependente deve ser não métrica, representando grupos de objetos que devem diferir nas variáveis independentes.

1. Escolha uma variável dependente que melhor represente diferenças de grupos de interesse.
2. Escolha uma variável dependente que melhor represente diferenças de grupos de interesse.
3. Defina grupos substancialmente distintos.
4. Minimize o número de categorias ao mesmo tempo que atenda aos objetivos da pesquisa, ao converter variáveis métricas para uma escala não métrica para uso como a variável dependente.
5. Considere o uso de grupos extremos para maximizar as diferenças de grupos.

> Variáveis independentes devem identificar diferenças entre pelo menos dois grupos para uso em análise discriminante.

A amostra deve ser grande o bastante para ter pelo menos uma observação a mais por grupo do que o número de variáveis independentes, e procurar por pelo menos 20 casos por grupo.

Deve haver 20 casos por variável independente, com um nível mínimo recomendado de 5 observações por variável; ter uma amostra grande o bastante para dividi-la em amostras de teste e de estimação, cada uma atendendo às exigências acima.

Atenção

A suposição mais importante é a igualdade das matrizes de covariância, o que afeta tanto a estimação quanto a classificação. A multicolinearidade entre as variáveis independentes pode reduzir sensivelmente o impacto estimado delas na função discriminante derivada, particularmente no caso de emprego de um processo de estimação stepwise.

Na estimação do modelo discriminante e avaliação do ajuste geral, para determinarmos a função discriminante, devemos decidir o método de estimação e então determinar o número de funções a serem retidas. Com as funções estimadas, o ajuste geral do modelo pode ser avaliado de diversas maneiras.

Primeiro, escores Z discriminantes, também conhecidos como os escores Z, podem ser calculados para cada objeto. A comparação das médias dos grupos (centroides) nos escores Z fornece uma medida de discriminação entre grupos.

A precisão preditiva pode ser medida como o número de observações classificadas nos grupos corretos, com vários critérios disponíveis para avaliar se o processo de classificação alcança significância prática ou estatística. Finalmente, diagnósticos por casos podem identificar a precisão de classificação de cada caso e seu impacto relativo sobre a estimação geral do modelo.

A estimação *stepwise* é uma alternativa à abordagem simultânea. Envolve a inclusão das variáveis independentes na função discriminante, uma por vez, com base em seu poder discriminatório. O método *stepwise* é útil quando o pesquisador quer considerar um número relativamente grande de variáveis independentes para inclusão na função.

Compare agora a análise discriminante e regressão logística:

Análise discriminante

A análise discriminante é apropriada quando a variável dependente é não métrica. No entanto, quando a variável dependente tem apenas dois grupos, a regressão logística pode ser preferida por duas razões.

A análise discriminante depende estritamente de se atenderem as suposições de normalidade multivariada e de igualdade entre as matrizes de variância covariância nos grupos, suposições que não são atendidas em muitas situações.

Regressão logística

A regressão logística não depende dessas suposições rígidas e é muito mais robusta quando tais pressupostos não são satisfeitos, o que torna sua aplicação apropriada em muitas situações.

A regressão logística tem uma única variável estatística composta de coeficientes estimados para cada variável independente, como na regressão múltipla. Tal variável estatística é estimada de uma maneira diferente. A regressão logística deriva seu nome da transformação logit usada com a variável dependente, criando diversas diferenças no processo de estimação.

Os coeficientes estimados para as variáveis independentes são estimados usando-se o valor *logit* ou a razão de desigualdades como medida dependente. Cada uma dessas formulações de modelo é exibida aqui:

$$\text{Logit}_{i} = \ln\left(\frac{\text{prob}_{\text{evento}}}{1-\text{prob}_{\text{evento}}}\right) = b_0 + b_1X_1 + \ldots + b_nX_n$$

Exemplo prático

Suponha um banco: Ele deseja fazer uma regra de classificação para o cliente que quer realizar um empréstimo. É importante que o banco saiba se esse cliente será inadimplente ou não. O banco poderá utilizar diversas informações de outros clientes que já tiveram empréstimos e honraram ou não com as suas obrigações.

As informações solicitadas pelo banco podem ser renda média mensal, profissão, número de cartões de crédito, estado civil, idade, número de filhos, história de empréstimos anteriores, casa própria ou não, marca e modelo do carro etc. Com base nessas variáveis, o banco consegue realizar a construção de uma regra que permitirá a classificação dos candidatos a empréstimos em possíveis inadimplentes ou não.

A hipótese a ser testada será: Dispomos de duas populações e um conjunto de observações independentes de cada população. Sendo assim, seja X uma variável aleatória; supomos para a primeira população distribuição normal, $N(\mu_1, \sigma^2)$, e, para a segunda, $N(\mu_2, \sigma^2)$.

X pode ser a renda dos clientes que estão pleiteando um empréstimo no banco. Em dados de anos anteriores, na população 1 estavam os adimplentes nos contratos de empréstimos, e na população 2 estavam inadimplentes. Com estas informações, queremos classificar os novos clientes, com base na renda deles.´

Calcula-se a razão de verossimilhança entre as duas populações, supondo a distribuição normal:

$$\lambda(x) = \frac{f_1(x)}{f_2(x)} = \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2\right]\right\}$$

Para uma renda x , se $\lambda(x) > 1$, o valor da função densidade da população 1 é maior do que o da população 2; então é razoável classificar o cliente como sendo da população 1 , isto é, que honra com seus empréstimos.

Se $\lambda(x) < 1$, então é razoável classificar o candidato como sendo da população 2 , isto é, inadimplente. Se $\lambda(x) = 1$, então o cliente pode ser classificado tanto na população 1 quanto na 2. Seriam necessárias mais informações.

No exemplo, acima suponhamos que $\mu_1 = 4, \sigma^2 = 4$ e $\mu_2 = 1$; se um cliente obtém renda de 3 mil reais, em qual população ele será classificado?

$$\lambda(3) = \exp\left\{-\frac{1}{2}\left[\left(\frac{3-4}{2}\right)^2 - \left(\frac{3-1}{2}\right)^2\right]\right\} = 3,08 > 1$$

Sendo assim, como $\lambda(3) > 1$, o valor da função densidade da população 1 é maior do que o da população 2; então é razoável classificar o cliente como sendo da população 1 , de clientes adimplentes, podendo ser concedido o empréstimo a ele.

Se fôssemos realizar análise discriminante no R , utilizaríamos a função Análise discriminante linear (LDA), conforme abaixo:

```
plain-text
lda(Y~modelo,CV=TRUE)
```

Ou seja:

```
plain-text
lda(x, ...)
#Método para uma fórmula
lda(fórmula, data, ..., subset, na.action)

# Método padrão
lda(x, grouping, prior = proportions, tol = 1.0e-4,
method, CV = FALSE, nu, ...)
# Método por matrizes
lda(x, grouping, ..., subset, na.action)
```

Para realizar a regressão logística no R, a seguir demonstramos a descrição da equação utilizando o comando com a sintaxe básica:

```
plain-text
glm(Y~modelo, family=binomial(link="logit"))
```

Utiliza-se uma função para Modelos Lineares Generalizados (glm - em inglês *Generalized Linear Models*), determinando a variável dependente (y_bin), as variáveis independentes ($x1 + x2 + x3$), a base de dados a ser utilizada (data=mydata) e a família dos modelos (family = binomial(link="logit")):

```
plain-text
logit=glm(y_bin~x1+x2+x3, data=mydata, family = binomial(link="logit"))
summary(logit)
```

No vídeo a seguir, confira uma exemplo prático.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Verificando o aprendizado

Questão 1

Em uma análise de risco de crédito de um conjunto de empresas foram consideradas duas variáveis, sendo elas o capital social da empresa e o número de ações negociadas na bolsa. Os scores discriminantes das empresas com bom risco crédito são acima de 2 , ou seja, $Z_{jk} > 2$. A função discriminante é apresentada abaixo:

$$< br > Z_{jk} = 0,5 + 0,6X_{1k} + 0,4X_{2k} < br >$$

Em que,

$Z_{jk} =$ escore Z discriminante da função discriminante j para o objeto k

$X_{1k} =$ Capital social

$X_{2k} =$ Número de ações negociadas em bolsa

Uma empresa com capital social de 5 milhões e com 2 milhões de ações negociadas em bolsa será classificada com um bom ou ruim risco de crédito?

Chave de resposta

A função discriminante é uma variável estatística das variáveis independentes selecionadas por seu poder discriminatório usado na previsão de pertinência ao grupo. O valor previsto da função discriminante é o escore Z discriminante, o qual é calculado para cada objeto (pessoa, empresa ou produto) na análise. Ele toma a forma da equação linear:

$$Z_{jk} = a + W_1X_{1k} + W_2X_{2k} + \ldots. + W_nX_{nk}$$

Em que,

$Z_{jk} =$ escore Z discriminante da função discriminante j para o objeto k

a = intercepto

$W_i =$ peso discriminante para a variável independente i

$X_{ik} =$ variável independente i para o objeto k

Os scores discriminantes das empresas com bom risco de crédito são acima de 2, ou seja, Zjk > 2. A função discriminante é apresentada abaixo:

$$Z_{jk} = 0,5 + 0,6X_{1k} + 0,4X_{2k}$$

Em que,

$Z_{jk} =$ escore Z discriminante da função discriminante j para o objeto k

$X_{1k} =$ Capital social

$X_{2k} =$ Número de ações negociadas em bolsa

Para uma empresa com capital social de 5 milhões e com 2 milhões de ações negociadas em bolsa, teremos o seguinte score:

$$Z_k = 0,5 + 0,6 \times 5 + 0,4 \times 2 = 4,3$$

Como $Z_k = 4,3 > 2$, então a empresa será classificada com um bom risco de crédito.

Questão 2

Qual a diferença entre análise discriminante e análise de regressão?

Chave de resposta

A análise discriminante é quase a mesma da análise de regressão, ou seja, a função discriminante é uma combinação linear (variável estatística) de medidas métricas para duas ou mais variáveis independentes, sendo usada para descrever ou prever uma única variável dependente.

A diferença chave é que a análise discriminante é adequada a problemas de pesquisa nos quais a variável dependente é categórica (nominal ou não métrica), ao passo que a regressão é usada quando a variável dependente é métrica.

A regressão logística é uma variante da regressão, tendo assim muitas semelhanças, exceto pelo tipo de variável dependente, que seria o tipo de regressão da análise discriminante.

Questão 3

Um banco deseja fazer uma regra de classificação para o cliente que quer abrir uma conta. O banco utilizou diversas informações de outros clientes que já tiveram contas e honraram ou não com as suas obrigações. Construa uma regra que permitirá a classificação dos clientes para a abertura da conta ou não.

Levando-se em consideração a variável valor a ser investido, suponhamos que $\mu_1 = 16, \sigma_2 = 4$ e $\mu_2 = 18$, se um cliente tem o valor de 16,5 mil reais.

Chave de resposta

Se $\lambda(x) > 1$, o valor da função densidade da população 1 é maior do que o da população 2; então é razoável classificar o cliente como sendo da população 1 , isto é, que honra com seus empréstimos. Se $\lambda(x) < 1$, então é razoável classificar o candidato como sendo da população 2 , isto é, inadimplente.

Se $\lambda(x) = 1$, então o cliente pode ser classificado tanto na população 1 quanto na 2. Levando-se em consideração a variável valor a ser investido, suponhamos que $\mu_1 = 16, \sigma_2 = 4$ e $\mu_2 = 18$, se um cliente tem o valor de 16,5 mil reais,

$$\lambda(16,5) = \exp\left\{-\frac{1}{2}\left[\left(\frac{16,5-16}{2}\right)^2 - \left(\frac{16,5-18}{2}\right)^2\right]\right\} = 1,284 > 1$$

O cliente pertence à população 1.

Conceito MANOVA

No vídeo a seguir, conheça o conceito de MANOVA.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

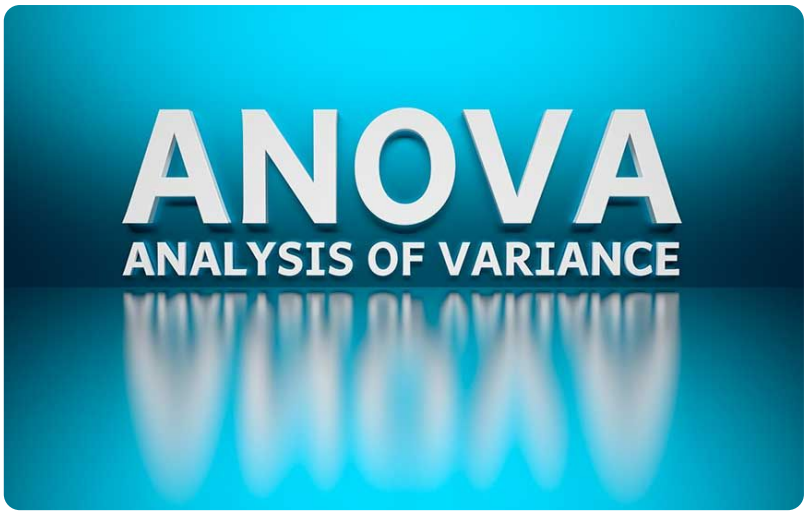
ANOVA é uma técnica de dependência que mede as diferenças para duas ou mais variáveis dependentes métricas, com base em um conjunto de variáveis categóricas (não métricas) que atuam como variáveis independentes.

A análise multivariada de variância é um teste que analisa a relação entre diversas variáveis de resposta e um conjunto comum de preditores ao mesmo tempo. Ela possui diversas vantagens ao invés de se usar várias análises de variância simples, como maior potência, pois é plausível usar a estrutura de covariância dos dados entre as variáveis de resposta para testar, ao mesmo tempo, a igualdade das médias.

> Se a variável de resposta estiver correlacionada, este dado adicional pode ajudar a detectar altercações pequenas para serem detectadas por meio de várias ANOVA individuais.

Outras vantagens são:

- ANOVA detecta padrões de respostas multivariadas, pois o fator pode afetar a relação entre as respostas, em vez de afetar uma única resposta.
- Ela controla a taxa de erro de família, pois a sua chance de rejeitar incorretamente a hipótese nula aumenta a cada ANOVA sucessiva. Fazer uma MANOVA para testar todas as variáveis de resposta ao mesmo tempo mantém a taxa de erro de família igual em seu nível alfa.
- Fornecer maior poder estatístico do que ANOVA quando o número de variáveis dependentes for 5 ou menos.
- Variáveis independentes não métricas criam grupos entre os quais as variáveis dependentes são comparadas; muitas vezes, os grupos representam variáveis experimentais ou "efeitos de tratamento".
- Pesquisadores devem incluir somente variáveis dependentes com forte suporte teórico.
- Controlar a taxa de erro experimental quando algum grau de intercorrelação entre variáveis dependentes está presente.



Sendo assim, a análise de variância multivariada é utilizada para comparar vetores de médias. As observações são provenientes de esboços estatísticos. A formulação de um teste estatístico para comparar vetores de médias depende da partição do total da variância em: Variância devido ao efeito de tratamentos e variância devido ao erro.

Um ponto relevante da análise multivariada é o aproveitamento da informação conjunta das variáveis envolvidas. Assim como ANOVA, MANOVA está interessada em diferenças entre grupos (ou tratamentos experimentais).

São representadas nesta forma geral:

ANOVA:

$$Y = X1 + X2 + \ldots + Xn$$

(Métrica) (Não métricas)

MANOVA:

$$Y1 + Y2 + \ldots + Yn = X1 + X2 + \ldots + Xn$$

(Métrica) (Não métricas)

MANOVA tem também um papel em planejamentos não experimentais, como em levantamentos de informações, em que grupos de interesse como sexo, realizar ou não uma ação, são definidos. As diferenças em qualquer número de variáveis métricas, como atitudes, satisfação e taxa de compras, são avaliadas quanto à significância estatística.

Para aplicação da análise de variância multivariada MAONOVA, devemos observar as seguintes suposições:

a) O modelo deve ser aditivo para os efeitos do tratamento, tanto os blocos quanto os erros.

b) Independência dos erros.

c) A matriz de covariância deve ser constante para todas as amostras.

d) Os erros ou resíduos devem ter distribuição multinormal.

Em resumo, existem vários benefícios de aplicação da análise de variância multivariada, sendos os principais:

- MANOVA pode detectar diferenças combinadas não encontradas nos testes univariados.
- Se múltiplas variáveis estatísticas são formadas, então elas podem fornecer dimensões de diferenças que podem distinguir entre os grupos melhor do que variáveis isoladas.
- Se o número de variáveis dependentes for mantido relativamente baixo (5 ou menos), o poder estatístico dos testes de MANOVA se iguala ou excede aquele obtido com uma única ANOVA.

As quatro medidas mais usadas para avaliar significância estatística entre grupos quanto às variáveis independentes são:

1. A maior raiz característica de Roy;

2. Teste lambda de Wilks;

3. Critério de Pillai;

4. T2 de Hotelling.

A seguir, aprofunde sobre as diferenças da MANOVA e ANOVA.

> **Conteúdo interativo**
>
> Acesse a versão digital para assistir ao vídeo.

Tanto para o teste lambda de Wilks quanto para o teste T_2 de Hotelling, a hipótese que queremos testar, ou seja, H_0 a ser testada, considera k tratamentos e p variáveis; essa hipótese é que os vetores de médias de tratamentos são iguais:

$$H_0 = \widetilde{\mu}_1 = \widetilde{\mu}_2 = \ldots = \widetilde{\mu}_k$$

$$H_0 = \begin{bmatrix} \mu_{11} \\ \ldots \\ \mu_{1p} \end{bmatrix} = \begin{bmatrix} \mu_{21} \\ \ldots \\ \mu_{2p} \end{bmatrix} = \ldots = \begin{bmatrix} \mu_{k1} \\ \ldots \\ \mu_{kp} \end{bmatrix} \quad \text{Sendo que } \mu_{kp} \text{ é a média da variável p do grupo } K.$$

As técnicas univariadas para análise de diferenças de grupos são o teste t, quando forem dois grupos, e a análise de variância, ANOVA, para dois ou mais grupos.

> Os procedimentos multivariados equivalentes são o T^2 de Hotelling ea análise multivariada de variância, MANOVA.

O teste t é utilizado para avaliar a significância estatística da diferença entre duas médias amostrais para uma única variável dependente. O teste t é um caso especial de ANOVA para dois grupos ou níveis de uma variável de tratamento.

Para o caso multivariado, o teste T^2 de Hotelling é utilizado para avaliar a significância estatística da diferença nas médias de duas ou mais variáveis entre dois grupos. É um caso especial de MANOVA usado com dois grupos ou níveis de uma variável de tratamento.

O teste T^2 de Hotelling segue uma distribuição conhecida sob a hipótese nula de nenhum efeito de tratamento sobre qualquer uma de um conjunto de medidas dependentes. Essa distribuição se transforma em uma distribuição F com pe $(N_1 + N_2 - 2 - 1)$ graus de liberdade após ajuste (em que $p =$ número de variáveis dependentes).

Para conseguir o valor crítico para o teste T^2 de Hotelling, encontramos o valor tabelado para Fcrit em um nível a especificado, e computamos T2crit como se segue:

$$T^2_{crit} = \frac{p(N_1+N_2-2)}{N_1+N_2-p-1} \times F_{crit} \quad \text{Sendo que } N_1 \text{ e } N_2 \text{ são o tamanho de cada amostra.}$$

Ao considerar o modelo linear multivariado na forma matricial:

$$Y = XB + \varepsilon$$

Em que,

Y é a matriz das observações;

X é a matriz dos delineamentos;

B é a matriz dos coeficientes;

E é a matriz dos erros.

Podemos encontrar a matriz da soma dos quadrados e produto de resíduos denotada por E:

$$E = \hat{\varepsilon}'\hat{\varepsilon}$$

Teremos assim o modelo multivariado:

$$A = H + B + E$$

Sendo que A, H, B e E serão as matrizes com dimensão $p \times p, A$ é a da soma de quadrados e produtos de totais, H dos tratamentos, B dos blocos e E dos resíduos.

O teste lambda de Wilks é o mais utilizado para verificar a hipótese H_0 da análise de variância multivariada MANOVA. O teste de Wilks é representado pela letra grega $\wedge$ (lambda maiúsculo) e pode ser definido de acordo com a seguinte estatística:

$$\Lambda = \frac{\det(E)}{\det(H+E)} = \frac{|E|}{|H+E|}$$

Atenção

Na presença de diferenças sistemáticas entre tratamentos, espera-se sempre obter , e tanto mais significativo quanto menor for seu valor. A regra de decisão é: Rejeita-se a hipótese nula ao nível de significância a se , caso contrário não se rejeita , sendo o teste significativo ao nível de significância .

Processo de decisão para MANOVA

O processo de executar uma análise multivariada de variância é semelhante ao encontrado em muitas outras técnicas multivariadas e, por isso, pode ser descrito por meio do processo de seis estágios para a construção de modelo.

O processo começa com a especificação dos objetivos da pesquisa. Segue então com várias questões do projeto que uma análise multivariada demanda e prossegue com uma análise das suposições inerentes a MANOVA. Com tais questões abordadas, o processo continua com a estimação do modelo MANOVA e a avaliação do ajuste geral do modelo.

Quando um modelo MANOVA aceitável é encontrado, os resultados podem ser interpretados em maiores detalhes. O passo final envolve esforços para validar os resultados para garantir generalização para a população.

Observe os estágios com mais detalhes:

Estágio 1

No estágio 1 a seleção de MANOVA é baseada no desejo de analisar uma relação de dependência representada como as diferenças em um conjunto de medidas dependentes ao longo de uma série de grupos formados por uma ou mais medidas independentes categóricas.

Estágio 2

No estágio 2 verificam-se questões do planejamento de pesquisa, como tamanho amostral adequado por grupo, o uso de covariáveis de seleção e de tratamentos (variáveis independentes). Então decidimos entre usar a ANOVA ou MANOVA.

Estágio 3

No estágio 3 analisamos as suposições, como independência, homogeneidade de matrizes de variância/covariância, normalidade, linearidade/multicolinearidade de variáveis dependentes e sensibilidade a observações atípicas.

Estágio 4

No estágio 4, ao estimar a significância de diferenças de grupos, selecionamos critérios para testes de significância, avaliamos o poder estatístico, aumentando-o, fazemos o uso em planejamento e análise, bem como verificamos os efeitos de multicolinearidade de variáveis dependentes.

Estágios 5 e 6

Nos estágios 5 e 6, de interpretação das variáveis e validação dos resultados, avaliamos as covariáveis e o impacto de variáveis independentes, repetimos o processo se for necessário e analisamos as amostras patrocinadas.

Aplicação da MANOVA

No vídeo a seguir, veja um exemplo de análise de variância múltipla.

Conteúdo interativo

Acesse a versão digital para assistir ao vídeo.

Considere que dois tratamentos (sexo e idade) são usados para examinar a escolha de um certo produto. Uma interação ordinal acontece, por exemplo, quando mulheres estão sempre mais propensas a comprar esse produto do que homens, mas a diferença entre homens e mulheres difere de acordo com a faixa etária do grupo.

Se desejamos testar as diferenças de grupos individualmente para cada uma das variáveis dependentes, podemos usar a raiz quadrada de T^2_{crit} como o valor crítico necessário para estabelecer significância.

Podemos utilizar o R para o cálculo desse valor:

$x <- \text{mvnorm}(n = n$, mean $= mu)$ hotellingOneSample(x)

O resultado traz o valor de T^2_{crit} e do p-valor:

```
plain-text
# # $statistic
# # [,1]
# # [,1] 804
# #
# # $pvalue
# # [,1]
# # [,1] 4e-15
```

Mas como é feita a análise? Através do p-valor.

Se estivermos trabalhando com uma significância de 0,05 (ou seja, 95% de probabilidade = 1-0,95=0,05 de significância ou com 99% de probabilidade = 1-0,99=0,01 de significância), se o p-valor for menor do que o nível de significância, rejeita-se a hipótese nula de que as médias são iguais.

No exemplo acima, para um nível de significância de 5%, o p-valor é de 0,00, que é menor do que 0,05. Então rejeita-se a hipótese nula de que as médias dos grupos de sexo são iguais.

Verificando o aprendizado

Quetsão 1

Descreva a análise de variância multivariada MANOVA e explique as principais diferença da ANOVA.

Chave de resposta

ANOVA é uma técnica de dependência que mede as diferenças para duas ou mais variáveis dependentes métricas, com base em um conjunto de variáveis categóricas (não métricas) que atuam como variáveis independentes; é um teste que analisa a relação entre diversas variáveis de resposta e um conjunto comum de preditores ao mesmo tempo.

ANOVA é uma extensão da análise de variância (ANOVA) para acomodar mais de uma variável dependente. Possui diversas vantagens ao invés de usar várias análises de variância simples, como maior potência, pois é plausível usar a estrutura de covariância dos dados entre as variáveis de resposta para testar, ao mesmo tempo, a igualdade das médias.

Se a variável de resposta estiver correlacionada, este dado adicional pode ajudar a detectar altercações pequenas para serem detectadas por meio de várias ANOVAs individuais. Assim como ANOVA, MANOVA está interessada em diferenças entre grupos (ou tratamentos experimentais). São representadas nesta forma geral:

ANOVA:

$$Y = X_1 + X_2 + \ldots + X_n$$

(Métrica) (Não métricas)

MANOVA:

$$Y_1 + Y_2 + \ldots + Y_n = X_1 + X_2 + \ldots + X_n$$

(Métrica) (Não métricas)

Questão 2

Descrever o processo de decisão para MANOVA:

Chave de resposta

O processo começa com a especificação dos objetivos da pesquisa. Segue então com várias questões do projeto que uma análise multivariada demanda e prossegue com uma análise das suposições inerentes a MANOVA. Com tais questões abordadas, o processo continua com a estimação do modelo MANOVA e a avaliação do ajuste geral do modelo.

Quando um modelo MANOVA aceitável é encontrado, os resultados podem ser interpretados em maiores detalhes. O passo final envolve esforços para validar os resultados para garantir generalização para a população.

Questão 3

Considere que dois tratamentos (vendas lojas físicas ou vendas lojas virtuais e idade) são usados para examinar o tipo de venda de um certo produto. Uma interação ordinal acontece, por exemplo, quando vendas são por meio de lojas físicas ou virtuais, mas a diferença entre esses meios de venda difere de acordo com a faixa etária do grupo.

Testar com probabilidade de 95% as diferenças de grupos individualmente para cada uma das variáveis dependentes, sabendo que o resultado traz o valor de T^2_{crit} e do p-valor:

$statistic

[,1]

[1,] 105

##

$pvalue

[,1]

[1,] 4e-3

Chave de resposta

Como estamos trabalhando com uma significância de 0,05 (ou seja, 95% de probabilidade = 1-0,95=0,05 de significância), se o p-valor for menor do que o nível de significância, rejeita-se a hipótese nula de que as médias são iguais.

No exemplo acima, para um nível de significância de 5%, o p-valor é de 0,004, que é menor do que 0,05. Então rejeita-se a hipótese nula de que as médias dos grupos de vendas em loja física ou *on-line* são iguais.

Considerações finais

O que você aprendeu neste conteúdo?

- A utilização de modelos estatísticos multivariados como suporte à tomada de decisões.
- A análise de regressão múltipla como ferramenta para prever variáveis contínuas.
- A análise discriminante e a regressão logística na classificação de grupos com variáveis categóricas.
- A MANOVA como método de comparação simultânea de médias de múltiplas variáveis dependentes.
- A importância das suposições estatísticas para a validade e robustez dos modelos.
- A integração das técnicas em contextos aplicados de negócios, saúde e ciências sociais

Explore +

- Pesquise na internet, sites, vídeos e artigos relacionados ao conteúdo visto.
- No site da UFMG existe o manual de introdução ao R com exemplos práticos de aplicação.
- Em caso de dúvidas, converse com seu professor online por meio dos recursos disponíveis no ambiente de aprendizagem.
- Acesse o site da UFMG e confira o manual de introdução ao R com exemplos práticos de aplicação.

Referências

HAIR JR., J.F. et al. **Análise Multivariada de Dados.** 6. ed. Porto Alegre, Bookman, 2009.

MINGOTI, S. **Análise de dados através de métodos de estatística multivariada.** Belo Horizonte: Editora UFMG, 2013.