ANALYTICS E MODELOS NÃO SUPERVISIONADOS DE MACHINE LEARNING

Clustering

Métodos Hierárquicos de Agrupamento

Métodos Não Hierárquicos (K-Means) de Agrupamento

Questão #1 — 1

O método de Ward é um dos métodos de agrupamento hierárquico. Com relação à este método, é **CORRETO** afirmar:

- Neste método, agrupa-se os indivíduos que possuem a mesma mediana.
- **Tem como base a análise da variância. ←**
- O critério de agrupamento deste método é a maior distância.
- Este método busca aumentar a variância dentro dos grupos.

Sobre a Análise de Cluster ou Análise de Agrupamentos, analise as afirmativas a seguir e escolha a alternativa correta:

1 - A análise de cluster ou análise de agrupamentos, por envolver cálculos de medidas de distância e variabilidade, deve ser feita apenas com variáveis numéricas.

2 - A análise de cluster ou análise de agrupamentos, por envolver cálculos de medidas de distância e variabilidade, deve ser feita apenas com variáveis categóricas e qualitativas.

- **1 - verdadeiro, 2 - falso ←**
- 1 - falso, 2 - falso
- 1 - falso, 2 - verdadeiro
- 1 - verdadeiro, 2 - verdadeiro

A análise do coeficiente R^2 nos permite obter algumas informações sobre os agrupamentos realizados. Com relação à este coeficiente, é **CORRETO** afirmar:

I. Quanto maior for o coeficiente R^2, melhor foi a forma de agrupamento realizado.

II. O SSR representa a variabilidade dentro do grupo.

III. O coeficiente R^2 não é apropriado para análise de variabilidade.

- **Somente I e II estão corretas. ←**
- Somente III está correta.
- Somente I está correta.
- Todas as alternativas estão corretas.

Considere um banco de dados que mostra, por aluno, a nota nas provas de Matemática e Português. Suponha que nosso objetivo é agrupar esses alunos por desempenho nas disciplinas para traçar um plano de reuniões com cada grupo.

Aluno	Matemática	Português
Adriana	9	7
Aline	5	4
Bruno	6	6
Michel	10	8
Regina	4	4
Zé	4	9

Sabemos que para definir os grupos é necessário calcular a distância euclidiana entre dois alunos, dada pela fórmula:

$$D = \sqrt{(x_{11} - x_{12})^2 + (x_{21} - x_{22})^2}$$

Sabendo disso, qual a distância euclidiana entre a Adriana e a Aline?

- 9,31
- **5,00 ←**
- 5,75
- 7,77

É exemplo de Técnica Supervisionada de Machine Learning:

- Análise de Cluster
- Análise de Correspondência Simples e Múltipla
- **Regressão Linear/ Logística ←**
- Principal Component Analysis (PCA)

É exemplo de Técnica não Supervisionadas de Machine Learning:

- Regressão Linear
- **Análise de Cluster/Agrupamento ←**
- Regressão Logística
- Árvore de decisão

O dendograma permite a visualização de como os agrupamentos foram acontecendo. Com relação ao dendograma, é **CORRETO** afirmar:

- Se a distância no dendograma aumentou, isso significa que tanto a média quanto a mediana aumentaram.
- **Se a distância no dendograma aumentou, isso significa que a variância (variabilidade) aumentou. ←**
- Se a distância no dendograma aumentou, isso significa que a média aumentou.
- Se a distância no dendograma aumentou, isso significa que a mediana aumentou.

A análise de cluster é uma técnica estatística que busca classificar os indivíduos em grupos, de forma que os indivíduos dentro de um mesmo cluster (grupo) sejam muito parecidos, e os indivíduos em diferentes clusters (grupos) sejam distintos entre si. Com relação à ferramenta de análise dendrograma, é **CORRETO** afirmar:

I. É uma ferramenta que possibilita a visualização de como os agrupamentos foram acontecendo.

II. É uma ferramenta numérica de análise de clusters (grupos).

III. O dendograma passa a ser inviável para grande quantidade de indivíduos na base de dados.

- Todas as alternativas estão corretas.
- **Somente I e III estão corretas. ←**
- Somente II está correta.
- Somente I está correta.

O coeficiente R^2 auxilia na análise de agrupamentos. Com relação à este coeficiente é **CORRETO** afirmar:

- O coeficiente R^2 analisa as médias dos grupos.
- **O coeficiente R^2 mostra o quanto da variabilidade nós conseguimos entender. ←**
- O coeficiente R^2 analisa tanto a média quanto a mediana dos grupos.
- O coeficiente R^2 analisa as medianas dos grupos.

As técnicas para escolha do número de clusters (grupos) nos auxilia na determinação da quantidade de grupos que devem ter na análise de agrupamentos. Com relação ao coeficiente de silhueta é **CORRETO** afirmar:

- **O coeficiente de silhueta é a medida da relação entre um ponto e os membros do grupo dele. ←**
- O coeficiente de silhueta compara o R^2 calculado com o seu esperado.
- No coeficiente de silhueta é necessário traçar a curva de wss de acordo com o número de clusters.
- O coeficiente de silhueta mostra o quanto da variabilidade nós conseguimos entender.

Sobre a Distância Euclidiana, analise as afirmativas a seguir e escolha a alternativa correta:

1 - Essa distância gera a distância linear entre quaisquer dois pontos em um campo com k dimensões.

2 - É uma generalização do Teorema de Pitágoras.

- 1 - falso, 2 - verdadeiro
- 1 - verdadeiro, 2 - falso
- 1 - falso, 2 - falso
- **1 - verdadeiro, 2 - verdadeiro ←**

Os métodos de agrupamento têm como objetivo agrupar indivíduos com características similares em grupos (clusters). Assim, o método centróide tem como objetivo:

- **Agrupar indivíduos que possuem a menor distância. ←**
- Agrupar indivíduos que possuem a maior média.
- Agrupar indivíduos que possuem a maior distância.
- Agrupar indivíduos que possuem a menor média.

É exemplo de Técnica não Supervisionadas de Machine Learning:

- Árvore de decisão
- Regressão Logística
- **Análise de Cluster/Agrupamento ←**
- Regressão Linear

Em relação às **técnicas de agrupamentos hierárquicos**, avalie as técnicas apresentadas:

1. **Complete Linkage** é a técnica de agrupamento que se baseia no cálculo do vizinho mais próximo

2. **Average Linkage** é a técnica de agrupamento que se baseia na média da distância

3. **Ward's method** é uma técnica de agrupamento não hierárquico

É correto afirmar que:

- somente os itens 2 e 3 estão corretos
- somente os itens 1 e 2 estão corretos
- **somente o item 2 está correto ←**
- todos os itens estão corretos

O processo de análise de cluster demanda alguns passos. Com relação a este processo, é **CORRETO** afirmar:

I. O primeiro passo é entender o problema a ser resolvido.

II. Uma vez que o problema tenha sido definido, é necessário procurar variáveis numéricas que possam ser importantes na análise de cluster.

III. Um ponto crítico na análise de cluster é a verificação se as variáveis estão na mesma escala.

- Somente I está correta.
- Somente III está correta.
- **Todas as alternativas estão corretas. ←**
- Somente II está correta.

Ao realizar a análise de cluster é necessário verificar se as variáveis estão na mesma escala, pois caso contrário, isso afetará a distância Euclidiana. Com relação à esta distância, é **CORRETO** afirmar:

I. Quando as variáveis estão em escalas diferentes, por exemplo: idade e renda, não é necessária a padronização, pois não haverá alteração na distância Euclidiana em comparação com a padronização destas variáveis.

II. Quando temos duas variáveis em escalas diferentes, a variável que tem menor valor terá maior impacto na distância Euclidiana.

III. Quando as variáveis estão em escalas diferentes é necessário padronizá-las.

- Somente I e III estão corretas.
- Somente I está correta.
- Todas as alternativas estão corretas.
- **Somente III está correta. ←**

Considere a seguinte rotina e, a seguir, assinale a alternativa que melhor se adequa ao que se está sendo comandado:

hc3 <- hclust(d, method = "average")

Considerando que:

d: refere-se a uma matriz de distâncias

- ( ) Com esta rotina, o R fará a análise de cluster hierárquico com base no método Single Linkage.
- ( ) Esta rotina fará uma análise de cluster não hierárquico com base no método Centróide.
- (•) **Por meio desta rotina, o R fará a análise de cluster hierárquico com base no método Avarage Linkage. ←**
- ( ) É certo que esta rotina se propõe a fazer uma análise de cluster não hierárquico com base no método Complete Linkage.

Leia atentamente o postulado por Fávero e Belfiore (2017) e, assinale a alternativa que melhor completa o trecho em branco.

"Dentre os esquemas de aglomeração não hierárquicos, o procedimento _____ é o mais utilizado por pesquisadores em diversos campos do conhecimento. Dado que a quantidade de clusters é definida preliminarmente pelo pesquisador, esse procedimento pode ser elaborado após a aplicação de um esquema hierárquico aglomerativo quando não se tem ideia da quantidade de clusters que podem ser formados e, nessa situação, o output obtido por esse procedimento pode servir de input para o não hierárquico."

- (•) **k-means ←**
- ( ) single linkage
- ( ) zscores
- ( ) de ponderação arbitrária

Assinale a alternativa que completa corretamente as afirmações a seguir. "Os esquemas de aglomeração podem ser classificados, basicamente, em dois tipos, conhecidos por _____ e _____. Enquanto que os primeiros caracterizam- se por privilegiar uma estrutura _____ (passo a passo) para a formação dos agrupamentos, os esquemas _____ utilizam algoritmos para maximizar a _____ dentro de cada agrupamento."

- ( ) não hierárquicos - hierárquico - não hierárquicos - hierárquicas - heterogeneidade
- (•) **hierárquicos - não hierárquicos - hierárquica - não hierárquicos - homogeneidade ←**
- ( ) não hierárquicos - hierárquico - não hierárquicos - hierárquicas - homogeneidade
- ( ) hierárquicos - não hierárquicos - hierárquica - não hierárquicos - heterogeneidade

Segundo Fávero e Belfiore (2019), o pesquisador interessado em aplicar uma análise de agrupamentos necessita, a partir da definição dos objetivos de pesquisa, escolher determinada medida de distância ou de semelhança, que servirá de base para que as observações sejam consideradas menos ou mais próximas, e determinado esquema de aglomeração, que deverá ser definido entre os métodos hierárquicos e não hierárquicos. Dessa forma, terá condições de analisar, interpretar e comparar os resultados.

Sobre os métodos hierárquicos e não hierárquicos, julgue os itens a seguir e selecione aquele que for mais adequado:

- Os esquemas não hierárquicos, partem de uma quantidade conhecida de clusters e, a partir de então, é elaborada a alocação das observações nesses clusters, com posterior avaliação da representatividade de cada variável para a formação deles.
- **Todas as alternativas estão corretas. ←**
- O resultado de um método hierárquico pode servir de input para a realização de um método não hierárquico, tornando a análise cíclica.
- Os esquemas hierárquicos permitem a identificação do ordenamento e da alocação das observações, oferecendo possibilidades para que o pesquisador estude, avalie e decida sobre a quantidade de agrupamentos formados.

Admita que uma amostra relativa ao volume de vendas de um grupo de 100 vendedores da empresa Alfa apresente distribuição normal. Tem-se que o desvio-padrão é R$ 27,00 e a média R$ 77,00. Qual o valor do Z-score de um vendedor que tenha feito uma venda de R$ 62,00?

Lembrando que a fórmula do Z score é: $$Z = \frac{(X-\mu)}{s}.$$

- O Z-score deste vendedor é 5.
- **O Z-score deste vendedor é -0,56. ←**
- O Z-score deste vendedor é 2,5.
- O Z-score deste vendedor é 1,56.

Não é exemplo de um método de encadeamento hierárquico:

- Complete linkage
- Ward's method
- **K-means ←**
- Single linkage

Assinale a alternativa correta:

- A técnica de Análise de Clusters é robusta quanto à presença de outliers.
- Antes de se elaborar uma Análise de Clusters, é ideal que os outliers da base de dados sejam identificados e removidos.
- **A exclusão, ou a retenção de outliers na base, dependerá dos objetivos de pesquisa e da natureza dos dados. ←**
- A padronização dos dados é técnica capaz de anular os efeitos dos outliers.

PCA (Principal Component Analysis)

Em relação às três assertivas a seguir, marque a alterativa correta:

I) Autovalores (eigenvalues) podem ser calculados pela soma em coluna das cargas fatoriais ao quadrado para um fator.

II) Autovalores (eigenvalues) representam a quantia de variância explicada por um fator.

III) Scores fatoriais, que entram no cálculo de determinado fator, são calculados a partir da definição do eigenvalue daquele mesmo fator.

- (●) **Todos os itens estão corretos. ←**
- ( ) Somente o item II está correto.
- ( ) Somente os itens I e III estão corretos.
- ( ) Somente os itens II e III estão corretos.

Imagine que, após a aplicação da técnica PCA em uma base de dados com cinco variáveis métricas, tenham sido extraídos dois fatores. As comunalidades resultantes do modelo citado, então, dirão respeito:

- ( ) À raiz quadrada dos eigenvalues.
- (●) **À variância total compartilhada entre cada uma das cinco variáveis e os dois fatores extraídos. ←**
- ( ) Às correlações existentes entre as cinco variáveis.
- ( ) Ao somatório das cargas fatoriais.

Analise os itens a seguir, e assinale a alternativa que contém apenas possíveis objetivos de uma Análise Fatorial PCA:

I – Redução dimensional dos dados;

II – Elaboração de rankings;

III – Confirmação de constructos;

IV – Predição para observações não presentes na amostra de treino do algoritmo.

- ( ) Apenas os itens I, III e IV estão corretos.
- ( ) Apenas os itens I e II estão corretos.
- ( ) Apenas o item I está correto.
- (●) **Apenas os itens I, II e III estão corretos. ←**

Ao fator com o maior percentual de variância compartilhada, damos o nome de:

- ( ) Autovalor.
- (●) **Fator principal. ←**
- ( ) Raiz latente.
- ( ) Comunalidade.

A Análise Fatorial por componentes principais:

- (●) **Pode ser aplicada para a criação de indicadores sociodemográficos, por exemplo, para posteriores plotagens em mapas em análise geoespacial. ←**
- ( ) Não pode ser considerada uma técnica de *machine learning unsupervised*.
- ( ) Não pode ser utilizada para a criação de *ronkings*.
- ( ) Não pode ser utilizada para a redução estrutural da base de dados.

Sobre a Análise Fatorial PCA, seria **correto** afirmar:

- ( ) Técnica ideal para se trabalhar com variáveis qualitativas.
- ( ) Objetiva o agrupamento de observações.
- ( ) Constitui-se em técnica supervisionada de machine learning.
- (●) **Constitui-se em técnica não-supervisionada de machine learning. ←**

Em relação à técnica de Análise Fatorial PCA, é **correto** afirmar:

- ( ) É uma técnica que não auxilia na redução estrutural de bases de dados.
- ( ) A esfericidade de Bartlett não indica a possibilidade de extração de fatores.
- ( ) O uso da esfericidade de Bartlett permite a inclusão de variáveis qualitativas.
- (●) **Os autovalores são extraídos de uma matriz de correlações de Pearson. ←**

Em relação à análise fatorial:

I) É muito útil para dados que tem correlação elevada e para quando se deseja criar novas variáveis que captem o comportamento do conjunto de dados.

II) Fatores são agrupamentos de variáveis segundo algum critério.

III) A técnica se encaixa em análise não supervisionada, o que significa que não é necessário ter uma variável resposta.

Assinale a alternativa **correta** em relação às três assertivas apresentadas:

- (●) **Todos os itens estão corretos. ←**
- ( ) Somente os itens I e III estão corretos.
- ( ) Somente os itens II e III estão corretos.
- ( ) Somente o item II está correto.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo. Ele é um modelo não supervisionado de PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo03.RData"

Para abri-lo, basta:

load("pca_arquivo03.RData")

summary(pca_arquivo03)

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **qual a quantidade máxima de fatores** poderia ser extraída?

Clique aqui para baixar o anexo.

- Nenhum.
- 2.
- 8.
- **11. ←**

As cargas fatoriais correspondem:

- Ao resultado da esfericidade de Bartlett.
- À soma das comunalidades.
- **Às correlações entre as variáveis originais e cada um dos fatores. ←**
- À razão entre os eigenvalues e os eigenvectors da matriz de correlação das variáveis.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo. Ele é um mapa de calor das correlações de uma base de dados, elaborado com auxílio da linguagem R.

Arquivo a ser baixado: "rho_arquivo01.RData".

Para abri-lo, basta:

load("rho_arquivo01.RData")

rho_arquivo01

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **observando apenas as correlações negativas**, com qual variável a variável *potássio_mg* guarda maior correlação?

Clique aqui para baixar o anexo.

- *fibras_g.*
- ***qtde_xicaras. ←***
- *sodio_mg.*

Imagine que, após a aplicação da técnica PCA em uma base de dados com cinco variáveis métricas, tenham sido extraídos dois fatores. **É possível afirmar que**:

- (●) **A divisão dos dois autovalores por 5, será uma *proxy* para indicar a quantidade de variabilidade total das variáveis originais que foram capturadas. ←**
- ( ) A soma dos dois autovalores será igual a 5.
- ( ) Pelo menos 3, das 5 variáveis, guardarão uma correlação mais alta com o primeiro fator.
- ( ) Pelo menos 4, das 5 variáveis, guardarão uma correlação mais alta com o primeiro fator.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo. Ele é um modelo não supervisionado de PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo02.RData"

Para abri-lo, basta:

```
load("pca_arquivo02.RData")
summary(pca_arquivo02)
```

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **ao se adotar o critério da raiz latente**, isto é, extrair fatores que possuam autovalores maiores que 1, pode-se dizer que a variável *conforto* possuirá uma correlação mais forte com quais dos fatores?

Clique aqui para baixar o anexo.

- ( ) F3
- (●) **F1 ←**
- ( ) F4

Em relação à PCA:

I) É objetivo da técnica dar nome aos fatores extraídos.

II) Fatores são agrupamentos de variáveis segundo algum critério.

III) O primeiro fator será, sempre, o que capturará a maior variância compartilhada pela base de dados.

Assinale a alternativa **correta** em relação às três assertivas apresentadas:

- ( ) Somente os itens I e III estão corretos.
- ( ) Todos os itens estão corretos.
- (●) **Somente os itens II e III estão corretos. ←**
- ( ) Somente o item II está correto.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo. Ele é um modelo não supervisionado de PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo01.RData"

Para abri-lo, basta:

load("pca_arquivo01.RData")

summary(pca_arquivo01)

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **ao se adotar o critério da raiz latente**, isto é, extrair fatores que possuam autovalores maiores que 1, dever-se-ia extrair quantos fatores?

Clique aqui para baixar o anexo.

- 3
- 1
- **2 ←**

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo. Ele é um modelo não supervisionado de PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo02.RData"

Para abri-lo, basta:

load("pca_arquivo02.RData")

summary(pca_arquivo02)

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, pode-se afirmar que o valor do *eigenvalue* (autovalor) associado ao fator principal é, aproximadamente, de:

Clique aqui para baixar o anexo.

- 1.5013.
- 1.9558.
- 0.4781.
- **3.825154. ←**

Home / Publicação de Avaliação / Unsupervised Machine Learning: Análise Fatorial e PCA II- 29/06/2021

PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo02.RData"

Para abri-lo, basta:

load("pca_arquivo02.RData")

summary(pca_arquivo02)

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **ao se adotar dois fatores**, qual o percentual de variância compartilhada da variável *atendimento* (comunalidade) foi capturada por esses dois fatores?

Clique aqui para baixar o anexo.

- -0.0647522.
- -0.3105135.
- **0.1006115. ←**
- 0.8727153.

PCA, estimado com auxílio da linguagem R.

Arquivo a ser baixado: "pca_arquivo02.RData"

Para abri-lo, basta:

load("pca_arquivo02.RData")

summary(pca_arquivo02)

***Para resolver a questão, é possível que somente os comandos anteriores não sejam suficientes para explorar o que o exercício pede. Utilize as técnicas e códigos aprendidos em aula, se necessário.**

De acordo com o arquivo baixado, **ao se adotar o critério da raiz latente**, isto é, extrair fatores que possuam autovalores maiores que 1, dever-se-ia extrair quantos fatores?

Clique aqui para baixar o anexo.

- **2 ←**
- 1
- 3
- 5

Análise de Correspondência Simples e Múltipla

Sobre a Análise de Correspondência é **correto** afirmar:

- **A Análise de Correspondências, utiliza de variáveis categóricas dispostas numa tabela de contingências, levando em conta medidas de associação entre suas linhas e colunas. ←**
- A Análise de Correspondências é ideal para modelar variáveis quantitativas.
- A Análise de Correspondências é uma técnica que objetiva o agrupamento das observações em razão de suas distâncias.
- A Análise de Correspondências é iniciada pela elaboração de uma matriz de correlações.

Como é calculado cada componente da matriz de resíduos (matriz R da aula)?:

- (●) **Valores observados menos valores esperados. ←**
- ( ) Valores esperados menos valores observados, divido pela raiz quadrada dos valores observados.
- ( ) Valores observados menos valores esperados, divido pela raiz quadrada dos valores esperados.
- ( ) Valores das correlações entre variáveis.

Em relação à técnica de Análise de Correspondências, é **correto** afirmar:

- ( ) É uma técnica adequada para modelarmos variáveis quantitativas.
- ( ) Estabelecido modelo, poderemos fazer inferências para observações não constantes na amostra utilizada para treinar o algoritmo.
- ( ) Um de seus objetivos é o agrupamento de observações.
- (●) **Um de seus objetivos é o estudo das associações entre categorias das variáveis utilizadas no modelo. ←**

Assinale a alternativa que melhor complementa o fragmento de texto a seguir:

"Imagine que um pesquisador tenha interesse em estudar a relação de interdependência entre duas variáveis categóricas, por exemplo, o comportamento de consumo, descrito pela preferência por determinados comportamentos de compra dos consumidores, e as vizinhanças desses consumidores. Nessa situação, _____".

- ( ) A Análise de Correspondências Múltiplas pode ser utilizada, uma vez que é uma **técnica multivariada** que permite investigar a associação entre duas, e somente duas, variáveis categóricas.
- ( ) A Análise de Correspondências Simples pode ser utilizada, uma vez que é uma **técnica multivariada** que permite investigar a associação entre duas, e somente duas, variáveis categóricas.
- ( ) A Análise de Correspondências Múltiplas pode ser utilizada, uma vez que é uma **técnica bivariada** que permite investigar a associação entre duas, e somente duas, variáveis categóricas.
- (●) **A Análise de Correspondências Simples pode ser utilizada, uma vez que é uma técnica bivariada que permite investigar a associação entre duas, e somente duas, variáveis categóricas. ←**

As **massas das linhas e das colunas**, podem ser entendidas como:

- (●) **São medidas influência de determinada categoria em relação às demais. ←**
- ( ) Correspondem à subtração dos valores observados pelos valores esperados.
- ( ) São maneiras de identificar se determinados cruzamentos de categorias, em linhas e colunas, possuem associação aleatória.
- ( ) São os somatórios das linhas e das colunas, individualmente.

Em relação à técnica de Análise de Correspondências, é **correto** afirmar:

- ( ) O uso do teste permite a inclusão de variáveis quantitativas.
- ( ) É uma técnica que não auxilia na redução estrutural de bases de dados.
- ( ) O teste não indica a possibilidade da elaboração de uma ANACOR.
- (●) **Os autovalores são extraídos por meio da decomposição da inércia total. ←**

Dos conjuntos de variáveis expostos a seguir, assinale a alternativa em que seja possível elaborar uma Análise de Correspondências:

- Quantidade de telefones celulares por município; percentual de aceitação de dado produto cosmético numa região; renda *per capta* de dado país.
- Nível de escolarização dos funcionários de um hospital; tempo gestacional de mamíferos; extensão das rodovias pavimentadas de alguns países.
- Números de telefone de algumas pessoas; as alturas de algumas pessoas; faixa de renda de algumas famílias.
- **Bairros de residência de algumas famílias; faixa de renda de algumas famílias; Classificação de astros em planetas. ←**

É certo que o teste Qui-Quadrado se propõe a:

- Verificar se a associação entre duas variáveis métricas se dá, ou não, de forma aleatória, a dado nível de significância.
- Verificar se a correlação entre duas variáveis métricas se dá, ou não, de forma aleatória, a dado nível de significância.
- **Verificar se a associação entre duas variáveis categóricas se dá, ou não, de forma aleatória, a dado nível de significância. ←**
- Verificar se a correlação entre duas variáveis categóricas se dá, ou não, de forma aleatória, a dado nível de significância.

Sobre o gráfico denominado **mapa perceptual** é correto afirmar:

I. No gráfico denominado mapa perceptual, temos representadas as coordenadas advindas dos dados, em linha e em coluna, de uma tabela cruzada a partir de técnicas de análise de correspondências.

II. A partir do mapa perceptual, é possível interpretar as similaridades e diferenças de comportamento entre variáveis e entre categorias.

III. Suas coordenadas são advindas de uma matriz de correlações entre variáveis.

Assinale a alternativa que traz os itens que estão corretos.

- **Somente os itens I e II estão corretos. ←**
- Somente o item II está correto.
- Somente o item III está correto.
- Somente os itens II e III estão corretos.

Sobre os resíduos padronizados ajustados, **é correto afirmar**:

- **A análise dos resíduos padronizados ajustados revelará os padrões característicos de cada categoria de uma variável segundo o excesso ou falta de ocorrências de sua combinação com cada categoria de outra variável. ←**
- Estuda se as associações entre as categorias das variáveis se associam, ou não, de forma aleatória.
- Indica a variância dos dados.
- Dizem respeito à diferença entre os valores observados e os valores esperados.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo02.RData")

De acordo com o arquivo baixado, quantas dimensões seriam necessárias para explicar a completude da inércia principal total dos dados?

Clique aqui para baixar o anexo.

- 5.
- 2.
- 3.
- **4. ←**

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo03.RData")

De acordo com o arquivo baixado, ao se adotar um mapa perceptual **tridimensional**, aproximadamente, qual o percentual da inércia principal total seria explicado?

Clique aqui para baixar o anexo.

- **84.31. ←**
- 62.78.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo03.RData")

De acordo com o arquivo baixado, ao se adotar um mapa perceptual **bidimensional**, aproximadamente, qual o percentual da inércia principal total seria explicado?

Clique aqui para baixar o anexo.

- 84.31.
- **62.78. ←**
- 39.61.
- 100.00.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo03.RData")

De acordo com o arquivo baixado, **respectivamente**, as coordenadas dos eixos das abcissas e das ordenadas para a categoria *Brazil**, são:

*Categoria das variáveis com grafia em inglês

- **-0.61494496 e 0.52175907. ←**
- -0.55852697 e 0.09164603.
- 0.3697534 e -0.444925181.
- -0.58832253 e 0.13198640.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo03.RData")

De acordo com o arquivo baixado, quantas dimensões seriam necessárias para explicar a completude da inércia principal total dos dados?

Clique aqui para baixar o anexo.

- 3.
- 5.
- **4. ←**
- 2.

no material complementar. Ele é um mapa perceptual bidimensional de uma ACM, estimada com auxílio da linguagem R.

Arquivo a ser baixado: **"acm_arquivo02.RData"**

Para abri-lo, basta:

load("acm_arquivo02.RData")

acm_arquivo02

De acordo com o arquivo baixado, pode-se sugerir que as intensidades menores do sintoma "febre" (*no fever* e *low fever)* estão mais fortemente associados a:

Clique aqui para baixar o anexo.

- Dengue.
- Coceira leve (mild itch).
- **Zika. ←**
- Chikungunya.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo02.RData")

De acordo com o arquivo baixado, ao se adotar um mapa perceptual tridimensional, aproximadamente, qual o percentual da inércia principal total seria explicado?

Clique aqui para baixar o anexo.

- 78.75.
- **99.68. ←**
- 100.00.
- 96.25.

Para abri-lo, basta:

library(tidyverse)

load("anacor_arquivo01.RData")

anacor_arquivo01.RData

De acordo com o arquivo baixado, **pode-se dizer** que as associações mais intensas (> 1,96) com o indicador CPC1, dar-se-ão para as:

Clique aqui para baixar o anexo.

- Universidades Públicas Federais.
- Universidades Privadas sem Fins Lucrativos.
- **Nenhuma das alternativas. ←**
- Universidades Públicas Estaduais.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ANACOR, estimado com auxílio da linguagem R.

Para abri-lo, basta:

load("anacor_arquivo02.RData")

De acordo com o arquivo baixado, respectivamente, as coordenadas dos eixos das abcissas e das ordenadas para a categoria Universidades Públicas Estaduais, são:

Clique aqui para baixar o anexo.

- (●) **0.13113995 e 0.21584486. ←**
- ( ) 0.16418011 e -0.12810522.
- ( ) 0.1598437 e -0.006749231.
- ( ) 0.41416700 e 0.31403773.

ATENÇÃO: Para responder a essa questão, baixe o arquivo de formato *.RData anexo à questão e/ou disponível no material complementar. Ele é um modelo não supervisionado de ACM, estimado com auxílio da linguagem R.

Arquivo a ser baixado: **"acm_arquivo01.RData"**

Para abri-lo, basta:

load("acm_arquivo01.RData")

De acordo com o arquivo baixado, caso assumíssemos um mapa perceptual bidimensional, qual seria, aproximadamente, o percentual da inércia principal total explicado?

Clique aqui para baixar o anexo.

- ( ) 20.63.
- (●) **52.29.**

ANALYTICS E MODELOS SUPERVISIONADOS DE MACHINE LEARNING: ESTIMAÇÃO POR OLS

Modelos Lineares de Regressão Simples e Múltipla

Modelos Não Lineares de Regressão e Transformação de Box-Cox;

Variáveis Explicativas Qualitativas e Variáveis Dummy;

Diagnósticos em Modelos de Regressão: Normalidade dos Resíduos;

Multicolinearidade e Heterocedasticidade;

Procedimento Stepwise.

Os principais critérios para a estimação de modelos de regressão por mínimos quadrados ordinários são:

- somatória dos erros sendo igual a zero e somatória dos erros ao quadrado sendo a máxima possível.
- somatória dos erros ao quadrado sendo igual a zero e somatória dos erros sendo a mínima possível.
- somatória dos erros sendo igual a zero e somatória dos erros ao quadrado sendo igual a zero.
- **somatória dos erros sendo igual a zero e somatória dos erros ao quadrado sendo a mínima possível. ←**

Sobre um modelo de regressão linear simples do tipo, marque a alternativa **correta**:

$$Y_i = \alpha + \beta . X_i + u_i$$

- o termo de erro u_i não tem qualquer função na equação.
- **o termo de erro u_i captura parte do comportamento da variável dependente *Y* que não foi devidamente explicado pela variável *X*. ←**
- o termo de erro u_i tem o valor igual a zero, necessariamente.
- o termo de erro u_i não representa o efeito de outras variáveis não incluídas na equação.

Assinale a alternativa **incorreta**:

- Modelos de regressão são técnicas que têm por finalidade, entre outros objetivos, a estimação de parâmetros para a definição de modelos com capacidade preditiva.
- **Modelos de regressão têm por finalidade principal a detecção de relações de causa e efeito sobre o fenômeno de estudo. ←**
- Modelos de regressão são adequados para fins preditivos dentro da interpolação dos dados presentes no dataset.
- Modelos de regressão são técnicas de machine learning consideradas supervisionadas.

São objetivos de uma regressão linear simples:

I. A estimação de uma equação linear que apresente a relação entre uma variável dependente e uma explicativa (preditora);

II. A estimação de uma equação exponencial que apresente a relação entre uma variável qualitativa e uma explicativa;

III. A estimação de uma equação dentro de um espaço de Hilbert utilizando operadores autoadjuntos ou hermitianos;

IV. O objetivo principal está na análise da relação entre duas variáveis. Essa análise sempre parte de uma variável chamada de dependente, e outra chamada de preditora (explicativa).

Marque a alternativa **correta**:

- Apenas II e IV estão corretas.
- **Apenas I e IV estão corretas. ←**
- Apenas I, II e IV estão corretas.
- Todas as afirmações estão corretas.

Assinale a alternativa **correta**:

- Modelos de regressão identificam relações de causa e efeito nos dados.
- Modelos de regressão são técnicas não supervisionadas de machine learning e, portanto, não são adequados para fins preditivos.
- Modelos de regressão são adequados para o estabelecimento de previsões e extrapolações para intervalos de dados que estejam fora da amplitude considerada no banco de dados.
- **Modelos de regressão simples representam a relação entre uma variável dependente e uma variável preditora. ←**

Sobre o método de mínimos quadrados ordinários para a estimação de parâmetros de modelos de regressão, é **correto** afirmar que:

- Somente podem ser aplicados para os casos em que as variáveis preditoras sejam quantitativas.
- Podem ser aplicados desde que a variável dependente seja dummy.
- **Podem ser aplicados desde que a variável dependente seja quantitativa. ←**
- Não podem ser aplicados para modelos de machine learning supervisionados.

Assinale a alternativa **correta**:

- Apenas a definição dos parâmetros já é suficiente para se determinar corretamente um modelo de regressão.
- Apenas a definição dos parâmetros e do R^2 já é suficiente para se determinar corretamente um modelo de regressão.
- Apenas a definição do R^2 já é suficiente para se determinar corretamente um modelo de regressão.
- **A definição dos parâmetros e do R^2 são insuficientes para se determinar corretamente um modelo de regressão, já que é preciso que sejam avaliadas as significâncias estatísticas dos parâmetros. ←**

Sobre um modelo de regressão linear simples marque a **alternativa verdadeira**.

- o objetivo da análise de regressão é encontrar um meio de condensar a informação contida em várias variáveis originais em um conjunto menor de variáveis estatísticas (fatores) com uma perda mínima de informação.
- é o método de análise apropriado quando o problema de pesquisa envolve mais de uma variável dependente métrica relacionada com apenas uma variável independente métrica.
- **é o método de análise apropriado quando o problema de pesquisa envolve a inferência do comportamento de uma única variável dependente quantitativa relacionada a uma variável explicativa. ←**
- o objetivo da análise de regressão é classi?car uma amostra de entidades (indivíduos ou objetos) em um número menor de grupos mutuamente excludentes, com base nas similaridades.

Trata-se de um modelo de **regressão linear simples**:

- Uma equação contendo diversas variáveis explicativas (variável X) multiplicadas, cada uma por um coeficiente angular, um coeficiente linear e um termo de erro.
- Uma equação contendo uma única variável qualitativa, um coeficiente linear e um termo de erro.
- **Uma equação contendo uma única variável explicativa (variável X) multiplicada por um coeficiente angular (beta), um coeficiente linear (alfa) e um termo de erro. ←**
- Um gráfico obtido através de um modelo linear.

Sejam as seguintes afirmações:

I) Os modelos de regressão inserem-se naquilo que é conhecido por GLM (Generalized Linear Models).

II) A estimação OLS é adequada para todo tipo de modelo de regressão, mesmo quando a variável dependente for qualitativa.

III) Muitas são as classes de estimações que se inserem nos modelos GLM, tais como os modelos de regressão simples e múltipla, os modelos logísticos binários e multinomiais e os modelos para dados de contagem.

Assinale a alternativa correta:

- Apenas I está correta.
- Apenas II e III estão corretas.
- Apenas I e II estão corretas.
- **Apenas I e III estão corretas. ←**

Sejam as seguintes afirmações:

I) As bandas dos intervalos de confiança de parâmetros beta em modelos regressivos aumentam quando se aumenta o nível de confiança.

II) Determinado parâmetro beta pode se tornar estatisticamente não significante ao aumentarmos o nível de confiança para fins preditivos.

III) Amostras reduzidas podem fazer com que o parâmetro alpha não se apresente estatisticamente significante em modelos regressivos.

- Apenas I está correta.
- **Todas estão corretas. ←**
- Apenas I e II estão corretas.
- Apenas I e III estão corretas.

Após a estimação de um modelo a partir da função "lm" no R, obteve-se o seguinte output:

Sejam as seguintes afirmações:

I. Os valores estimados dos coeficientes do modelo de regressão são alpha = 11,97192; beta1 (idade) = 0,09970; beta2 (horas) = -0,40134.

II. Este modelo permite a elaboração de previsões da variável cpi, desde que os valores inseridos em idade e horas estejam dentro do range, ou amplitude, dessas variáveis em nosso banco de dados.

III. Os parâmetros das variáveis preditoras idade e horas são estatisticamente significantes ao nível de confiança de 95% (nível de significância de 5%).

Assinale a alternativa correta:

Assinale a alternativa correta:

```
Call:
lm(formula = cpi ~ . - pais, data = paises)

Residuals:
   Min     1Q Median     3Q    Max
-4.123 -1.440 -0.316  1.570  4.251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.97192    5.16537   2.318  0.02487 *
idade        0.09970    0.03266   3.052  0.00373 **
horas       -0.40134    0.13467  -2.980  0.00455 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.249 on 47 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.2951
F-statistic: 11.26 on 2 and 47 DF,  p-value: 0.0001013
```

- Somente I e III estão corretas.
- Somente I está correta.
- Somente II e III estão corretas.
- **Todas as afirmações estão corretas. ←**

Em relação a modelos de regressão estimados pelo método de Mínimos Quadrados Ordinários, é correto afirmar que:

- Os testes t e F não são utilizados para se avaliar a adequação de um modelo regressivo.
- **A somatória dos termos de erro é sempre igual a zero. ←**
- O incremento amostral não favorece a significância estatística do parâmetro correspondente ao intercepto.
- O parâmetro correspondente ao intercepto (alpha) deverá sempre ser excluído do modelo quando este não se mostrar estatisticamente significante.

Muitas são as formas de replicação (multiplicação da quantidade de observações para as mesmas variáveis) de uma base de dados no R.

Em aula, vimos a função:

- **slice. ←**
- replicate.
- mice.
- multiplicate.

Sobre o parâmetro correspondente ao intercepto (alpha), é correto dizer que:

- ○ Sempre deverá ser excluído do modelo final.
- ○ Não existe parâmetro alpha em modelos regressivos.
- ○ Deverá ser excluído do modelo final quando não se mostrar estatisticamente significante.
- ● **Nunca deverá ser excluído do modelo final, já que sua não significância estatística é decorrente de problemas relacionados ao tamanho da amostra. ←**

Após a estimação de um modelo a partir da função "lm" no R, obteve-se o seguinte output:

É correto afirmar que o percentual de variância da variável Y (cpi) que é capturado pelo comportamento de variação das variáveis preditoras X (idade e horas) é igual a:

```
Call:
lm(formula = cpi ~ . - pais, data = paises)

Residuals:
   Min     1Q Median     3Q    Max
-4.123 -1.440 -0.316  1.570  4.251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.97192    5.16537   2.318  0.02487 *
idade        0.09970    0.03266   3.052  0.00373 **
horas       -0.40134    0.13467  -2.980  0.00455 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.249 on 47 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.2951
F-statistic: 11.26 on 2 and 47 DF,  p-value: 0.0001013
```

- ● **32,39%. ←**
- ○ 11,97%.
- ○ 29,51%.

Após a estimação de um modelo a partir da função "lm" no R, obteve-se o seguinte output:

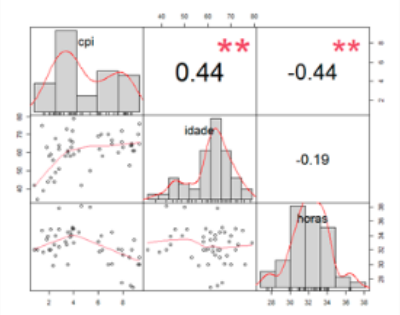
Qual o valor previsto de cpi, em média, para uma observação com idade = 40 e horas = 30?

```
Call:
lm(formula = cpi ~ . - pais, data = paises)

Residuals:
   Min     1Q Median     3Q    Max
-4.123 -1.440 -0.316  1.570  4.251

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.97192    5.16537   2.318  0.02487 *
idade        0.09970    0.03266   3.052  0.00373 **
horas       -0.40134    0.13467  -2.980  0.00455 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.249 on 47 degrees of freedom
Multiple R-squared:  0.3239,    Adjusted R-squared:  0.2951
F-statistic: 11.26 on 2 and 47 DF,  p-value: 0.0001013
```

- 2,40.
- 5,39.
- 7,47.
- **3,92. ←**

A partir de um banco de dados com três variáveis (cpi, idade e horas), elaborou-se o seguinte gráfico com a matriz de correlações entre cada par de variáveis.

Se estimarmos um modelo de regressão entre idade (Y) e horas (X), o R^2 será de:



- 19,36%.
- **3,61%. ←**
- 44,00%.
- 8,37%.

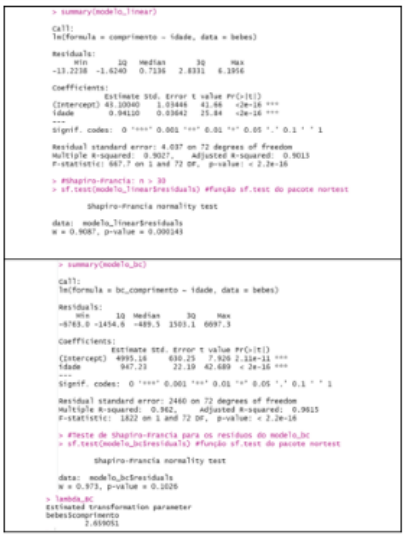
Foi elaborado um teste de Shapiro-Francia à distribuição dos resíduos de um modelo de regressão estimado por OLS. O resultado do teste é apresentado a seguir. A partir do resultado do teste, adotando-se uma significância estatística de 5%, é possível afirmar que:

```
        Shapiro-Francia normality test

data:  modelo_linear$residuals
W = 0.9087, p-value = 0.000143
```

- (●) **A distribuição dos resíduos não é aderente à distribuição normal. ←**
- ( ) A distribuição dos resíduos é aderente à distribuição t-Student.
- ( ) A distribuição dos resíduos é aderente à distribuição normal.
- ( ) O resultado é inconclusivo, já que o teste de Shapiro-Francia não serve para se avaliar aderência à normalidade.

Dois modelos de regressão foram estimados, sem e com transformação de Box-Cox, respectivamente (modelo_linear e modelo_bc). A figura a seguir apresenta os outputs dos modelos, com os respectivos testes de Shapiro-Francia dos resíduos e o lambda de Box-Cox para o segundo modelo.

Pergunta-se: qual a equação do modelo preditivo mais adequado neste caso?



```
> summary(modelo_linear)

Call:
lm(formula = comprimento ~ idade, data = bebes)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2238  -1.6240   0.7136   2.8331   6.1916

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.10040    1.03446   41.66   <2e-16 ***
idade        0.94100    0.03642   25.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.037 on 72 degrees of freedom
Multiple R-squared:  0.9027,    Adjusted R-squared:  0.9013
F-statistic: 667.7 on 1 and 72 DF,  p-value: < 2.2e-16

> #Shapiro-Francia: n > 30
> sf.test(modelo_linear$residuals) #função sf.test do pacote nortest

        Shapiro-Francia normality test

data:  modelo_linear$residuals
W = 0.9087, p-value = 0.000143
```

```
> summary(modelo_bc)

Call:
lm(formula = bc_comprimento ~ idade, data = bebes)

Residuals:
    Min      1Q  Median      3Q     Max
-6763.0 -1454.6  -489.5  1503.1  6697.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4995.16     630.25   7.926 2.11e-11 ***
idade         947.23      22.10  42.689  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2460 on 72 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9615
F-statistic:  1822 on 1 and 72 DF,  p-value: < 2.2e-16

> #Teste de Shapiro-Francia para os resíduos do modelo_bc
> sf.test(modelo_bc$residuals) #função sf.test do pacote nortest

        Shapiro-Francia normality test

data:  modelo_bc$residuals
W = 0.973, p-value = 0.1026

> lambda_bc
Estimated transformation parameter
bebes$comprimento
        2.659051
```

- ( ) $comprimento_i = 0{,}94 + 43{,}10 * idade_i$
- (●) **$(comprimento_i^{2{,}659} - 1) / 2{,}659 = 4.995{,}16 + 947{,}23 * idade_i$ ←**
- ( ) $(comprimento_i^{2{,}659} - 1) / 2{,}659 = 947{,}23 + 4.995{,}16 * idade_i$
- ( ) $comprimento_i = 43{,}10 + 0{,}94 * idade_i$

Considerando os modelos de regressão do tipo GLM, no caso de ser necessário acrescentar uma variável preditora categórica com mais de uma categoria, como devemos proceder?

- ( ) Não será possível fazer esta regressão.
- ( ) Neste caso, devemos incluir n dummies, em que n é a quantidade de categorias existentes na variável original.
- ( ) Neste caso, devemos incluir n + 1 dummies, em que n é a quantidade de categorias existentes na variável original.
- (●) **Neste caso, devemos incluir n - 1 dummies, em que n é a quantidade de categorias existentes na variável original. ←**

O que são variáveis dummy?

- São variáveis selecionadas como preditoras quantitativas.
- São variáveis dependentes métricas.
- São medidas que expressam o grau de dispersão de um conjunto de dados.
- **São variáveis categóricas que representam um atributo por meio de combinação binária (0 para a ausência ou 1 para presença). ←**

Em relação à multicolinearidade, podemos afirmar que:

- É um fenômeno decorrente da correlação significante entre variável dependente e variáveis preditoras.
- É um fenômeno decorrente da correlação significante entre variáveis preditoras e termos de erro.
- É um fenômeno que aparece em modelos regressivos apenas após a elaboração de transformações de Box-Cox.
- **É um fenômeno decorrente da correlação significante entre variáveis preditoras. ←**

Sejam as seguintes afirmações:

I. Quando a distribuição dos resíduos não se mostrar aderente à normalidade, procedimentos de normalização da variável dependente (ex.: Box-Cox) podem ser úteis para se estimar um modelo não linear.

II. A aderência da distribuição dos resíduos à normalidade, para amostras grandes, pode ser identificada por meio do teste de Shapiro-Francia.

III. Comportamentos não lineares de determinados fenômenos não podem ser identificados por meio de modelos de regressão, já que não geram coeficientes de determinação R^2.

- Apenas I está correta.
- **Apenas I e II estão corretas. ←**
- Todas as afirmações estão corretas.
- Apenas II e III estão corretas.

Sobre modelos de regressão estimados pelo critério dos mínimos quadrados ordinários:

I. Quando uma variável X apresentar uma correlação baixa e não estatisticamente significante com uma variável Y, necessariamente ela deverá estar presente no modelo final preditivo.

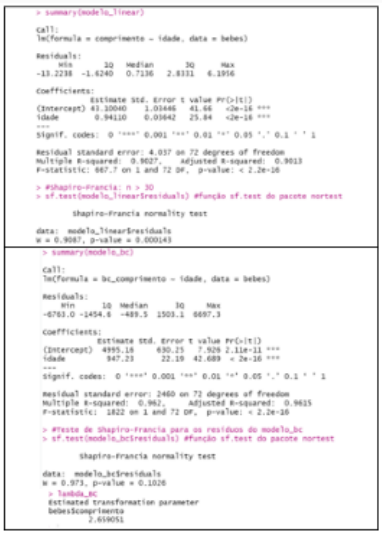
II. Todas as variáveis preditoras qualitativas de um modelo final de regressão, necessariamente, devem apresentar parâmetros estatisticamente significantes.

III. Quando uma das variáveis preditoras não estiver presente no modelo final múltiplo, necessariamente, esta variável não apresenta relação significante individual com a variável Y.

- Somente I é verdadeira.
- Todas as afirmações são verdadeiras.
- Todas as afirmações são falsas.
- **Somente II é verdadeira. ←**

Dois modelos de regressão foram estimados, sem e com transformação de Box-Cox, respectivamente (modelo_linear e modelo_bc). A figura a seguir apresenta os outputs dos modelos, com os respectivos testes de Shapiro-Francia dos resíduos e o lambda de Box-Cox para o segundo modelo.

Pergunta-se: qual o valor mais adequado da estimativa aproximada da variável comprimento para uma idade de 40 semanas?



```
> summary(modelo_linear)

Call:
lm(formula = comprimento ~ idade, data = bebes)

Residuals:
     Min       1Q   Median       3Q      Max
-13.2238  -1.6240   0.7136   2.8331   6.1916

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.10040    1.03446   41.66   <2e-16 ***
idade        0.94110    0.03642   25.84   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.037 on 72 degrees of freedom
Multiple R-squared:  0.9027,    Adjusted R-squared:  0.9013
F-statistic: 667.7 on 1 and 72 DF,  p-value: < 2.2e-16

> #Shapiro-Francia: n > 30
> sf.test(modelo_linear$residuals) #função sf.test do pacote nortest

        Shapiro-Francia normality test

data:  modelo_linear$residuals
W = 0.9087, p-value = 0.000143
```

```
> summary(modelo_bc)

Call:
lm(formula = bc_comprimento ~ idade, data = bebes)

Residuals:
    Min      1Q  Median      3Q     Max
-6763.0 -1454.6  -489.5  1503.1  6697.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4995.16     630.25   7.926 2.11e-11 ***
idade         947.23      22.19  42.689  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2460 on 72 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9615
F-statistic:  1822 on 1 and 72 DF,  p-value: < 2.2e-16

> #Teste de Shapiro-Francia para os resíduos do modelo_bc
> sf.test(modelo_bc$residuals) #função sf.test do pacote nortest

        Shapiro-Francia normality test

data:  modelo_bc$residuals
W = 0.973, p-value = 0.1026
> lambda_BC
Estimated transformation parameter
bebes$comprimento
         2.659051
```

- ○ 40 cm.
- ○ 60 cm.
- ○ 20 cm.
- ● **80 cm.** ←

QUESTÃO #10

Estimou-se um modelo de regressão por OLS com transformação de Box-Cox, e os outputs obtidos encontram-se na figura a seguir:

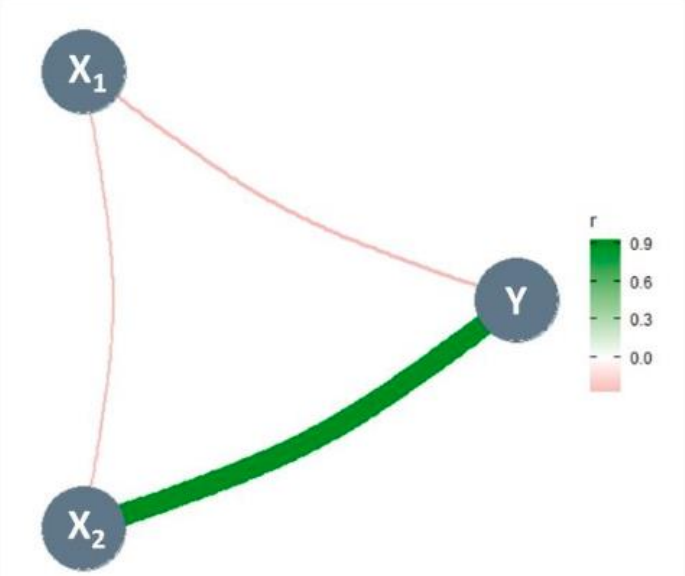
Pergunta-se: qual o valor da estimativa aproximada da variável comprimento para uma idade de 20 semanas?

```
Call:
lm(formula = bc_comprimento ~ idade, data = bebes)

Residuals:
    Min      1Q  Median      3Q     Max
-6763.0 -1454.6  -489.5  1503.1  6697.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4995.16     630.25   7.926 2.11e-11 ***
idade         947.23      22.19  42.689  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2460 on 72 degrees of freedom
Multiple R-squared:  0.962,     Adjusted R-squared:  0.9615
F-statistic:  1822 on 1 and 72 DF,  p-value: < 2.2e-16

> lambda_BC
Estimated transformation parameter
bebes$comprimento
         2.659051
```

- 42 cm.
- **64 cm. ←**
- 20 cm.
- 90 cm.

Seja o seguinte diagrama de correlações de Pearson entre cada par de variáveis (Y, X_1 e X_2), sendo Y a variável dependente e X_1 e X_2 as variáveis preditoras de determinado modelo de regressão linear múltipla:

Por meio da análise visual deste diagrama, é possível afirmar que:



- **Praticamente não há problemas de multicolinearidade no modelo estimado. ←**
- Os problemas de multicolinearidade existentes no modelo estimado são decorrentes da baixa correlação entre as variáveis X_1 e X_2.
- Os problemas de multicolinearidade existentes no modelo estimado são decorrentes da alta correlação entre as variáveis Y e X_2.
- Há indícios de existência de fortes problemas de multicolinearidade no modelo estimado.

Em relação à multicolinearidade, podemos afirmar que:

- É um fenômeno decorrente da correlação significante entre a variável dependente e os termos de erro.
- É um fenômeno que aparece em modelos regressivos apenas após a elaboração do teste de Shapiro-Francia.
- **É a consequência da existência de correlação alta entre duas ou mais variáveis explicativas. ←**
- É um fenômeno decorrente da correlação significante entre termos de erro com defasagens temporais.

Sobre a heterocedasticidade, é correto afirmar que:

- É um fenômeno que aparece em modelos regressivos apenas após a elaboração de transformações de Box-Cox.
- É um fenômeno que aparece em modelos regressivos que contenham apenas variáveis preditoras dummies.
- **É a consequência da existência de correlação significante entre os termos de erro e uma ou mais variáveis preditoras. ←**
- É um fenômeno decorrente da correlação significante entre a variável dependente e as variáveis explicativas.

O diagnóstico de heterocedasticidade em modelos regressivos pode ser realizado por meio do teste de:

- Box-Cox.
- Galton-Pearson-Gosset-Snedecor.
- **Breusch-Pagan. ←**
- Shapiro-Francia.

Podem ser fatores geradores da multicolinearidade:

I) Existência de variáveis que apresentam a mesma tendência durante alguns períodos, em decorrência da seleção de uma amostra que inclua apenas observações referentes a estes períodos.

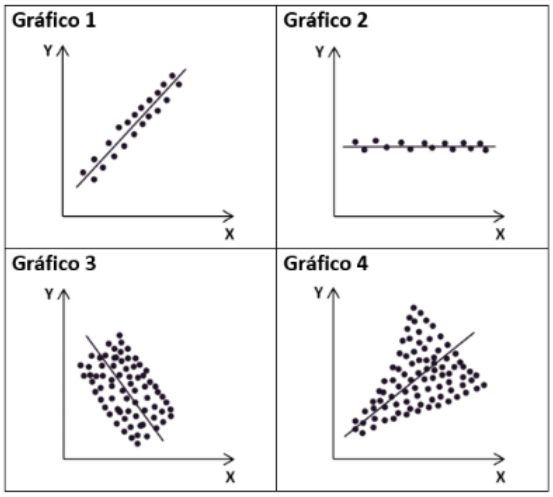
II) Utilização de amostras com reduzido número de observações.

III) Utilização de valores defasados em algumas das variáveis explicativas como "novas" explicativas.

Assinale a alternativa correta:

- ○ Apenas II e III estão corretas.
- ○ Apenas I e II estão corretas.
- ◉ **Todas as afirmações estão corretas. ←**
- ○ Apenas I e III estão corretas.

Assinale o gráfico que apresenta o maior indício de existência de heterocedasticidade quando da estimação de um modelo de regressão que considera as variáveis *Y* (dependente) e *X* (preditora).



- ◉ **Gráfico 4 ←**
- ○ Gráfico 3
- ○ Gráfico 2
- ○ Gráfico 1

Sobre a estatística VIF, é correto dizer que:

- ○ Só são aceitos modelos de regressão múltipla com valores de VIF maiores que 100 para todas as variáveis preditoras.
- ○ É algebricamente o inverso de $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável Y como dependente dos termos de erro.
- ○ É algebricamente o inverso do R^2 obtido ao se estimar um modelo de regressão com determinada variável Y como dependente das demais variáveis preditoras.
- ◉ **É algebricamente o inverso de $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável X como dependente das demais variáveis preditoras. ←**

Quezia - qsalles90@gmail.com

Questão #8

1

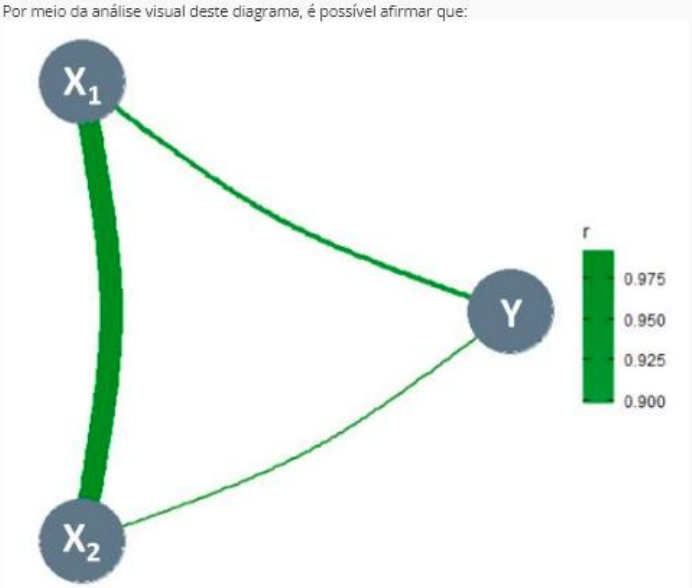
A multicolinearidade:

- ○ Não existe este fenômeno em modelos supervisionados de machine learning.
- ○ Não gera erros de predições.
- ◉ **Pode gerar interpretações erradas do modelo pela eventual distorção dos sinais dos parâmetros. ←**
- ○ Não gera sinais inesperados dos parâmetros estimados.

Quezia - qsalles90@gmail.com

Seja o seguinte diagrama de correlações de Pearson entre cada par de variáveis (Y, X_1 e X_2), sendo Y a variável dependente e X_1 e X_2 as variáveis preditoras de determinado modelo de regressão linear múltipla:

Por meio da análise visual deste diagrama, é possível afirmar que:



- ○ Os problemas de multicolinearidade existentes no modelo estimado são decorrentes da baixa correlação entre as variáveis Y e X_2.
- ◉ **Há indícios de existência de fortes problemas de multicolinearidade no modelo estimado em razão da alta correlação entre as variáveis X_1 e X_2.** ←

Sobre a estatística Tolerance, é correto dizer que:

- ○ É algebricamente o inverso de $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável X como dependente das demais variáveis preditoras.
- ○ É algebricamente o inverso de $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável Y como dependente das demais variáveis preditoras.
- ○ É algebricamente igual a $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável Y como dependente das demais variáveis preditoras.
- ◉ **É algebricamente igual a $(1 - R^2)$, em que o R^2 é obtido ao se estimar um modelo de regressão com determinada variável X como dependente das demais variáveis preditoras.** ←

ANALYTICS E MODELOS SUPERVISIONADOS DE MACHINE LEARNING: ESTIMAÇÃO POR MÁXIMA VEROSSIMILHANÇA

Modelos Logísticos Binários e Multinominais

A técnica de regressão logística binária possui muitas aplicações, como:

I) determinar a probabilidade de infarto do miocárdio de determinado paciente com base em resultados de seus exames e em seus hábitos de vida.

II) determinar a quantidade de acidentes em uma estrada por mês.

III) determinar a probabilidade de ocorrência de sinistro para determinado cliente de uma seguradora.

IV) diferenciar os clientes adimplentes dos inadimplentes em relação a empréstimos bancários.

Assinale a alternativa correta:

- ◉ **As afirmações I, III e IV estão corretas.** ←
- ○ Todas as afirmações estão corretas.
- ○ Apenas as afirmações II e IV estão corretas.
- ○ Nenhuma afirmação está correta.

Em matrizes de confusão, o que significa a especificidade?

- (●) **Taxa de acerto do modelo para as observações classificadas como "não evento". ←**
- ( ) Taxa de acerto do modelo para as observações classificadas como "evento".
- ( ) Taxa global de acerto do modelo.
- ( ) Taxa de erro do modelo.

D

Escolha a alternativa **incorreta** sobre a função logística:

- (●) **A variável dependente se apresenta na forma quantitativa e com média maior que zero. ←**
- ( ) Para se analisar a qualidade do ajuste do modelo para determinado cutoff, podemos utilizar uma matriz de confusão.
- ( ) Esta função é definida para que se estabeleça a probabilidade de ocorrência de determinado evento e a importância das variáveis explicativas para esta ocorrência.
- ( ) A estimação dos parâmetros desta função é um processo iterativo para maximizar o acerto da probabilidade de ocorrência de um evento à sua real ocorrência, por meio do Método de Máxima Verossimilhança.

Assumindo que um logito possua o valor igual a Z = 0. Em um modelo de regressão logística binária, qual será a probabilidade de ocorrência do evento em estudo?

- (●) **0,5 ←**
- ( ) 0,7311
- ( ) 0,9933
- ( ) 1

Em matrizes de confusão, o que significa a sensitividade?

- ( ) Taxa de acerto do modelo para as observações classificadas como "não evento".
- ( ) Taxa global de acerto do modelo.
- ( ) Taxa de erro do modelo.
- (●) **Taxa de acerto do modelo para as observações classificadas como "evento". ←**

Assumindo que um logito possua o valor igual a Z = 1. Em um modelo de regressão logística binária, qual será a probabilidade de ocorrência do evento em estudo?

- ( ) 0,5
- ( ) 0,2689
- ( ) 1
- (●) **0,7311 ←**

O logaritmo natural da chance de ocorrência de uma resposta do tipo "sim" é definido como:

- (●) **Logito. ←**
- ( ) Probabilidade.
- ( ) Probito.
- ( ) Acurácia.

Os parâmetros de modelos de regressão logística são estimados por:

- ( ) Máximos Quadrados Ordinários.
- ( ) Mínima Verossimilhança.
- ( ) Mínimos Quadrados Ordinários.
- (●) **Máxima Verossimilhança. ←**

A respeito de modelos de regressão logística, julgue as seguintes afirmações:

I) A variável dependente de um modelo de regressão logística binária apresenta distribuição Bernoulli.

II) Uma característica da variável dependente em uma regressão logística binária é o fato de ser qualitativa com mais de duas categorias.

III) A variável dependente de um modelo de regressão logística multinomial apresenta distribuição binomial.

Assinale a alternativa correta:

- (●) **Somente as afirmações I e III estão corretas. ←**
- ( ) Somente a afirmação III está correta.
- ( ) Somente a afirmação I está correta.
- ( ) Somente a afirmação II está correta.

Para um modelo para avaliar a probabilidade de que atletas concluam determinada prova de corrida de rua (evento = conclusão da prova), considere o seguinte logito obtido na estimação do modelo de regressão logística binária:

$Z = \alpha + \beta_1$.sexo(feminino = 1) + β_2.idade

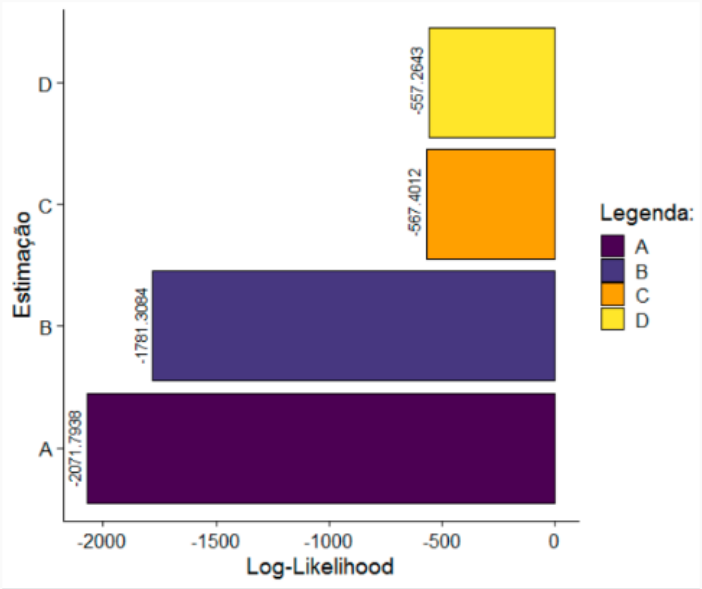
Sendo $\alpha = 0{,}34$; $\beta_1 = 0{,}14$; $\beta_2 = -0{,}047$

Uma pessoa do sexo feminino e com 37 anos de idade apresenta qual probabilidade de conclusão da prova?

- (●) **0,221 ←**
- ( ) 0,333
- ( ) 0,064
- ( ) 0,779

Foram estimados quatro modelos de regressão logística multinomial (A, B, C e D), com quatro diferentes grupos de variáveis preditoras. Os valores de log-likelihood de cada modelo encontram-se no gráfico a seguir:

Qual o modelo com melhor qualidade do ajuste para fins preditivos?



- ( ) Modelo A.
- ( ) Modelo B.
- (●) **Modelo D. ←**

Dada a seguinte matriz de confusão (com lógica igual à apresentada em aula) obtida após a estimação de um modelo de regressão logística binária e com cutoff = 0,5, escolha a alternativa que apresenta a eficiência global do modelo (indicador de acurácia).

```
Confusion Matrix

       TRUE FALSE
TRUE   1470   210
FALSE   210  1110
```

- (●) 86,00% ←
- ( ) 12,40%
- ( ) 14,00%
- ( ) 70,20%

Dada a seguinte matriz de confusão (com lógica igual à apresentada em aula) obtida após a estimação de um modelo de regressão logística binária e com cutoff = 0,5, escolha a alternativa que apresenta a sensitividade do modelo.

```
Confusion Matrix

       TRUE FALSE
TRUE   1470   210
FALSE   210  1110
```

- (●) 87,50% ←
- ( ) 12,50%
- ( ) 30,20%
- ( ) 77,40%

Para o estudo das probabilidades de não se chegar atrasado às aulas (categoria de referência), de se chegar atrasado à primeira aula (primeira categoria alternativa) ou de se chegar atrasado à segunda aula (segunda categoria alternativa) por parte de alunos de determinada escola, em função da distância (em km) de onde partem até a escola (variável preditora *dist*) e da quantidade de semáforos por que passam ao longo do trajeto (variável preditora *sem*), estimou-se um modelo de regressão logística multinomial. Os outputs com os parâmetros (todos estatisticamente significantes ao nível de confiança de 95%) encontram-se a seguir:

Pergunta-se: qual a probabilidade de que determinado aluno não chegue atrasado às aulas, sabendo-se que durante o trajeto de 22 km há 12 semáforos?

```
Coefficients:
                                (Intercept)      dist      sem
chegou atrasado à primeira aula   -33.06853 0.5574586 1.666924
chegou atrasado à segunda aula    -62.21623 1.0766952 2.891689
```

- (●) 68,00% ←
- ( ) 30,52%
- ( ) 1,48%
- ( ) 41,77%

Dada a seguinte matriz de confusão (com lógica igual à apresentada em aula) obtida após a estimação de um modelo de regressão logística binária e com cutoff = 0,5, escolha a alternativa que apresenta a especificidade do modelo.

```
Confusion Matrix

       TRUE FALSE
TRUE   1470   210
FALSE   210  1110
```

- ( ) 40,20%
- (●) 84,09% ←
- ( ) 70,42%
- ( ) 15,91%

Para se avaliar a probabilidade de se chegar atrasado à aula (evento = atrasado; não evento = não atrasado) por parte de alunos de determinada escola, em função da distância (em km) de onde partem até a escola (variável preditora *dist*) e da quantidade de semáforos por que passam ao longo do trajeto (variável preditora *sem*), estimou-se um modelo de regressão logística binária. Os outputs encontram-se a seguir:

Pergunta-se: qual a probabilidade de que determinado aluno chegue atrasado, sabendo-se que durante o trajeto de 6 km há 10 semáforos?

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.16654    8.44197  -3.100  0.00194 **
dist          0.19038    0.07637   2.493  0.01267 *
sem           2.36288    0.79512   2.972  0.00296 **
```

- ○ 80,15%
- ○ 50,77%
- ○ 4,84%
- ◉ **19,85% ←**

Sobre um modelo de regressão logística multinomial, é **correto** afirmar que:

- ○ Os parâmetros são estimados pelo método de mínimos quadrados ordinários.
- ○ Deve-se definir um ponto de corte (cutoff) para classificação das observações.
- ◉ **A variável dependente segue uma distribuição binomial. ←**
- ○ Não é possível se estabelecer um critério para classificação das observações após a estimação do modelo, já que não pode ser definida uma matriz de confusão.

Um indicador bastante útil para se avaliar a eficiência de modelos de regressão logística binária, independentemente do cutoff, é:

- ◉ **Área abaixo da curva ROC. ←**
- ○ Sensitividade.
- ○ Acurácia.
- ○ Especificidade.

Sobre a regressão logística multinomial, é correto afirmar que:

I) A variável dependente se apresenta na forma qualitativa com mais de duas categorias.

II) Para uma variável qualitativa com três categorias, serão definidos dois logitos.

III) Os parâmetros são estimados por máxima verossimilhança.

Assinale a alternativa correta:

- ○ Somente as afirmações I e II estão corretas.
- ○ Nenhuma afirmação está correta.
- ○ Somente as afirmações II e III estão corretas.
- ◉ **Todas as afirmações estão corretas. ←**

Para se avaliar a probabilidade de se chegar atrasado à aula (evento = atrasado; não evento = não atrasado) por parte de alunos de determinada escola, em função da distância (em km) de onde partem até a escola (variável preditora *dist*) e da quantidade de semáforos por que passam ao longo do trajeto (variável preditora *sem*), estimou-se um modelo de regressão logística binária. Os outputs encontram-se a seguir:

Pergunta-se: qual a probabilidade de que determinado aluno chegue atrasado, sabendo-se que durante o trajeto de 10 km há 10 semáforos?

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -26.16654    8.44197  -3.100  0.00194 **
dist          0.19038    0.07637   2.493  0.01267 *
sem           2.36288    0.79512   2.972  0.00296 **
```

- (●) **34,66% ←**
- ( ) 4,84%
- ( ) 65,34%
- ( ) 50,77%

Modelos de Regressão para Dados de Contagem (Poisson e Binomial Negativo);

Modelos Inflacionados de Zeros;

Os parâmetros de um modelo de regressão do tipo Poisson são estimados pelo método de:

- ( ) Máximos Quadrados Ordinários.
- ( ) Mínima Verossimilhança.
- ( ) Mínimos Quadrados Ordinários.
- (●) **Máxima Verossimilhança. ←**

A função e o respectivo pacote para a elaboração direta do teste para verificação de existência de superdispersão nos dados da variável dependente para a estimação de modelos de contagem, no R, são:

- ( ) função **hiperdisp** do pacote **megadisp**.
- ( ) função **superdisp** do pacote **hiperdisp**.
- ( ) função **megadisp** do pacote **ultradisp**.
- (●) **função overdisp do pacote overdisp. ←**

Dado o seguinte resultado de um teste para verificação de existência de superdispersão na variável dependente de determinado modelo.

A partir deste output, é correto dizer que:

```
        Overdispersion Test - Cameron & Trivedi (1990)

data:  corruption
Lambda t test score: = 2.7538, p-value = 0.006253
alternative hypothesis: overdispersion if lambda p-value is less than or equal to the s
tipulated significance level
```

- ( ) Verifica-se a existência de equidispersão nos dados da variável dependente.
- ( ) A partir deste output não se pode concluir nada a respeito de uma eventual superdispersão nos dados da variável dependente, já que (lambda – theta = delta).
- ( ) Verifica-se a existência de dispersão reversa nos dados das variáveis preditoras.
- (●) **Verifica-se a existência de superdispersão nos dados da variável dependente. ←**

Sobre os modelos de regressão para dados de contagem, podemos avaliar a qualidade do ajuste do modelo por meio do seguinte indicador:

- ( ) Lambda (λ) de Box-Cox.
- ( ) Área abaixo da curva ROC.
- (●) **Valor de Log-Likelihood. ←**
- ( ) p-value da estatística t de Student.

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão Poisson, considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não *enforcement* legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy *post*: *yes* = há obrigatoriedade do pagamento; no = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se a seguir:

Pergunta-se: qual a quantidade esperada de violações de trânsito para um país cujo corpo diplomático seja composto por 28 membros, considerando inexistência de *enforcement* legal (*post* = "no", ou seja, dummy *postyes* = 0) e índice de corrupção igual a 1?

```
Call:
glm(formula = violations ~ staff + post + corruption, family = "poisson",
    data = corruption)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.1425  -2.8326  -0.6008  -0.3940  24.6141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.212739   0.031107   71.13   <2e-16 ***
staff        0.021870   0.001228   17.81   <2e-16 ***
postyes     -4.296762   0.197446  -21.76   <2e-16 ***
corruption   0.341765   0.027495   12.43   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6397.7  on 297  degrees of freedom
Residual deviance: 3644.0  on 294  degrees of freedom
AIC: 4151.6
```

- ○ 47,54
- ○ 29,32
- ○ 17,93
- ◉ **23,73 ←**

São exemplos de variáveis com dados de contagem:

I) Quantidade de vezes que pacientes idosos vão ao médico por ano.

II) Quantidade de ofertas públicas de ações que são realizadas em uma amostra de países desenvolvidos e emergentes por ano.

III) Quantidade de apartamentos à venda por bairro.

IV) Faixa de renda (definida em labels) de uma amostra de consumidores.

Assinale a alternativa correta:

- ◉ **Somente as afirmações I, II e III estão corretas. ←**
- ○ Somente as afirmações II e IV estão corretas.
- ○ Todas as afirmações estão corretas.
- ○ Nenhuma afirmação está correta.

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão Poisson, considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não *enforcement* legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy *post*: *yes* = há obrigatoriedade do pagamento; no = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se a seguir:

Pergunta-se: qual a equação do modelo de regressão Poisson que deverá ser utilizada para fins preditivos? O subscrito i refere-se à linha (*row*) do dataset.

```
Call:
glm(formula = violations ~ staff + post + corruption, family = "poisson",
    data = corruption)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-9.1425  -2.8326  -0.6008  -0.3940  24.6141

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.212739   0.031107   71.13   <2e-16 ***
staff        0.021870   0.001228   17.81   <2e-16 ***
postyes     -4.296762   0.197446  -21.76   <2e-16 ***
corruption   0.341765   0.027495   12.43   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 6397.7  on 297  degrees of freedom
Residual deviance: 3644.0  on 294  degrees of freedom
AIC: 4151.6
```

- (●) **$\ln(violations_i) = 2{,}212739 + 0{,}021870 \cdot (staff_i) - 4{,}296762 \cdot (post = \text{“yes”}_i) + 0{,}341765 \cdot (corruption_i)$ ←**
- ( ) $violations_i = 0{,}021870 \cdot (staff_i) - 4{,}296762 \cdot (post = \text{“yes”}_i) + 0{,}341765 \cdot (corruption_i)$
- ( ) $violations_i = 2{,}212739 + 0{,}021870 \cdot (staff_i) - 4{,}296762 \cdot (post = \text{“yes”}_i) + 0{,}341765 \cdot (corruption_i)$
- ( ) $e^{(violations_i)} = 2{,}212739 + 0{,}021870 \cdot (staff_i) - 4{,}296762 \cdot (post = \text{“yes”}_i) + 0{,}341765 \cdot (corruption_i)$

Uma variável com dados de contagem apresenta as seguintes características:

- (●) **É quantitativa, apresenta dados discretos e não negativos, e é definida uma exposição. ←**
- ( ) É quantitativa, apresenta dados contínuos e negativos, e é definida uma exposição.
- ( ) É qualitativa, apresenta dados contínuos e negativos, e é definida uma exposição.
- ( ) É quantitativa, apresenta dados contínuos e não negativos, e não se consegue definir uma exposição.

O principal teste para verificação de existência de superdispersão nos dados da variável dependente é o:

- (●) **Teste de Cameron e Trivedi. ←**
- ( ) Teste de Lambert.
- ( ) Teste de Vuong.
- ( ) Teste de Shapiro-Francia.

Qual a característica da superdispersão em determinada variável dependente com dados de contagem?

- ( ) Média estatisticamente superior à variância.
- (●) **Variância estatisticamente superior à média. ←**
- ( ) Variância estatisticamente superior ao desvio-padrão.
- ( ) Variância estatisticamente superior à entropia conjunta das variáveis preditoras.

A definição sobre a existência ou não de uma quantidade excessiva de zeros na variável dependente *Y* de um modelo de regressão para dados de contagem é verificada por meio de um teste específico, conhecido por:

- ( ) Teste de Cameron e Trivedi.
- ( ) Teste de Erlang.
- ( ) Teste de Lambert.
- (●) **Teste de Vuong. ←**

Em relação especificamente aos modelos de regressão Poisson inflacionados de zeros, podemos afirmar que:

I) A probabilidade p de ocorrência de nenhuma contagem para dada observação i (i = 1, 2, ..., n, em que n é o tamanho da amostra), ou seja, $p(Y_i = 0)$, é calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem.

II) Deve-se definir a probabilidade p_{logit} de não ocorrer nenhuma contagem devido exclusivamente ao componente dicotômico, bem como a probabilidade p de ocorrência de determinada contagem m (m = 1, 2, ...), ou seja, $p(Y_i = m)$, que segue a própria expressão da probabilidade da distribuição Poisson, multiplicada por $(1 - p_{logit})$.

III) São úteis apenas para variáveis preditoras qualitativas.

Assinale a alternativa correta:

- Apenas as afirmações II e III estão corretas.
- Apenas as afirmações I e III estão corretas.
- Todas as afirmações estão corretas.
- **Apenas as afirmações I e II estão corretas. ←**

Em relação aos modelos de regressão do tipo binomial negativo com inflação de zeros, podemos afirmar que:

I) A probabilidade p de ocorrência de nenhuma contagem para dada observação i, ou seja, $p(Y_i = 0)$, é calculada levando-se em consideração a soma de um componente dicotômico com um componente de contagem.

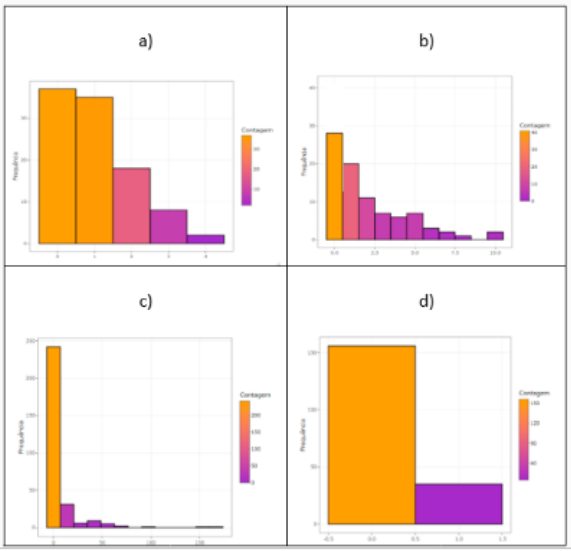
II) A probabilidade p de ocorrência de determinada contagem m (m = 1, 2, ...), ou seja, $p(Y_i = m)$, segue a expressão da probabilidade da distribuição Poisson-Gama.

III) Apresentam parâmetros estimados por máxima verossimilhança.

Assinale a alternativa correta:

- Apenas as afirmações I e II estão corretas.
- Apenas as afirmações II e III estão corretas.
- Apenas as afirmações I e III estão corretas.
- **Todas as afirmações estão corretas. ←**

Assinale a alternativa que mostra um histograma de determinada variável com indícios de existência de superdispersão e inflação de zeros nos dados.



- Histograma **d**.
- Histograma **b**.
- Histograma **a**.
- **Histograma c. ←**

São modelos supervisionados de machine learning para dados de contagem:

- (●) **Regressão Poisson, Regressão Binomial Negativa, Regressão Poisson com Inflação de Zeros e Regressão Binomial Negativa com Inflação de Zeros. ←**
- ( ) Análise de Clusters e Análise Fatorial por Componentes Principais.
- ( ) Regressão Gaussiana e Regressão estimada por OLS.
- ( ) Regressão Logística Binária e Regressão Logística Multinomial.

Escolha a alternativa que apresenta um código do R para se estimar um modelo de regressão para dados de contagem do tipo binomial negativo:

- ( ) modelo <- nbin(formula = y ~ x1 + x2 + x3, data = dataset)
- (●) **modelo <- glm.nb(formula = y ~ x1 + x2 + x3, data = dataset) ←**
- ( ) modelo <- glm(formula = y ~ x1 + x2 + x3, data = dataset, family = "negbin")
- ( ) modelo <- glm(formula = y ~ x1 + x2 + x3, data = dataset, family = "poissongamma")

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão binomial negativo considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não *enforcement* legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy *post*: "*yes*" = há obrigatoriedade do pagamento; "*no*" = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se a seguir:

Pergunta-se: qual a equação do modelo de regressão binomial negativo que deverá ser utilizada para fins preditivos? O subscrito *i* refere-se à linha (*row*) do dataset.

```
Call:
glm.nb(formula = violations ~ staff + post + corruption, data = corruption,
    init.theta = 0.4770222578, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9682  -0.6754  -0.5194  -0.2938   4.2880

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.946898   0.161575  12.050  < 2e-16 ***
staff        0.040018   0.008899   4.497 6.90e-06 ***
postyes     -4.274635   0.265986 -16.071  < 2e-16 ***
corruption   0.452655   0.114607   3.950 7.83e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.477) family taken to be 1)

    Null deviance: 570.78  on 297  degrees of freedom
Residual deviance: 239.13  on 294  degrees of freedom
AIC: 1144.8

Number of Fisher Scoring iterations: 1


              Theta:  0.4770
          Std. Err.:  0.0554

 2 x log-likelihood:  -1134.8020
```


- ( ) $violations_i = 1{,}946868 + 0{,}040018 \,.\, (staff_i) - 4{,}274635 \,.\, (\text{post} = \text{“}yes\text{”}_i) + 0{,}452655 \,.\, (corruption_i)$
- ( ) $e^{(violations_i)} = 1{,}946868 + 0{,}040018 \,.\, (staff_i) - 4{,}274635 \,.\, (\text{post} = \text{“}yes\text{”}_i) + 0{,}452655 \,.\, (corruption_i)$
- ( ) $violations_i = 0{,}040018 \,.\, (staff_i) - 4{,}274635 \,.\, (\text{post} = \text{“}yes\text{”}_i) + 0{,}452655 \,.\, (corruption_i)$
- (●) **$\ln(violations_i) = 1{,}946868 + 0{,}040018 \,.\, (staff_i) - 4{,}274635 \,.\, (\text{post} = \text{“}yes\text{”}_i) + 0{,}452655 \,.\, (corruption_i)$ ←**

A variável dependente quantitativa a ser considerada em determinado modelo de regressão para dados de contagem apresenta as seguintes estatísticas descritivas.

A priori, dentre as alternativas propostas a seguir, qual seria o modelo mais adequado a ser estimado?

Média	Variância
6.496644	331.6178

- ( ) Logístico multinomial.
- ( ) Poisson.
- ( ) Logístico binário.
- (●) **Binomial negativo. ←**

Com o intuito de se estudar e projetar a quantidade de violações de trânsito (variável dependente *violations*) na cidade de Nova York por parte de membros do corpo diplomático de países pertencentes às Nações Unidas, foi estimado um modelo de regressão binomial negativo, considerando, como variáveis preditoras, a quantidade de membros no corpo diplomático em cada país (variável *staff*), o índice de corrupção de cada país (variável *corruption*) e o fato de haver ou não *enforcement* legal quanto à obrigatoriedade de se pagar a multa em caso de violação (variável dummy *post*: "*yes*" = há obrigatoriedade do pagamento; "*no*" = não há obrigatoriedade do pagamento). Os outputs do referido modelo, obtidos no R, encontram-se a seguir:

Pergunta-se: qual a quantidade esperada de violações de trânsito para um país cujo corpo diplomático seja composto por 28 membros, considerand inexistência de *enforcement* legal (*post* = "*no*", ou seja, dummy *postyes* = 0) e índice de corrupção igual a 1?

```
Call:
glm.nb(formula = violations ~ staff + post + corruption, data = corruption,
    init.theta = 0.4770222578, link = log)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9682  -0.6754  -0.5194  -0.2938   4.2880

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.946898   0.161575  12.050  < 2e-16 ***
staff        0.040018   0.008899   4.497 6.90e-06 ***
postyes     -4.274635   0.265986 -16.071  < 2e-16 ***
corruption   0.452655   0.114607   3.950 7.83e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(0.477) family taken to be 1)

    Null deviance: 570.78  on 297  degrees of freedom
Residual deviance: 239.13  on 294  degrees of freedom
AIC: 1144.8

Number of Fisher Scoring iterations: 1


              Theta:  0.4770
          Std. Err.:  0.0554

 2 x log-likelihood:  -1134.8020
```

- ○ 11,93
- ○ 59,32
- ● 33,78 ←
- ○ 77,52

Modelos Multinível

Dado o seguinte modelo multinível, conforme discutido em aula:

em que Y é a variável dependente, X é uma variável preditora de nível 1 (indivíduos i), W é uma variável preditora contextual (grupos contextuais j), γ são os parâmetros a serem estimados, ν_0 e ν_1 representam, respectivamente, os termos aleatórios de intercepto e inclinação de nível 2, e ε representa os termos de erro idiossincráticos (nível 1).

Assinale a alternativa correta:

$$Y_{ij} = \underbrace{\gamma_{00} + \gamma_{10} \cdot X_{ij} + \gamma_{01} \cdot W_j + \gamma_{11} \cdot W_j \cdot X_{ij}}_{\textbf{Efeitos fixos}} + \underbrace{\nu_{0j} + \nu_{1j} \cdot X_{ij} + \varepsilon_{ij}}_{\textbf{Efeitos Aleatórios}}$$

- ○ Nenhuma das anteriores.
- ○ Se as variâncias dos termos aleatórios ν_0 e ν_1 para os grupos j forem estatisticamente iguais a zero, procedimentos tradicionais de estimação dos parâmetros do modelo, como mínimos quadrados ordinários, não serão adequados.
- ● **Se as variâncias dos termos aleatórios ν_0 e ν_1 para os grupos j forem estatisticamente diferentes de zero, procedimentos tradicionais de estimação dos parâmetros do modelo, como mínimos quadrados ordinários, não serão adequados. ←**
- ○ Se as variâncias dos termos aleatórios ν_0 e ν_1 para os grupos j forem estatisticamente diferentes a zero, procedimentos tradicionais de estimação dos parâmetros do modelo, como mínimos quadrados ordinários, serão adequados.

São considerados desafios atuais em modelagem multinível:

I) Interações profundas entre variáveis e capacidade computacional de processamento.

II) Métodos de estimação dos parâmetros.

III) Clusterização da amostra.

Assinale a alternativa correta:

- ( ) Apenas as afirmações II e III estão corretas.
- ( ) Apenas as afirmações I e II estão corretas.
- (•) **Todas as afirmações estão corretas. ←**
- ( ) Apenas as afirmações I e III estão corretas.

São modelos supervisionados de machine learning:

- (•) **Modelos Lineares Generalizados e Modelos Multinível. ←**
- ( ) Análise de Conglomerados e Modelos Multinível.
- ( ) Análise de Correspondência e Modelos Lineares Generalizados.
- ( ) Modelos Lineares Generalizados e Análise Fatorial por Componentes Principais.

São nomenclaturas para a modelagem multinível:

- ( ) Modelos para Dados de Contagem, Modelos Poisson, Modelos Binomiais Negativos, Modelos Inflacionados de Zeros.
- ( ) Modelos Não Supervisionados de Machine Learning.
- (•) **GLMM, Modelagem Hierárquica, GLLAMM, Random Coefficients Modeling, Mixed Modeling. ←**
- ( ) GLM, Modelos de Regressão Simples e Múltipla, Modelos Logísticos e Modelos Poisson.

Em relação aos modelos multinível, é correto dizer que:

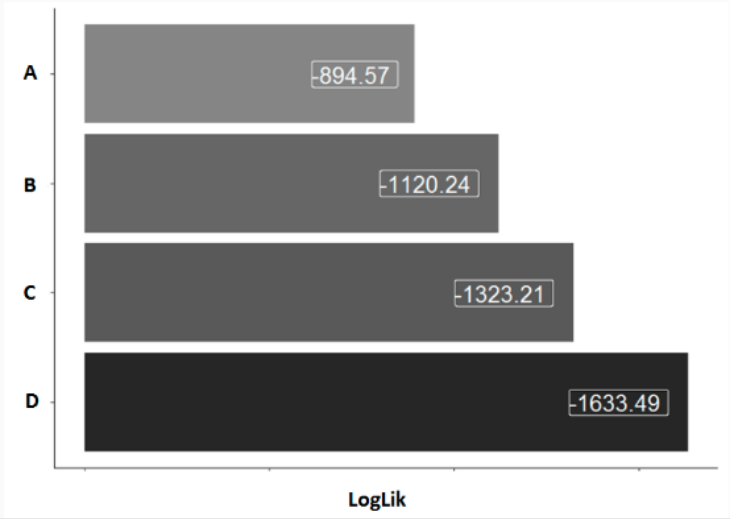
- ( ) São estimados pelo método de Minimum Variance Quadratic Unbiased Estimation (MIVQUE).
- ( ) São estimados pelo método de Máximos Quadrados Não Ordinários.
- (•) **São estimados pelo método de Máxima Verossimilhança no Conceito Restrito (REML). ←**
- ( ) São estimados pelo método de Mínimos Quadrados Ordinários.

Assinale a alternativa que **NÃO** traz um exemplo de observações aninhadas em contextos (nesta ordem), ou seja, que **NÃO** tornaria possível a definição direta de níveis hierárquicos para a estimação de uma modelagem multinível.

- ( ) Municípios e Países.
- ( ) Empresas e Setores.
- ( ) Alunos e Escolas.
- (•) **Seres Humanos e Mariposas. ←**

Quatro diferentes modelos (A, B, C e D) foram estimados, a partir de diferentes critérios e considerando, ou não, a perspectiva multinível. O gráfico a seguir apresenta os valores de LogLik para cada um deles, sabendo-se que todos os LogLiks são estatisticamente diferentes entre si (informação obtida a partir de diferentes *likelihood ratio tests*).

Para fins preditivos, qual deverá ser o modelo escolhido pelo critério do LogLik?



- (●) A. ←
- ( ) D.
- ( ) B.
- ( ) C.

Dadas as seguintes afirmações:

I) Em modelagem multinível, o procedimento de inserção de *dummies* de grupo não torna possível a identificação dos efeitos contextuais, visto que não se separam os efeitos observáveis dos não observáveis sobre a variável dependente.

II) Em modelagem multinível, o procedimento de inserção de *dummies* de grupo é a melhor decisão a ser tomada.

III) Apenas a inserção de *dummies* de grupo, em modelagem multinível, faz com que sejam capturados os efeitos contextuais sobre a variável dependente.

Assinale a alternativa correta:

- (●) **Apenas a afirmação I está correta. ←**
- ( ) Todas as afirmações estão corretas.
- ( ) Apenas as afirmações I e III estão corretas.
- ( ) Apenas as afirmações II e III estão corretas.

Em relação a modelagens multinível, é correto afirmar que:

I) Deve haver aninhamento entre observações e contextos a que pertencem estas observações, fato que caracteriza o nível hierárquico.

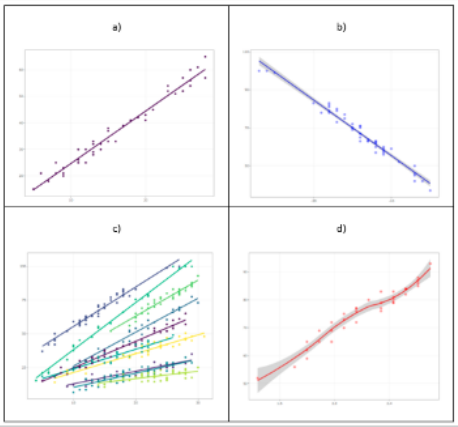
II) Quando não houver grupos diretamente observáveis, técnicas não supervisionadas de machine learning, como análise de clusters, poderão ser úteis para se definirem contextos (grupos) latentes, ou não observáveis.

III) Apresentam componentes de efeitos fixos e de efeitos aleatórios em sua expressão algébrica.

Assinale a alternativa correta:

- ( ) Apenas as afirmações I e II estão corretas.
- ( ) Apenas as afirmações I e III estão corretas.
- ( ) Apenas as afirmações II e III estão corretas.
- (●) **Todas as afirmações estão corretas. ←**

Assinale a alternativa que mostra, visualmente, o conceito de modelagem multinível.



- d.
- b.
- a.
- c. ←

Sejam as seguintes afirmações sobre os modelos multinível.

I) O modelo final pode ser obtido após a aplicação de um procedimento *Stepwise*.

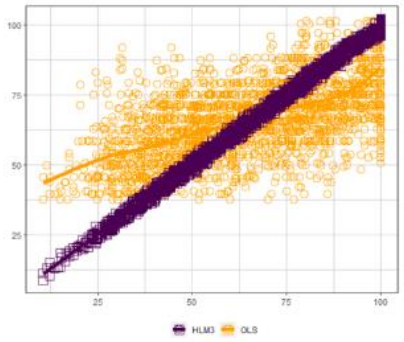
II) O modelo final pode ser obtido após a aplicação de uma *Step-Up Strategy*, já que procedimentos tradicionais *Stepwise* não são adequados em modelos com componentes fixos e aleatórios.

III) Os parâmetros podem ser estimados pelo método REML (*restricted estimation of maximum likelihood*).

Assinale a alternativa correta:

- Todas as afirmações estão corretas.
- **Apenas as afirmações II e III estão corretas. ←**
- Apenas as afirmações I e III estão corretas.
- Apenas a afirmação I está correta.

A partir de um dataset com estrutura hierárquica nos dados, foram estimados dois modelos (HLM3 e OLS). A figura a seguir apresenta a configuração da relação entre os *fitted values* da variável dependente obtidos em cada caso (eixo das ordenadas) e os valores reais da variável dependente (eixo das abcissas).



A partir da análise visual do gráfico, é possível afirmar que:

- nenhuma das alternativas anteriores.
- os dois modelos apresentam, provavelmente, os mesmos valores de LL.
- **o modelo HLM3, provavelmente, apresenta maior valor de LL. ←**
- o modelo OLS, provavelmente, apresenta maior valor de LL.

A partir de determinado dataset com estrutura hierárquica, e com base na nomenclatura utilizada em aula, foi estimado inicialmente um modelo multinível nulo com dois níveis (estudantes aninhados em escolas), sendo gerado o respectivo objeto **modelo_nulo_hlm2**. Os outputs obtidos encontram-se a seguir:

```
> summary(modelo_nulo_hlm2)
Linear mixed-effects model fit by REML
  Data: estudante_escola
       AIC      BIC    logLik
  2838.015 2849.648 -1416.007

Random effects:
 Formula: ~1 | escola
        (Intercept) Residual
StdDev:    20.34946 11.95508

Fixed effects:  desempenho ~ 1
               Value Std.Error  DF  t-value p-value
(Intercept) 42.38711  6.468659 348 6.552689       0

Standardized Within-Group Residuals:
         Min           Q1          Med           Q3          Max
-3.517143601 -0.426183104  0.004840772  0.450713210  3.567711200

Number of Observations: 358
Number of Groups: 10
> #Verificando a funcionalidade da função 'stderr_nlme' desenvolvida
> stderr_nlme(modelo_nulo_hlm2)
  RE.Components Variance.Estimatives  Std.Err.         z p.value
1      Var(v0j)             414.1005 197.09749  2.100993   0.036
2        Var(e)             142.9239  10.83481 13.191175   0.000
```

Pergunta-se: qual o percentual da variação da variável dependente (desempenho) que é devido às diferenças existentes entre escolas (efeito contextual escola), ou seja, a ICC (correlação intraclasse) do nível escola?

- ○ 34,74%
- ○ 99,99%
- ○ 12,47%
- ◉ 74,34% ←

Seguindo a nomenclatura adotada em aula, apresenta-se a seguinte especificação de um modelo multinível.

$$Y_{ij} = \beta_{0j} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \nu_{0j}$$

A partir da especificação apresentada, é possível definir o modelo proposto como sendo um:

- ○ Modelo nulo HLM3 com medidas repetidas.
- ◉ **Modelo nulo HLM2.** ←
- ○ Modelos HLM2 com inclinações aleatórias.
- ○ Modelos HLM3 com interceptos e inclinações aleatórios.

Seguindo a nomenclatura adotada em aula, apresenta-se a seguinte especificação de um modelo multinível.

$$Y_{tjk} = \beta_{0jk} + \beta_{1jk}.mês_{jk} + \varepsilon_{tjk}$$

$$\beta_{0jk} = \gamma_{00k} + \nu_{0jk}$$

$$\beta_{1jk} = \gamma_{10k} + \nu_{1jk}$$

$$\gamma_{00k} = \delta_{000} + \tau_{00k}$$

$$\gamma_{10k} = \delta_{100} + \tau_{10k}$$

que resulta na seguinte expressão:

$$Y_{tjk} = \delta_{000} + \delta_{100}.mês_{jk} + \tau_{00k} + \tau_{10k}.mês_{jk} + \nu_{0jk} + \nu_{1jk}.mês_{jk} + \varepsilon_{tjk}$$

A partir da especificação apresentada, é possível definir o modelo proposto como sendo um:

- ○ Modelo HLM2 com interceptos aleatórios.
- ○ Modelo nulo HLM2.
- ○ Modelo HLM3 nulo.
- ◉ **Modelo HLM3 com medidas repetidas (tendência linear) e com interceptos e inclinações aleatórios.** ←

Dado o seguinte código de programação no R, utilizado para se estimar um modelo a partir de um dataset com dados de evolução temporal anual (medidas repetidas) de indivíduos aninhados em determinados grupos:

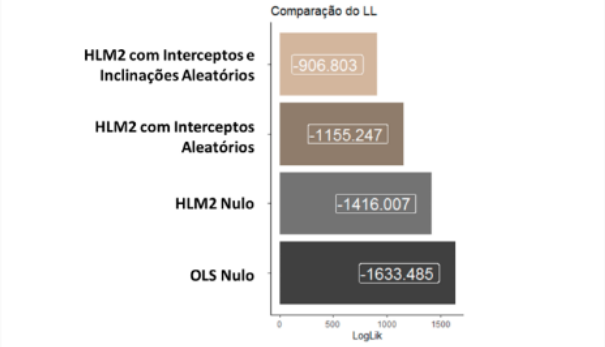
```
modelo <- lme(fixed = Y ~ ano + X + W + X:ano + W:ano,
              random = list(grupo  = ~ ano, indivíduo = ~ ano),
              data = dataset,
              method = "REML")
```

Sabe-se que Y é a variável dependente, X é a variável preditora de indivíduos e W é a variável preditora de grupos.

O código apresentado refere-se a uma modelagem do tipo:

- ○ HLM1.
- ○ HLM2 com inflação de zeros.
- ◉ **HLM3 com medidas repetidas. ←**
- ○ OLS.

A partir da estimação de quatro modelos por meio de um dataset com estrutura hierárquica nos dados, obtiveram-se os respectivos valores de LogLik, conforme mostra a figura a seguir:



Qual o modelo mais adequado para fins de melhor ajuste entre os valores previstos da variável dependente (*fitted values*) e valores reais?

- ◉ **HLM2 com Interceptos e Inclinações Aleatórios. ←**
- ○ OLS Nulo.
- ○ HLM2 com Interceptos Aleatórios.
- ○ HLM2 Nulo.

Para o estudo da evolução temporal do desempenho de estudantes pertencentes a diferentes escolas, estimou-se um modelo HLM3 com medidas repetidas. Neste caso, é correto dizer que os níveis 1, 2 e 3 desta modelagem serão caracterizados, respectivamente, por:

- ○ estudantes, escolas e evolução temporal.
- ○ evolução temporal, escolas e estudantes.
- ○ escolas, estudantes e evolução temporal.escolas, estudantes e evolução temporal.
- ◉ **a evolução temporal, estudantes e escolas. ←**

Seguindo a nomenclatura adotada em aula, apresenta-se a seguinte especificação de um modelo multinível.

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \varepsilon_{ij}$$

$$\beta_{0j} = \gamma_{00} + \nu_{0j}$$

$$\beta_{1j} = \gamma_{10}$$

que resulta na seguinte expressão:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \nu_{0j} + \varepsilon_{ij}$$

A partir da especificação apresentada, é possível definir o modelo proposto como sendo um:

- ◉ **Modelo HLM2 com interceptos aleatórios no nível 2. ←**
- ○ Modelo nulo HLM2.
- ○ Modelos HLM3 com medidas repetidas.
- ○ Modelos HLM3 com interceptos e inclinações aleatórios.

GESTÃO EM DATA SCIENCE & ANALYTICS

Transformação Digital

Em seu livro sobre Organizações Positivas, Robert E. Quinn apresenta cinco grandes dimensões dentro das quais há práticas positivas que podem melhorar o ambiente organizacional. Leia as dimensões a seguir e escolha a melhor alternativa:

1) Senso de Propósito Compartilhado
2) Diálogos Autênticos
3) Cinco Forças de Michael Porter
4) Visualização de Possibilidades
5) Foco no Bem Comum
6) Confiança no Processo Emergente
7) 4 P's do Marketing de Philip Kottler

- ○ 2, 3, 4, 5 e 6 são as cinco dimensões de positividade em organizações.
- ○ 1, 2, 3, 4 e 5 são as cinco dimensões de positividade em organizações.
- ○ 1, 2, 3, 5, e 7 são as cinco dimensões de positividade em organizações.
- ◉ **1, 2, 4, 5 e 6 são as cinco dimensões de positividade em organizações. ←**

A partir do *Competing Values Framework*, assinale a alternativa que descreve a estrutura de um ambiente direcionado à colaboração:

- ○ Lugar altamente estruturado e formal. Procedimentos e regras governam comportamentos.
- ○ Lugar dinâmico, empreendedor e criativo. Inovação e tomadas de risco são praticadas pelos indivíduos.
- ○ A realização do trabalho é o foco do direcionamento de resultados da organização.
- ◉ **A organização é um lugar aberto e amigável para se trabalhar, onde as pessoas compartilham muito de si mesmas. ←**

O que significa o termo "*stakeholders*" ou "partes interessadas"?

- ○ é a sociedade em geral.
- ○ são os colaboradores da organização.
- ○ são os acionistas do negócio.
- ◉ **são todos aqueles (pessoas ou grupos de pessoas) que impactam ou são impactadas pelos resultados da organização. ←**

A partir do *Framework* de Tensões Organizacionais, assinale a alternativa que contém somente aspectos da zona negativa:

- ○ Controle de custos e aspereza.
- ○ Controle gerencial e interesse próprio.
- ◉ **Exclusão e tendência à conformidade. ←**
- ○ Conflito e foco em resultado.

Dentre os 8 erros mais comuns dos processos de mudança, John Kotter afirma que o último erro seria, "não institucionalizar a mudança na cultura da organização", que está relacionado principalmente a:

- ○ não padronizar os novos processos operacionais padrão.
- ○ não automatizar a coleta de dados e a geração de um *scorecard* de desempenho.
- ○ não mexer na estrutura organizacional da empresa.
- ◉ **não incorporar a mudança no jeito de ser da organização. ←**

De forma mais simples, Kurt Lewin define três estágios do processo de mudança, que são:

- ○ Executar | Checar | Agir
- ○ Definir | Melhorar | Checar
- ○ Planejar | Executar | Controlar
- ◉ **Descongelar | Mudar | Recongelar ←**

A partir do *Competing Values Framework*, identifique a alternativa que apresenta os quadrantes que denominam os perfis culturais:

- Relações de competição, sistemas abertos, processos internos e metas racionais.
- Relações humanas, sistemas controlados, processos internos e metas de liderança.
- Relações de competição, sistemas abertos, processos externos e metas de liderança.
- **Relações humanas, sistemas abertos, processos internos e metas racionais. ←**

São considerados erros no processo de mudança:

- Empoderar as pessoas para trabalharem na nova visão, comunicar a visão, não formar uma poderosa coalisão.
- **Não comunicar a visão, declarar a vitória cedo demais, não institucionalizar a mudança na cultura da organização. ←**
- Não estabelecer um senso de urgência, não criar uma visão, institucionalizar as mudanças.
- Não formar uma poderosa coalisão, estabelecer um senso de urgência, declarar a vitória cedo demais.

O *Competing Values Framework* é um modelo que traz quatro tipologias de Cultura Organizacional. Qual alternativa a seguir mais bem descreve esses quatro tipos?

- Adocracia (cultura estruturação, excelência, controle) | Mercado (cultura de realização, agilidade, competição) | Hierarquia (cultura de grupo, colaboração) | Clã (cultura de criação, inovação, desenvolvimentista)
- Adocracia (cultura estruturação, excelência, controle) | Mercado (cultura de realização, agilidade, competição) | Hierarquia (cultura de criação, inovação, desenvolvimentista)| Clã (cultura de grupo, colaboração)
- **Adocracia (cultura de criação, inovação, desenvolvimentista) | Mercado (cultura de realização, agilidade, competição) | Hierarquia (cultura estruturação, excelência, controle) | Clã (cultura de grupo, colaboração) ←**
- Adocracia (cultura de grupo, colaboração) | Mercado (cultura de criação, inovação, desenvolvimentista) | Hierarquia (cultura de realização, agilidade, competição) | Clã (cultura estruturação, excelência, controle)

Para se construir a capacidade de mudança, são necessárias intervenções focadas nos membros, na estrutura e na cultura da organização, e tal construção envolve três áreas principais, exceto:

- uma **cultura** organizacional que facilite o aprendizado.
- um **contexto** organizacional que sustente a mudança.
- a **implementação** da mudança.
- **uma tecnologia de comunicação online. ←**

Metodologias Ágeis

Qual a quantidade recomendada de pessoas para uma equipe que trabalhe em decisões do sistema **caótico**?

- Até 2 pessoas.
- Até 20 pessoas.
- Até 150 pessoas.
- **Até 5 pessoas. ←**

Assinale a alternativa que preenche a lacuna corretamente:

"Algo é conhecido _____ quando é conhecido independentemente da experiência e através do pensamento apenas, da razão, dedução e reminiscência".

- majoritariamente
- **a priori ←**
- marginalmente
- a posteriori

Dados necessários para resolver problemas **incognoscíveis**, encontram-se no:

- **Futuro. ←**
- Passado.
- Não podem ser solucionados em hipótese alguma.
- Presente.

Qual a quantidade recomendada de pessoas para uma equipe que trabalhe em decisões do sistema **complexo**?

- Não há limites.
- Até 20 pessoas.
- Até 150 pessoas.
- **Até 15 pessoas.** ←

Dados necessários para resolver problemas **conhecidos**, encontram-se no:

- Não podem ser solucionados em hipótese alguma.
- Futuro.
- Presente.
- **Passado.** ←

Assinale a alternativa que apresenta *a* quantidade *de* domínios do framework Cynefin, apresentados em aula:

- 2.
- 6.
- 3.
- **5.** ←

São os três tipos de problemas organizacionais:

- **Conhecidos, Conhecíveis e Incognoscíveis.** ←
- Conhecidos, Conhecíveis e Desejáveis.
- Conhecidos, Conhecíveis e Caóticos.
- Conhecidos, Complexos e Incognoscíveis.

Assinale a alternativa que preenche a lacuna corretamente:

"Algo é conhecido _____ quando é conhecido através da experiência (sensorial ou introspectiva), da experimentação".

- a priori
- majoritariamente
- **a posteriori** ←
- marginalmente

"Não há qualquer relação de causa-efeito conhecida; gera crise se for acidental; quando contido, é ótimo para inovação; não é fácil criar e é impossível manter". De que tipo de **sistema** estamos falando?

- Catastrófico.
- Efeito Borboleta.
- Teoria da Complexidade.
- **Sistema Caótico.** ←

Aqui é o cenário ideal para o empirismo e adaptação. O caminho é construído à medida que caminhamos. Possui as respostas nas práticas do presente. De que domínio estamos falando?

- Óbvio.
- Complicado.
- Caótico.
- **Complexo.** ←

Assinale a alternativa que representa as **três "famílias" de desperdícios** na metodologia *Lean*:

- ○ Desordem, Caos e Preditivo.
- ◉ **Muda, Mura e Muri.** ←
- ○ Claro, Complicado e Complexo.
- ○ Projetos, Programas e Portfólios.

Assinale a alternativa que apresenta três dentre os sete desperdícios do Lean:

- ○ Superprodução, Transporte e Velocidade.
- ○ Defeitos, Processamento excessivo e Adaptabilidade.
- ◉ **Superprodução, Transporte e Espera.** ←
- ○ Inventário, Espera e Agilidade.

Em um planejamento ágil, o *Roadmap* consiste:

- ○ No acompanhamento mensal do progresso.
- ○ No que validarmos com o usuário por meio de um lançamento.
- ◉ **No plano que desenvolveremos para alcançar a Visão do produto.** ←
- ○ No acompanhamento diário do progresso.

Qual a melhor definição de ***Timebox***?

- ○ a quantidade média de duração de um evento, a partir de uma série histórica.
- ◉ **a quantidade máxima de tempo que um evento deve durar.** ←
- ○ a quantidade exata de tempo que um evento deve durar.
- ○ a quantidade mínima de tempo que um evento deve durar.

Assinale a alternativa que **NÃO** apresenta um dos sete desperdícios do Lean:

- ◉ **Agilidade.** ←
- ○ Superprodução.
- ○ Espera.
- ○ Transporte.

Assinale a alternativa que apresenta os **4 P'S do "Jeito Toyota"**:

- ◉ ***Problem Solving, People and Partners, Process, Philosophy*** ←
- ○ *Program, People and Partners, Process, Philosophy*
- ○ *Problem Solving, Pattern, Process, Philosophy*
- ○ *Problem Solving, Program, Process, Philosophy*

Em um planejamento ágil, assinale a alternativa que define o conceito de **Daily**:

- ○ O plano que desenvolveremos para alcançar a visão do produto.
- ◉ **O acompanhamento diário do progresso.** ←
- ○ O que validarmos com o usuário por meio de um lançamento.
- ○ Como, incrementalmente, construiremos nosso lançamento.

A perda causada por constantes interrupções, chamadas "*switching*", em uma situação na qual 5 atividades são feitas simultaneamente, segundo Gerald Weinberg, no livro *Quality software Management: Systems Thinking*, é de:

- (●) **75%.** ←
- ( ) 45%.
- ( ) 25%.
- ( ) 55%.

Em um planejamento ágil, o conceito de ***Sprint*** representa:

- ( ) O que validarmos com o usuário por meio de um lançamento.
- ( ) Um acompanhamento diário do progresso.
- (●) **Como, incrementalmente, construiremos nosso lançamento.** ←
- ( ) A visão do projeto.

Assinale a alternativa que **NÃO** apresenta um dos 4 P'S do "Jeito Toyota":

- ( ) *Problem Solving.*
- ( ) *Process.*
- ( ) *People and Partners.*
- (●) ***Program.*** ←

Business Intelligence

Qual dessas alternativas é uma opção viável para mostrar relação entre duas variáveis quantitativas contínuas:

- ( ) Gráfico de barras.
- ( ) Gráfico de colunas.
- (●) **Gráfico de dispersão.** ←
- ( ) Gráfico de setores (pizza).

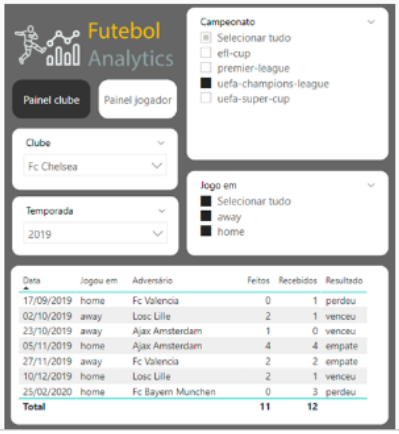
Em se tratando de **Business Intelligence**, podemos afirmar que:

- (●) **Refere-se ao processo de coleta, organização, análise, compartilhamento de informações que oferecem suporte à gestão de negócios.** ←
- ( ) Não é necessário haver conhecimento do dado para se apresentar uma informação.
- ( ) Refere-se ao processo de coleta, organização, análise, compartilhamento de informações. Porém, sem nenhuma relação com os objetivos do negócio.
- ( ) Refere-se a um software de criação de gráficos.

De acordo com a aula, um possível fluxo utilizado em projetos de **Business Intelligence**, visando apresentar uma informação a partir de um conjunto de dados, pode ser definido como:

- (●) **Conhecer os dados, definir o objetivo, organizar os dados, aplicar as análises e apresentar a informação.** ←
- ( ) Organizar os dados, conhecer os dados, definir o objetivo, aplicar as análises e apresentar a informação.
- ( ) Conhecer os dados, organizar os dados, aplicar as análises, definir o objetivo, e apresentar a informação.
- ( ) Organizar os dados, apresentar a informação, definir o objetivo, conhecer os dados e aplicar análises.

No exemplo a seguir envolvendo segmentação de dados. Qual seria a quantidade de gols feitos para o clube *Fc Chelsea*, caso o usuário do painel alterasse o segmentador de dados ***Jogo em*** para selecionar apenas os jogos classificados como *away*?



Data	Jogou em	Adversário	Feitos	Recebidos	Resultado
17/09/2019	home	Fc Valencia	0	1	perdeu
02/10/2019	away	Losc Lille	2	1	venceu
23/10/2019	away	Ajax Amsterdam	1	0	venceu
05/11/2019	home	Ajax Amsterdam	4	4	empate
27/11/2019	away	Fc Valencia	2	2	empate
10/12/2019	home	Losc Lille	2	1	venceu
25/02/2020	home	Fc Bayern Munchen	0	3	perdeu
Total			**11**	**12**	

- ( ) 3.
- ( ) 1.
- (•) **5. ←**
- ( ) 2.

Assinale a alternativa de software que **não** é uma opção viável para o desenvolvimento de um projeto de Business Intelligence:

- ( ) Power BI.
- (•) **Power Point.** ←
- ( ) Apache Superset.
- ( ) Tableau.

Considere as opções a seguir:

I. Quantidade de itens vendidos;

II. Quantidade de faltas de um aluno;

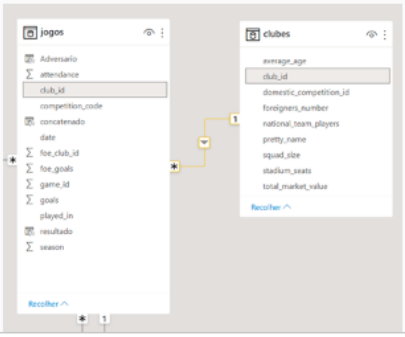
III. Quantidade de acidentes;

IV. Número de filhos.

Assinale a alternativa que apresenta apenas exemplos de variáveis quantitativas discretas.

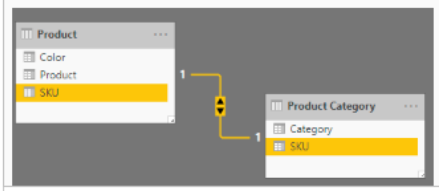
- ( ) Apenas I, II e III.
- ( ) Apenas a IV.
- ( ) Apenas I e III.
- (•) **Todas as alternativas.** ←

Em um projeto de Power BI, a tabela clubes e a tabela jogos estão relacionadas através de uma relação de cardinalidade pelo atributo *club_id*, que representa o identificador único de cada clube. Qual a relação de cardinalidade existente entre as tabelas?



- (●) **Um para muitos.** ←
- ( ) Um para um.
- ( ) Muitos para muitos.
- ( ) Nenhuma das alternativas.

Em um projeto de Power BI, a tabela Product e a tabela ProductCatetgory estão relacionadas através de uma relação de cardinalidade pelo atributo *SKU*. Qual a relação de cardinalidade existente entre as tabelas?



- ( ) Um para muitos.
- (●) **Um para um.** ←
- ( ) Muitos para muitos.
- ( ) Nenhuma das alternativas.

No exemplo a seguir envolvendo segmentação de dados. Qual seria a quantidade de gols recebidos para o clube *Fc Liverpool*, caso o usuário do painel alterasse o segmentador de dados ***Jogo em*** para selecionar apenas os jogos classificados como *home*?

- ( ) 8.
- ( ) 1.
- (●) **2.** ←
- ( ) 6.

Considere as opções a seguir:

I. Cor dos olhos;

II. Grau de escolaridade;

III. Estágio de doença;

IV. Altura.

Assinale a alternativa que apresenta apenas exemplos de variáveis qualitativas ordinárias.

- I e II.
- II e IV.
- I e IV.
- **II e III.** ←

Qual função DAX deve ser utilizada para criar uma medida que retorne a contagem de linhas de uma tabela?

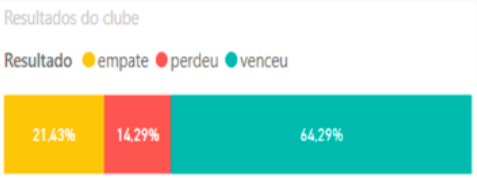
- FILTERROWS
- **COUNTROWS** ←
- DISTINCTCOUNT
- CALCULATEROWS

Foi criada uma coluna calculada chamada "RISCO" na tabela apresentada abaixo, utilizando a seguinte expressão: IF([VALOR]>4, "SIM", "NÃO"). Qual o output esperado em cada uma das linhas dessa coluna, na ordem correta?

VALOR	TIPO
5	A
4	A
3	B
5	B
6	A
2	A

- "NÃO","NÃO","NÃO","SIM","NÃO","NÃO"
- **"SIM","NÃO","NÃO","SIM","SIM","NÃO"** ←
- "SIM","SIM","SIM","SIM","NÃO","NÃO"
- "SIM","SIM","NÃO","NÃO","SIM","NÃO"

Analisando o gráfico dos resultados do Chelsea em 2020 na Champions League, é possível afirmar que:



- **Ele venceu mais da metade dos jogos que disputou.** ←
- Ele perdeu mais jogos do que empatou.
- Ele venceu todos os jogos que disputou.
- Ele perdeu mais da metade dos jogos que disputou.

Na situação da figura abaixo temos um gráfico de linha com o valor de mercado do jogador Kylian Mbappé ao longo do tempo. Os dados de valor de mercado estão apresentados na unidade de libras esterlinas. Qual foi o maior valor de mercado que esse jogador teve em sua carreira?

- 81 Mi
- 4 Mi
- **180 Mi** ←
- 144 Mi

Utilizando os dados apresentados na imagem abaixo, ao criar uma medida com a seguinte fórmula: SUM(VENDAS[VALOR_VENDA]). Qual deve ser o resultado esperado para essa medida? Obs.: nome da tabela da imagem: "VENDAS".

DATA	PRODUTO_ID	VALOR_VENDA
01/05/2022	1	R$355,9
03/05/2022	1	R$355,9
06/05/2022	3	R$99,9
06/05/2022	2	R$199,9
10/06/2022	2	R$199,9
15/05/2022	3	R$99,9
24/05/2022	3	R$99,9
28/05/2022	1	R$355,9
03/06/2022	1	R$355,9
03/06/2022	2	R$199,9
05/06/2022	3	R$99,9
10/06/2022	3	R$99,9

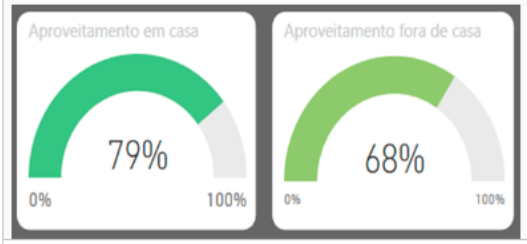
- R$ 3.270,00
- **R$ 2.522,80** ←
- R$ 1.200,00
- R$ 1.959,00

Utilizando os dados apresentados na imagem abaixo, ao criar um medida com a seguinte fórmula: CALCULATE(SUM(TESTE[VALOR]), TESTE[TIPO]="B"). Qual deve ser o resultado esperado para essa medida? Obs.: nome da tabela da imagem: "TESTE".

VALOR	TIPO	RISCO	RISCO 2
5	A	SIM	ALTO
4	A	NÃO	MEDIO
3	B	NÃO	MEDIO
5	B	SIM	ALTO
6	A	SIM	ALTO
2	A	NÃO	SEGURO

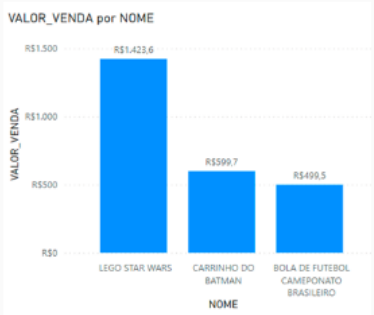
- 4
- 3
- **8** ←
- 6

Analisando o gráfico dos aproveitamentos de vitórias dentro e fora de casa do Liverpool em 2021 na Premier League, é possível afirmar que:



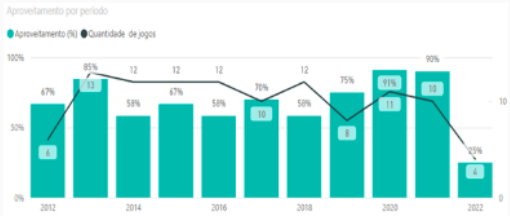
- (●) **O aproveitamento de vitórias do time é melhor quando ele joga em casa.** ←
- ( ) O aproveitamento de vitórias do time é igual dentro e fora de casa.
- ( ) O time não teve vitórias no campeonato.
- ( ) O aproveitamento de vitórias do time é melhor quando ele joga fora de casa.

Complete a frase que define a estrutura do gráfico na imagem abaixo. A estrutura apresentada é de um gráfico de _____, com o atributo "NOME" no eixo ___ e "VALOR_VENDA" no eixo _____. Assinale a alternativa que preenche as três lacunas, na ordem correta.



- ( ) Linha, Y, X
- ( ) Barras, Y, X
- ( ) Dispersão, X, Y
- (●) **Colunas, X, Y** ←

Observando o gráfico misto de colunas com linha abaixo, são apresentadas as seguintes informações: Aproveitamento de vitórias do Bayern de Munique em porcentagem ao longo dos anos e a quantidade de jogos disputados. Qual foi o ano em que esse clube teve o seu melhor aproveitamento, independentemente da quantidade de jogos?



- ( ) 2017
- (●) **2020** ←
- ( ) 2018
- ( ) 2021

Quais funções DAX devem ser utilizadas para somar e retornar a média (média aritmética) de todos os números de uma coluna, respectivamente?

- ( ) AVERAGE e SUM
- ( ) COUNTROWS e SUM
- ( ) FILTER e CALCULATE
- (●) **SUM e AVERAGE** ←

Tecnologia da Informação e Inovação Tecnológica

O uso eficiente da tecnologia permite que as empresas consigam:

I – Solucionar problemas e melhorar o processo de tomada de decisões

II- Ter melhor controle dos processos e melhorar o fluxo de informações

III – Reduzir os lucros e aumentar os custos de manutenção

Em relação às afirmações acima, assinale a alternativa correta:

- ○ Apenas I e III são verdadeiras
- ○ I e II são falsas
- ○ Apenas II e III são verdadeiras
- ◉ **Apenas I e II são verdadeiras ←**

Na etapa de **informação** do ciclo da gestão do conhecimento, estão inclusas as seguintes características:

- ◉ **KPI, dashboards, interpretação dos dados ←**
- ○ Otimização do negócio, KPI, dashboards e aplicação de padrões
- ○ Tomada de decisões de negócio, dashboards e interpretação dos dados
- ○ Otimização do negócio, assimilação de competências

Sobre o TOE (*Technology Organization Environment Framework*), afirma-se que:

I – É uma Teoria usada para explicar o contexto de uma organização na tomada de decisão;

II- O fator de tecnologia do TOE trata da avaliação de tecnologias disponíveis na empresa;

III – O fator ambiental do TOE compreende a influência de características gerais e rede social interna em relação ao comportamento da organização para adoção de tecnologias;

Em relação às afirmações acima, assinale a alternativa correta:

- ○ Apenas II e III são verdadeiras
- ○ I e II são falsas
- ○ Apenas I e III são verdadeiras
- ◉ **Apenas I e II são verdadeiras ←**

Empresas inovadoras possuem competências estratégicas e organizacionais. Assinale a alternativa que NÃO apresenta um exemplo de competência organizacional:

- ○ Capacidade de envolver toda a empresa no processo de mudança
- ◉ **Capacidade de identificar tendências de mercado ←**
- ○ Investimento em recursos humanos
- ○ Disposição e capacidade de gerenciamento de riscos

Na etapa de **sabedoria** do ciclo da gestão do conhecimento, estão inclusas as seguintes características:

- ○ Assimilação de competências, obtenção de vantagem competitiva e interpretação dos dados
- ○ Otimização do negócio, KPI, dashboards e aplicação de padrões
- ○ Tomada de decisões de negócio, dashboards e interpretação dos dados
- ◉ **Otimização do negócio, assimilação de competências e tomadas de decisões de negócio ←**

Assinale a etapa que apresenta os maiores níveis de valor e maturidade no **Ciclo da Gestão do Conhecimento**:

- ○ Informação
- ◉ **Sabedoria ←**
- ○ Dados
- ○ Conhecimento

Empresas inovadoras possuem competências estratégicas e organizacionais. Assinale a alternativa que **NÃO** se enquadra como uma competência estratégica:

- ○ Capacidade de reunir, processar e assimilar informações
- ◉ **Investimento em recursos humanos ←**
- ○ Visão de longo prazo
- ○ Capacidade de identificar tendências de mercado

A tecnologia da informação desempenha um papel fundamental para ajudar as organizações a:

- ◉ **Organizar dados e informações que serão utilizados por toda a empresa. ←**
- ○ Manter a estrutura burocrática existente.
- ○ Aumentar o retrabalho.
- ○ Trabalhar apenas com tarefas não rotineiras.

Considere as assertivas a seguir:

I. Relacionamento com Clientes

II. Empoderamento dos Funcionários

III. Otimização de operações

IV. Transformar produtos e serviços

Assinale a alternativa que apresenta as assertivas que apresentam **Pontos críticos para Transformação Digital**:

- ○ Apenas I e II
- ◉ **I, II, III e IV ←**
- ○ Apenas I, II e III
- ○ Apenas II, III e IV