SME0620 - Trabalho de Pesquisa

Análise de Dados com Modelos de Regressão, ANOVA e Teste Qui-Quadrado

Alunos:

Giovanni dos Santos – 13695341

Luciano Gonçalves Lopes Filho - 13676520

Professor: Vicente Garibay Cancho

São Carlos, 2024

Conteúdo

1 Introdução

A base de dados que tomamos como ponto de partida para o nosso trabalho busca reunir informações estatísticas relevantes para o estudo da saúde cardíaca em uma variedade de países. Este levantamento de dados é fundamental para entender as diferenças e similaridades nos perfis de saúde cardiovascular ao redor do mundo. Sob essa perspectiva, nosso objetivo é realizar uma avaliação estatística detalhada desses dados, aplicando métodos rigorosos de análise para obter insights valiosos.

Para atingir esse objetivo, utilizamos diversas técnicas estatísticas avançadas, incluindo a regressão, a análise de variância (ANOVA) e o teste qui-quadrado. Cada um desses métodos oferece uma abordagem única para examinar os dados e revelar padrões significativos. A regressão nos permite modelar e prever a relação entre variáveis, enquanto a ANOVA nos ajuda a comparar médias entre diferentes grupos para identificar variações significativas. O teste qui-quadrado, por sua vez, é empregado para avaliar associações entre categorias e determinar se as distribuições observadas diferem das esperadas.

Com a aplicação dessas técnicas, pretendemos traçar conclusões sólidas sobre os grupos que requerem maior atenção em relação à saúde cardíaca. Identificar esses grupos é crucial para direcionar esforços de prevenção e intervenção de maneira eficaz. Ao compreender quais segmentos da população estão mais vulneráveis a doenças cardíacas, podemos desenvolver estratégias específicas para mitigar riscos e promover a saúde cardiovascular.

2 Metodologia

2.1 Considerações

Os métodos estatísticos utilizados no trabalho estão descritos abaixo. Dessa maneira, vale ressaltar que nos utilizamos da linguagem R para realizar os cálculos e determinações necessários.

2.2 Definição das variáveis e suas categorias

Trazendo um pouco do trabalho anterior, aqui está a definição das variáveis que compõe a base de dados que utilizamos:

Tabela 1: Variáveis Qualitativas

Variável	**Subcategoria**	**Descrição**
Sexo (Sex)	Nominal	1 = macho 0 = fêmea
Tipo de dor no peito (Chest Pain Type)	Nominal	1 = Angina típico 2 = Angina atípico 3 = Dor não-anginosa 4 = assintomático
Açúcar no sangue em jejum (Fasting Blood Sugar)	Nominal	$nível > 120mg/dl$ 1 = verdadeiro, 0 = falso
Eletrocardiograma em repouso (Resting Eletrocardiogram Results)	Nominal	0 = normal 1 = ondas ST-T anormais 2 = Hipertrofia do ventrículo esquerdo
Angina decorrente de exercício (Exercise Induced Angina)	Nominal	1 = sim 0 = não
Inclinação de ST no exercício (ST slope)	Nominal	1 = crescente 2 = estática 3 = decrescente
Classe (Target)	Nominal	1 = doença cardíaca 0 = normal

Tabela 2: Variáveis Quantitativas

Variável	**Subcategoria**
Idade (Age)	Discreta
Pressão do sangue em repouso (Resting bp s)	Discreta
Colesterol no soro sanguíneo (Cholesterol)	Discreta
Máxima taxa de batimentos cardíacos atingida (Max heart rate)	Discreta
Depressão do Seguimento ST (oldpeak)	Contínua

2.3 Regressão

Em um contexto de duas variáveis, uma independente e uma dependente da outra, podemos traçar a correlação entre essas variáveis, o que indica o quanto de "influência"que a variável independente analisada causa na dependente. A correlação linear entre duas variáveis pode ser descrita da seguinte forma:

$$\rho_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{1}$$

onde:

- X_i e Y_i são os valores das variáveis X e Y respectivamente,
- $\overline{X}$ é a média dos valores de X,
- $\overline{Y}$ é a média dos valores de Y,
- n é o número de pares de valores.

Para valores de correlação menores que 50%, temos uma correlação fraca, enquanto para acima disso temos uma correlação forte, o que permite que façamos uso do método de regressão linear para descrever a relação entre as variáveis. A forma com que a regressão linear funciona é buscando uma equação de reta que represente a relação aproximada das variáveis:

$$Y = a + bX \tag{2}$$

Os coeficientes são estimados usando o método dos mínimos quadrados, que minimiza a soma dos quadrados dos resíduos (diferenças entre os valores observados e os valores previstos). As fórmulas para β_0 e β_1 são:

$$b = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{3}$$

$$a = \overline{Y} - b\overline{X} \tag{4}$$

onde:

- X_i e Y_i são os valores das variáveis X e Y respectivamente,
- $\overline{X}$ é a média dos valores de X,
- $\overline{Y}$ é a média dos valores de Y,
- n é o número de pares de valores.

2.4 ANOVA

A ANOVA consiste em uma análise de variância, com o objetivo de comparar a média de 3 ou mais grupos. Nesse contexto, desde que garantirmos que a distribuição analisada se adeque a normalidade, essa análise divide-se em duas: a variação dentro dos grupos e entre eles. Para realizar a análise inter-grupos, utiliza-se da hipótese nula, a qual diz que não há diferenças significativas entre as médias dos grupos, e também da hipótese 1, a qual diz que essas diferenças não são desprezíveis.

Sob a ótica do elucidado acima, a ANOVA utiliza-se da estatística F, a qual é calculada da seguinte forma:

$$F = \frac{\mathrm{MS}_{\mathrm{entre}}}{\mathrm{MS}_{\mathrm{dentro}}} \tag{5}$$

onde:

- $\mathrm{MS}_{\mathrm{entre}}$ é a média dos quadrados entre os grupos
- $\mathrm{MS}_{\mathrm{dentro}}$ é a média dos quadrados dentro dos grupos

Estas médias dos quadrados são calculadas como:

$$\mathrm{MS}_{\mathrm{entre}} = \frac{\mathrm{SS}_{\mathrm{entre}}}{\mathrm{df}_{\mathrm{entre}}} \tag{6}$$

$$\mathrm{MS}_{\text{dentro}} = \frac{\mathrm{SS}_{\text{dentro}}}{\mathrm{df}_{\text{dentro}}} \tag{7}$$

onde:

- $\mathrm{SS}_{\text{entre}}$ é a soma dos quadrados entre os grupos,
- $\mathrm{SS}_{\text{dentro}}$ é a soma dos quadrados dentro dos grupos,
- $\mathrm{df}_{\text{entre}}$ são os graus de liberdade entre os grupos, calculados como $k-1$, onde k é o número de grupos,
- $\mathrm{df}_{\text{dentro}}$ são os graus de liberdade dentro dos grupos, calculados como $N-k$, onde N é o número total de observações.

Com os valores de F, de df_{entre} e de df_{dentro}, podemos determinar o valor de F crítico nos baseando em algum valor arbitrário de significância. Dessa forma, pode-se comparar para ver se o nosso valor de F pertence ao intervalo $0 < F < F_{crítico}$ ou se não. No primeiro caso, podemos afirmar que a análise de variância se trata da Hipótese Nula, ou seja, que não há diferenças estatiscamente relevantes entre as médias dos grupos. Em contrapartida, no segundo caso a situação é a da Hipótese 1, assim as diferenças são relevantes.

2.5 Teste Qui-Quadrado

O Teste Qui-Quadrado consiste em uma análise de frequências entre diferentes grupos, buscando entender se as diferenças entre elas são relevantes ou não. Nesse cenário, para o funcionamento correto do teste, é necessário que as frequências esperadas não sejam muito pequenas, assim não podendo serem menores que 5. No caso de serem menor do que 10, é possível aplicar a Correça de Yates para resolver a situação. Para determinar o valor necessário para a análise, utiliza-se da seguinte fórmula:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{8}$$

onde:

- O_i é o valor observado na categoria i,
- E_i é o valor esperado na categoria i,
- n é o número de categorias.

Os valores esperados (E_i) são calculados com base na hipótese nula. Para o caso de uma tabela de contingência, os valores esperados são calculados como:

$$E_{ij} = \frac{R_i \cdot C_j}{N} \tag{9}$$

onde:

- E_{ij} é o valor esperado na célula da linha i e coluna j,
- R_i é o total da linha i,
- C_j é o total da coluna j,
- N é o total geral das observações.

Tendo esse valor determinado, podemos utilizar dos graus de liberdade e de uma significância arbitrária escolhida para determinar o valor de $\chi^2_{crítico}$ através da tabela. Dessa forma, caso o valor encontrado de χ^2 encontre-se no intervalo $0 < \chi^2 < \chi^2_{crítico}$, as diferenças entre as frequências podem ser desconsideradas. Entretanto, no caso contrário elas devem ser levadas em consideração.

3 Resultados

3.1 Regressão Linear

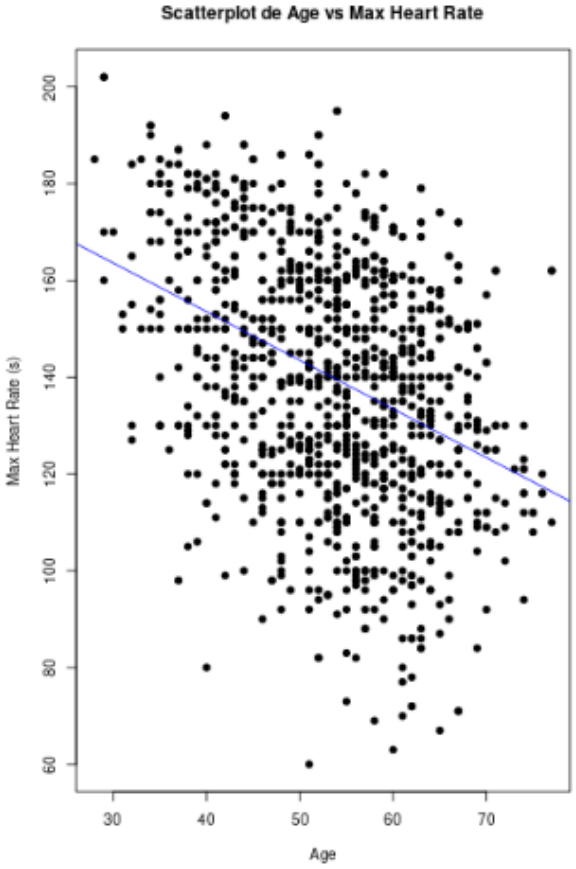

Figura 1: Gráfico de Dispersão com a melhor reta

Tabela 3: Coeficientes do Modelo de Regressão

Termo	**Estima**	**Erro Padrão**	**Valor t**	$Pr > \|t\|$
α	193.73729	4.00996	48.31	$< 2 \times 10^{-16}$
β	-1.00529	0.07354	-13.67	$< 2 \times 10^{-16}$

Os dados relativos a regressão linear estão dispostos na tabela e gráfico acima. Podemos perceber pelo gráfico que as variáveis "idade"e "máxima

frequência cardíaca"tem baixa correlação ($|\rho| = 0.368$). Ainda assim, podemos perceber que a frequência máxima cardíaca decai com a idade. Tendência refletida pela melhor reta e seu coeficiente $\beta = -1.00529$.

3.2 Teste ANOVA

Tabela 4: Coeficientes do ANOVA

	df	**Soma dos Quadrados**	**Somas das Medias**	**Valor F**
Entre	4	14379	3595	47.46
Dentro	1185	89749	76	

O teste de ANOVA foi feito considerando que o dataset estudado é uma colagem de pesquisas feitas em 5 diferentes lugares: Cleveland-OH, Hungria, Suica, Long Beach-VA, e Irvine-CA. Dessa forma, gostariamos de saber se as idades dos participantes dessa pesquisa mantinham uma média padrão ou se havia uma discrepância muito notória.

Considerando uma significancia de $\alpha = 0,05$ e lembrando que o dataset tem 4 graus de liberdade entre grupos e 1185 dentro dos grupos, chegamos ao valor valor de F critico de $F_{crit} = 2,3794$. Como $F > F_{crit}$, rejeitamos a hipótese nula. A discrepância entre as médias pode ter diversos motivos, dentre eles as politicas de saúde largamente diferente entre as localidades, que podem fazer com que pessoas mais novas se submetam a exames e sejam diagnosticadas mais cedo.

3.3 Teste Qui-Quadrado

Tabela 5: Tabela de Contingencia

χ^2	**df**	**P-Valor**
28.455	6	$7.713 * 10^{-5}$

Para o teste de qui-quadrado, escolhemos as variaveis "resting.ecg", que analisa o tipo de curva do eletrocardiograma durante o movimento de sistole (contração do músculo cardíaco), e "chest.pain.type", que descreve o tipo de dor do paciente, ambas qualitativas. Como o P-valor eh menor que 0,05, podemos rejeitar a hipótese nula de que não ha associação entre o tipo de sistole e o tipo de dor, sugerindo que o tipo de dor pode estar associado ao tipo de contração do músculo cardíaco.

4 Conclusão

Primeiramente, é necessário notar as limitações do dataset utilizado neste trabalho, o qual pode ser colocado como marginalmente satisfatório para fins de um estudo introdutório de estatística básica, porém se apresenta extremamente raso e limitado para uma análise detalhada dos fatores causadores de doenças cardíacas ao redor do mundo. Ainda assim, a análise, mesmo que simples desses dados, levanta algumas observações que demonstram a invariável importância da conscientização da população acerca da realização de exames preventivos das doenças de coração, que agem de forma silenciosa entretanto devastadora no longo prazo.

5 Fontes

World Health Organization. WHO Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. Geneva: World Health Organization, 2020.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

https://www.kaggle.com/datasets/mexwell/heart-disease-dataset?select=documentation.pdf