Enap

Monitoramento da biodiversidade: gestão, análise e síntese dos dados

Módulo 5

Noções de estatística no âmbito do Programa Monitora III

Fundação Escola Nacional de Administração Pública

Presidente
Diogo Godinho Ramos Costa

Diretor de Educação Continuada
Paulo Marques

Coordenador-Geral de Educação a Distância
Carlos Eduardo dos Santos

Conteudista/s
Jumara M. Souza (conteudista ICMBio, 2020)

Equipe responsável:
Ana Paula Medeiros Araújo (produção gráfica, 2020)
Bruna W. F. Miranda (CGGP/ICMBio, 2020)
Guilherme Telles (implementação Articulate, 2020)
Juliana Bermudez (revisão textual, 2020)
Kamila S. N. Oliveira (pedagoga ICMBio, 2020)
Lavínia Cavalcanti Martini Teixeira dos Santos (coordenadora, 2020)
Michelli Lopes (implementação Moodle, 2020)
Priscila Campos Pereira (coordenadora, 2020)
Rosana L. S. Siqueira (CGGP/ICMBio, 2020)
Sheila Rodrigues de Freitas (coordenação web, 2020)
Tathiana C. de Souza (coordenadora ComobOMOB/ICMBio, 2020)
Ugo José B. Bezerra (coordenador substituto ComobOMOB/ICMBio, 2020)
Vanessa Mubarak Albim (diagramação, 2020)

Curso produzido em Brasília 2020.

Desenvolvimento do curso realizado no âmbito do acordo de Cooperação Técnica FUB / CDT / Laboratório Latitude e Enap.

Enap Escola Nacional de Administração Pública

Enap, 2020

Enap Escola Nacional de Administração Pública
Diretoria de Educação Continuada
SAIS - Área 2-A - 70610-900 — Brasília, DF

Sumário

Módulo 5 Noções de estatística no âmbito do Programa Monitora III

Unidade 1 - Noções sobre estatística inferencial: testes estatísticos

Ao final desta unidade, você deverá ser capaz de aplicar os testes estatísticos aos dados coletados pelo Programa Monitora.

1.1 Contextualização

Você pode estar se perguntando: “por que precisamos aprender a executar as análises estatísticas se existem vários programas que fazem isso automaticamente?”

Apesar de utilizarmos os programas estatísticos durante as análises, é muito importante saber como o processo é feito. Embora os softwares estatísticos façam todos os cálculos, na maioria dos casos, eles não analisam se, por exemplo, o tipo do dado inserido faz sentido no contexto da análise.

Vale ressaltar que muitos dos programas estatísticos exigem que tenhamos ao menos uma noção básica sobre a análise que pretendemos aplicar. Por isso, vamos conhecer alguns testes estatísticos utilizados nas análises inferenciais conduzidas no Monitora.

De forma simples, as análises estatísticas podem ser divididas em paramétricas e não paramétricas.

- **Análises paramétricas**
  São um grupo de análises para as quais os dados atendem as premissas de distribuição normal ou aproximadamente normal e homogeneidade das variâncias (homocedasticidade). Como exemplo, podemos citar o teste t e a análise de variância.

- **Análises não paramétricas**
  Quando os dados não atendem às premissas supracitadas, são utilizadas as análises não paramétricas, para as quais não é necessário conhecer a distribuição dos dados. Como exemplo, podemos citar o teste de Kruskal-Wallis.

Vamos focar nos cenários mais comuns dentro do nosso monitoramento. Começaremos apresentando as análises de teste t e Anova, detalhando o raciocínio e as equações envolvidas. Por fim, falaremos brevemente sobre as análises não paramétricas.

1.2 Teste t de Student

O teste t é aplicado quando pretendemos analisar as diferenças entre dois grupos de dados.

DESTAQUE

Basicamente, ele consiste em um teste de comparação de médias entre amostras e pode ser calculado de formas diferentes para dados dependentes (pareados) e independentes (não pareados), bem como variâncias iguais (homogêneas) e diferentes.

Aqui, iremos conhecer o teste t para dados independentes e variâncias similares. Ele é calculado por meio da seguinte equação:

$$tcalc=(\bar{X}_A-\bar{X}_B)/\sqrt{(S_0^2(1/n_A+1/n_B))}$$

Onde:

- $\bar{X}_A$ e $\bar{X}_B$ são as médias dos dois conjuntos de dados (A e B).
- n_A e n_B são os números de elementos de A e B.
- S_0^2 é a variância comum às duas amostras, dada por:

$S_0^2=((n_A-1)S_A^2+(n_B-1)S_B^2)/(n_A+n_B-2)$, sendo S_A^2 e S_B^2 as variâncias das amostras.

Acompanhe o exemplo a seguir:

Após a conclusão do segundo ano de coleta de dados do protocolo de caranguejo-uçá, subprograma Marinho Costeiro, no Parna XYZ, a equipe gestora tem interesse em saber se as mudanças observadas nos valores do diâmetro das galerias (DG) de um ano para o outro são significativas. Os dados coletados estão tabelados a seguir:

Valores do diâmetro da galeria de caranguejo-uçá coletados nos anos 1 e 2 no Parna XYZ.

Ano 1	Ano 2
10	18
12	13
11	11
10	11
15	14
14	17
16	12
18	10
	10

Hipóteses:

H0 – Não há diferença entre os valores coletados de DG nas duas campanhas.
H1 – Há diferença entre os valores coletados de DG nas duas campanhas.

Para os cálculos, a letra A será referente aos parâmetros do Ano 1 e a letra B aos parâmetros do Ano 2. Após calcular a média e a variância dos dados amostrados, sintetizados na tabela a seguir, vamos calcular o valor de t.

	Ano 1 (A)	Ano 2 (B)
Média ($\bar{X}$)	13,25	12,88
Nº de galerias (n)	8	9
Variância (S^2)	8,78	8,61

Média, variância e número de galerias amostradas em dois anos no Parna XYZ.

Primeiro, vamos calcular a variância comum:

$$S_0{}^2=((n_A-1)S_A{}^2+(n_B-1)S_B{}^2/(n_A+n_B-2)=((8-1)*8{,}78+(9-1)*12{,}88)/(8+9-2)=8{,}69$$

Substituindo $S_0{}^2$ na equação do teste t, temos:

$$tcalc=(\bar{X}_A-\bar{X}_B)/(\sqrt{(S_0{}^2(1/n_A+1/n_B))}=(13{,}25-12{,}88)/\sqrt{(8{,}69*(1/8+1/9))}=0{,}25$$

Após calcularmos o valor de t, é hora de tomarmos a decisão de rejeitar ou não a hipótese nula. Para isso, precisamos observar se o resultado da análise está ou não na região crítica da curva de distribuição. Isso é feito por meio da comparação do valor de t crítico (um valor tabelado) com o valor de t calculado.

DESTAQUE

Assim, se o t calculado for maior que o t crítico, rejeitamos H0. Caso contrário, t calculado ≤ t crítico, aceitamos H0. programas de monitoramento, como o Monitora, contam uma história.

Para encontrar o valor de t tabelado, precisamos do nível de significância (que adotaremos aqui como 0,05) e do grau de liberdade (GL), dado por n_A+n_B-2, que é igual a 15.

Na tabela com os valores de t crítico, disponível no arquivo *Valores críticos* de *t* e *F* na biblioteca do curso, você vai encontrar que o valor de t tabelado para α de 0,05 e GL de 15 é igual a 2,13. Como nosso t calculado é menor que o t crítico, aceitamos H0.

DESTAQUE

Com isso, podemos concluir que a diferença entre as médias de diâmetro das galerias coletadas nos anos 1 e 2 não é significativa.

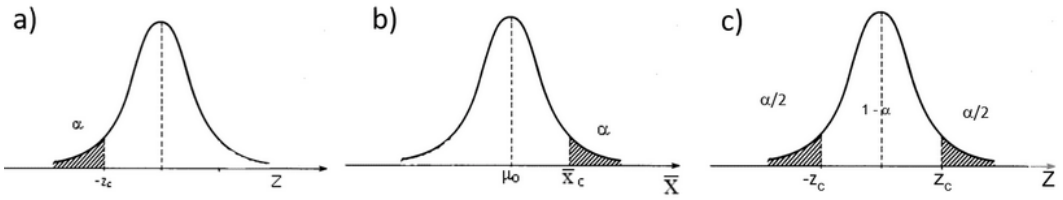
SAIBA MAIS

A distribuição t é tabelada em função do tamanho da amostra (n) ou em função dos graus de liberdade da distribuição (n-1). Cada linha da tabela refere-se a uma distribuição particular e cada coluna refere-se a um determinado nível de significância. Conforme a tabela, o nível de significância poderá ser unilateral (uma cauda) ou bilateral (duas caudas). Dizemos que um teste é de uma cauda quando esperamos que os valores para a hipótese alternativa se concentrem em um dos lados do gráfico de distribuição, conforme demonstrado na figura aseguir. Por exemplo, quando estabelecemos, por meio da hipótese alternativa, que é esperada uma redução ou um aumento no indicador monitorado.

Já no teste de duas caudas, nossa hipótese alternativa é de que há diferença entre os valores que estamos analisando, mas não definimos se essa diferença é um aumento ou uma redução no valor do indicador, conforme demonstrado na figura a seguir.

Perceba que a definição das caudas é feita no momento de construção das hipóteses e baseia-se no conhecimento sobre o conjunto de dados e na

pergunta que pretendemos responder. Em caso de dúvida, considere sua análise como sendo de duas caudas.



Nas publicações, encontraremos o valor de P. Ele já é calculado pelos programas estatísticos e não é necessário consultar uma tabela para tomar a decisão.

No caso do nosso exemplo, o valor-P é 0,80. Comparando esse valor com α, temos P > 0,05, assim tomamos a decisão de aceitar a hipótese nula.

Por ser robusto, os resultados do teste t ainda são confiáveis, mesmo que os dados se distanciem de uma distribuição normal, porém as amostras precisam ter tamanhos iguais e o teste ser bilateral.

O exemplo que acabamos de resolver mostra uma das várias possibilidades de abordagem do teste t.

1.3 Análise de variância (Anova)

Conforme abordado, o teste t foi desenhado para compararmos as médias de duas amostras. No entanto, podem existir situações em que queremos comparar mais de duas, por exemplo, as médias de três ou quatro estações amostrais.

Nessas situações, não é aconselhável usarmos o teste t, pois é muito trabalhoso comparar as médias par a par, além de aumentar as chances de cometermos o erro tipo I: concluir que há diferença quando ela não existe.

DESTAQUE

A solução para isso é a análise de variância, também conhecida como Anova, que, em apenas um teste, compara diversas médias e indica se há diferença entre elas.

Vale ressaltar que a Anova indica que há diferença, mas não qual das médias é diferente.

A Anova é uma análise que envolve variáveis quantitativas dependentes e categóricas independentes. Ela é dita de um fator (*one way*) quando testa uma variável quantitativa contra uma única variável categórica e de dois fatores (*two-way*) quando envolve mais de uma variável categórica. Vamos conhecer a Anova **one way**.

DESTAQUE

Basicamente, na Anova, divide-se a fonte de variância em dois componentes:

I. Variância entre grupos: resultado do tratamento ou do que estamos analisando.

II. Variância dentro do grupo: representa a variância aleatória, do erro ou residual.

Depois de calculadas as variâncias, é calculada a razão F de variâncias (ou teste F) por meio da seguinte fórmula:

$$F = Variância\ entre\ grupos / Variância\ dentro\ do\ grupo$$

Com o resultado do teste F, podemos calcular o valor de P e comparar o valor do F calculado com o tabelado, decidindo sobre a hipótese nula.

Diversos cálculos são utilizados para saber os valores das variâncias. Primeiro, é calculada a soma dos quadrados (SQ) para estimar o quanto cada conjunto de dados difere em relação a sua média. Então, calcula-se o quadrado médio (QM), que é a variância do conjunto de dados. Os quadrados médios da variância entre e dentro dos grupos serão utilizados no teste F. Acompanhe o exemplo a seguir:

Após a finalização das coletas do alvo plantas, componente Florestal, nas três estações amostrais (EA1, EA2 e EA3) da Resex K, a equipe gestora começou a indagar se havia algum efeito das EAs sobre os dados coletados, principalmente para os dados de altura das plantas. Desse modo, podemos pensar nas seguintes hipóteses:

H0 – Não há diferença entre as alturas das plantas para as diferentes EAs.
H1 – Há diferença entre as alturas das plantas nas diferentes EAs.

Vamos adotar 0,05 como nosso nível de significância ($\alpha = 0,05$). Nesse exemplo, nossa variável categórica independente é a estação amostral com três níveis e a variável resposta é a altura das plantas.

A seguinte tabela resume o nosso conjunto de dados:

	EA1	EA2	EA3	Total
	8,2	14	2,8	
	11	8,2	3	
	22	6,3	12,5	
	14	2,2	6,5	
	8,2	6,8	9,3	
	6,3	3,4	7,8	
	5,2	5,2	15,3	
	4,3	15,8	14,3	
	6,8	4,2	8,9	
	12	4,6	14	
	13,5	3,2	5	
	14,8	16,8	6,3	
	11,8	6,8	5,2	
	8,2	2,5	17,9	
	4,5	2,8	14,2	
Nº de indivíduos				
(n)	15	15	15	45
ΣX	150,8	102,8	143	396,6
ΣX^2	1838,96	1033,86	1690,24	4563,06
Média ($\bar{X}$)	10,05	6,85	9,53	
Desvio padrão	4,8	4,85	4,83	

Altura das plantas amostradas em três estações amostrais da Resex K.

A partir desses dados, vamos calcular a Anova.

Constante de correção (C)

$$\boldsymbol{C}=(\Sigma x)^2/\Sigma n_i = (\Sigma x)^2/\Sigma n_i = (396{,}6)^2/45=3495{,}37$$

Soma dos quadrados (SQ)

$$\boldsymbol{SQtotal}= \Sigma x^2 - C = 4563{,}06 - 3495{,}37 = 1067{,}69$$

$$\boldsymbol{SQentre}= \Sigma((\Sigma x_i)/n_i) - C = (150{,}8^2 + 102{,}8^2 + 143^2)/15 - 3495{,}37 = 88{,}46$$

$$\boldsymbol{SQdentro}= SQ_{total} - SQ_{entre} = 1067{,}69 - 88{,}46 = 979{,}22$$

Grau de liberdade (GL)

$$\boldsymbol{GLtotal}=(\Sigma n_i)-1=45-1=44$$

$$\boldsymbol{GLentre}=n^{\underline{o}}\ de\ tratamentos-1=3-1=2$$

$$\boldsymbol{GLdentro}=(\Sigma n_i)-n^{\underline{o}}\ de\ tratamentos=45-3=42$$

Quadrado médio (QM)

$$\boldsymbol{QM}_{entre}=\boldsymbol{SQ}_{entre}/\boldsymbol{GL}_{entre}=88{,}46/2=44{,}23$$

$$\boldsymbol{QM}_{dentro}=\boldsymbol{SQ}_{dentro}/\boldsymbol{GL}_{dentro}=979{,}22/42=23{,}31$$

Teste F

$$\boldsymbol{F}_{calc}=\boldsymbol{QM}_{entre}/\boldsymbol{QM}_{dentro}=44{,}23/23{,}31=1{,}89$$

Esses resultados comumente são apresentados em uma tabela como esta:

Fonte de variação	Graus de liberdade (GL)	E Soma dos quadrados (SQ)	Quadrado médio (QM)	Razão F	Valor de P
Entre grupos	2	88,46	44,23	1,89	0,16
Dentro do grupo (resíduos)	42	979,22	23,31		
Total	44	1067,69			

Tabela resumo da Anova one way para os dados da altura de plantas amostradas na Resex K.

Após calculamos o valor da estatística do teste, nesse caso, o valor de F, é hora de tomarmos uma decisão sobre a hipótese nula. Para isso, precisamos observar se o resultado da análise está ou não na região crítica da curva de distribuição.

DESTAQUE

Isso é feito por meio da comparação do valor de F tabelado com o valor de F calculado, de modo que: se F calculado > F tabelado, rejeitamos H0; caso contrário (F calculado < F tabelado), aceitamos H0.

Na tabela de distribuição F, disponível no arquivo *Valores críticos de t e F* na biblioteca do curso, o valor tabelado está relacionado ao grau de liberdade do numerador (GL entre = 2) e do denominador (GL dentro = 42) da razão F. Assim, considerando os graus de liberdade e α, temos que o F tabelado para o nosso exemplo é 4,03.

DESTAQUE

Como nosso valor de F calculado (1,89) é menor que o tabelado (4,03), nós não rejeitamos a hipótese nula. Ou seja, as alturas das plantas não são significativamente diferentes entre as estações amostrais.

Mas e se o resultado indicasse uma diferença, como saberíamos qual das estações tem a média diferente? Quando houver uma diferença significativa entre as médias, é necessário conduzir um teste a posteriori. Existe uma variedade desses testes, sendo os mais comuns os de Tukey, Duncan e Scheffé.

Agora que conhecemos as principais abordagens paramétricas usadas pelo Monitora, é hora de conversarmos sobre as análises não paramétricas.

1.4 Análises não paramétricas

As análises não paramétricas são abordagens estatísticas que não seguem o pressuposto de que os dados necessitam apresentar distribuição normal ou aproximadamente normal, nem o de homocedasticidade dos dados. Por isso, os testes não paramétricos também são conhecidos como testes de distribuição livre, pois não dependem do conhecimento da distribuição dos dados.

Quase todas as análises paramétricas têm seu correspondente não paramétrico, como o teste t (paramétrico) tem no teste U de Wilcoxon-Mann-Whitney seu correspondente não paramétrico. Para a Anova, o correspondente não paramétrico é o teste de Kruskal-Wallis.

DESTAQUE

Assim, quando não se conhece a distribuição dos dados ou quando eles não atendem aos pressupostos dos testes clássicos, recomenda-se o uso dos testes não paramétricos. Além disso, eles também são utilizados quando lidamos com variáveis nominais, ou seja, categóricas.

Tenha atenção, pois utilizar testes não paramétricos quando os dados atendem aos pressupostos de normalidade e homocedasticidade diminui a eficiência da análise.

Além disso, sempre que possível, recomendamos a escolha de uma abordagem paramétrica, pois apresenta um poder de teste maior.

IMPORTANTE

Então, antes de aplicar uma abordagem não paramétrica, tente transformar os dados para normalizar a distribuição e diminuir as variâncias, ou seja, tente transformá-los para que tenham uma distribuição normal, atingindo os requisitos necessários de um teste paramétrico.

As principais transformações aplicadas a conjuntos de dados são:

- **Logarítmica**
  A transformação logarítmica é usada para corrigir distribuições assimétricas e remover a dependência entre média e variância, além de homogeneizar variâncias entre grupos.

- **Raiz quadrada**
  A transformação de raiz quadrada é usada principalmente para dados de contagem, como os coletados no protocolo de peixes que integra o componente Igarapé. Ela também tem o objetivo de normalizar os dados e diminuir a variância.

- **Arco-seno da raiz quadrada ou angular**
  A transformação arco-seno da raiz quadrada ou angular é muito aplicada para dados de porcentagem, como no caso do indicador de proporções relativas de borboletas frugívoras.

Embora sejam as principais, outras transformações podem ser aplicadas a seu conjunto de dados.

DESTAQUE

O mais importante é que, independente da transformação utilizada, precisamos ter atenção para que os dados, depois de transformados, não percam seu significado biológico. Isso vale para a hora de produção dos gráficos, quando devemos utilizar os dados não transformados.

Unidade 2 - Noções sobre modelos lineares

Ao final desta unidade, você deverá ser capaz de explicar a aplicação de modelos lineares no âmbito do Programa Monitora.

2.1 Contextualização

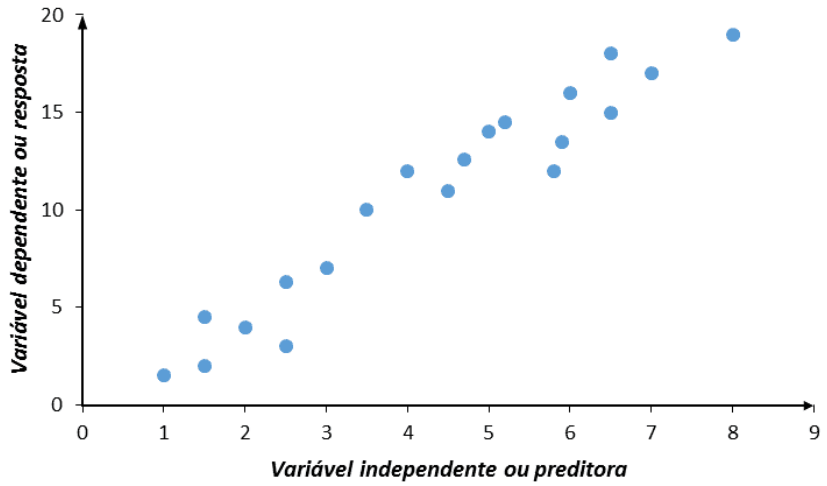
Os modelos tentam prever o comportamento de uma ou mais variáveis em função de outra. Os modelos lineares abarcam uma série de análises em que as variáveis possuem uma relação

linear e que a variável resposta é contínua. Isso inclui teste t, Anova, análise de covariância e regressões.

Vamos conhecer um pouco sobre a regressão linear simples e os modelos lineares generalizados. Mas antes, relembre esses dois conceitos importantes:

- **Variável dependente ou resposta**
  É referente àquilo que monitoramos, ou seja, é a variável sobre a qual buscamos entender o comportamento. Geralmente, é associada à letra Y, pois, numa representação gráfica, ela deve ser plotada sempre no eixo vertical, conhecido como eixo y.

- **Variável independente, explanatória ou preditora**
  É referente àquilo que provoca a variação sobre a variável Y, ou seja, é sobre essa variável que modelamos a resposta da variável dependente. Geralmente, é associada à letra X, pois, em uma representação gráfica, ela deve ser plotada sempre no eixo horizontal, conhecido como eixo x.



2.2 Regressão linear simples

Utilizamos a regressão para analisar a existência de relações, geralmente de causa e efeito, entre conjuntos de variáveis quantitativas, ou seja, buscamos entender se o valor de uma variável causa ou afeta o valor da outra.

DESTAQUE

Por exemplo, por meio de uma regressão, podemos avaliar se o tamanho do corpo do caranguejo afeta o diâmetro da sua galeria.

Na regressão linear simples, é produzida uma equação da reta (um modelo) que descreve a relação linear entre uma variável preditora (independente) e uma variável resposta (dependente). Essa equação é dada por:

$$Y=\beta_0+\beta_1 X$$

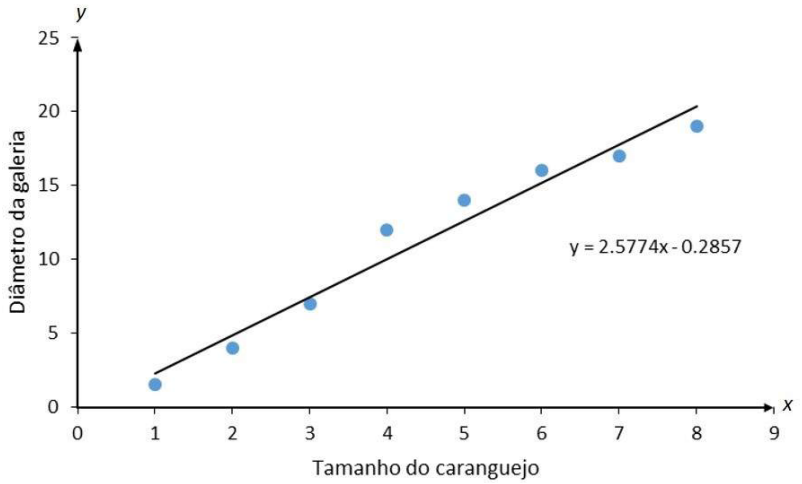
Onde:

- Y é a variável resposta.
- X a variável preditora.
- β_0 representa o intercepto ou coeficiente linear.
- β_1 corresponde ao coeficiente angular (ou inclinação da reta).

Os coeficientes são calculados pelo método dos mínimos quadrados, porém não entraremos nos detalhes desses cálculos. Na maior parte dos programas estatísticos, e até mesmo no Excel, é possível realizar uma análise de regressão. O nosso foco será na interpretação dos resultados.

Acompanhe o exemplo a seguir:

Imagine que estamos estudando o mangue da UC M e buscamos analisar a relação entre o tamanho do corpo do caranguejo (variável independente) e o diâmetro da sua galeria (variável dependente). Com os dados coletados para essas variáveis, realizamos uma análise de regressão simples e construímos o seguinte gráfico:



Tamanho do caranguejo em relação ao diâmetro da galeria na UC M.

A equação da reta, resultado da análise de regressão com os dados coletados, é Y = 2,5774X – 0,2857. Nesse exemplo, o intercepto (β_0) é -0,2857, que corresponde ao ponto no qual a reta toca o eixo y quando X = 0. A inclinação da reta (β_1) é 2,5774 e indica o número de unidades que vamos aumentar ou diminuir em nossa variável Y para cada unidade de X.

IMPORTANTE

Ou seja, para cada aumento no tamanho do caranguejo, o diâmetro da galeria aumenta 2,57 vezes.

Além disso, o coeficiente angular (β_1) também descreve a direção da relação entre as variáveis. Quando $\beta_1 > 0$, como é o nosso caso, um aumento ou uma diminuição em uma variável provocará na outra variável uma mudança no mesmo sentido.

Já se $\beta_1 < 0$, existe uma relação negativa entre as variáveis, ou seja, o aumento em uma variável provocará uma diminuição na outra e vice-versa.

Se o valor da inclinação for zero, a reta será uma linha horizontal e podemos afirmar que não há relação entre as variáveis.

Quando tratamos com dados biológicos, raramente a reta da regressão estará perfeitamente sobre todos os dados. Isso porque a reta representa uma média dos valores esperados de Y para cada valor de X. Como já sabemos, pode existir uma variabilidade nos dados que não é captada pela média.

DESTAQUE

Essa variabilidade associada a cada um dos valores médios de Y estimados é o que chamamos de erro ou resíduo (ε).

Geralmente, a capacidade explicativa do modelo é interpretada pelo coeficiente de determinação (R^2).

DESTAQUE

O R^2 representa a porcentagem de variação que foi explicada pelo modelo e varia de 0 a 1. Quanto maior o R^2, maior o poder de explicação do modelo; quanto menor o R^2, menor será a relação entre as variáveis estudadas.

No nosso exemplo, o R^2 foi 0,96, ou seja, a regressão explica 96% da variação do diâmetro da galeria.

O R^2 é tão importante na interpretação de um modelo quanto a inclinação da reta e o valor de P. Este último é encontrado pelo teste de hipóteses que pode ser calculado por um teste t ou um teste F. Em ambos os casos, a hipótese nula é que a inclinação da reta (β_1) populacional é igual a zero, indicando que não há relação entre as variáveis.

Lembre-se que estamos tratando da estatística inferencial. Os valores da reta foram estimados a partir de dados de uma amostra da população. Por isso, usamos um teste de hipótese para determinar se as relações que encontramos são devido ao acaso ou realmente ocorrem na população.

No nosso exemplo, o valor de P foi 0,14 x 10^{-4}. Assim, rejeitamos a hipótese nula, pois, considerando um nível de significância de 0,05, temos que P < 0,05, ou seja, existe relação significativa entre as variáveis. Podemos, então, concluir que o diâmetro da galeria depende do tamanho do caranguejo.

Qualquer programa estatístico que você utilizar fornecerá os mesmos outputs para a análise de regressão, que são: a inclinação da reta (*slope*), o intercepto, o erro associado a ambos os coeficientes, o R^2 e o valor de P.

2.3 Pressupostos e diagnósticos da regressão linear

Agora que já entendemos os parâmetros envolvidos na regressão linear simples, é hora de conversarmos sobre algumas boas práticas que ajudam na análise do modelo.

DESTAQUE

Todo modelo tem pressupostos que não devem ser ignorados. Se não os considerarmos, os resultados do modelo podem não ser válidos.

No caso dos modelos lineares, além dos dados terem uma distribuição normal, existe a exigência da homocedasticidade da variância e normalidade dos resíduos (média = 0).

Os resíduos são a diferença entre os valores observados e os preditos em um modelo. Acompanhe esse conceito no seguinte gráfico:

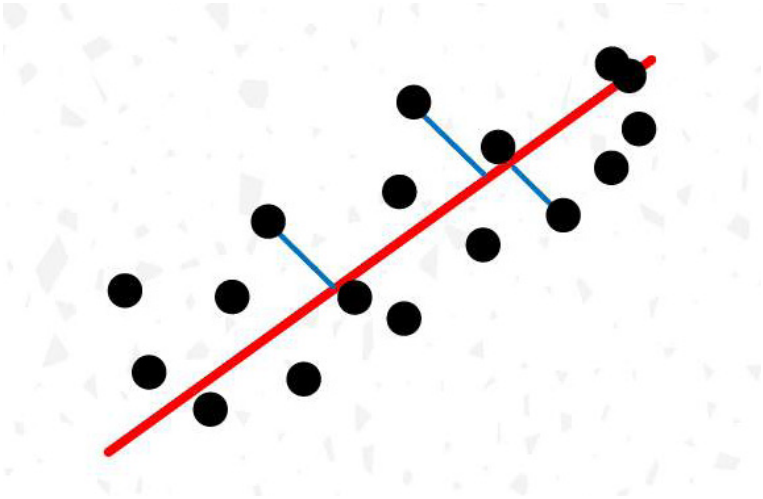


Gráfico dos valores observados (pontos pretos) e valores preditos (linha vermelha).

A linha vermelha representa os valores preditos e os pontos pretos são os valores observados. A distância entre eles (linha azul) representa o resíduo.

Os modelos lineares, tais como correlações, regressões e Anovas, assumem o pressuposto de normalidade dos resíduos e homocedasticidade da variância. Portanto, os resíduos devem ter uma distribuição normal, assim como a própria variável resposta. Isso significa que as distâncias entre observação e predição devem agrupar-se em torno de uma tendência central.

DESTAQUE

A homocedasticidade da variância significa que os resíduos do modelo devem ter variância constante ou igual em relação à variável preditora, que pode ser contínua ou categórica. No caso da variável categórica, assume-se que a variável resposta deve ter variância constante entre os grupos da preditora.

Se plotarmos os resíduos em relação ao valor esperado de Y em um gráfico, espera-se que ele tenha a seguinte aparência, caso o modelo linear se ajuste aos dados:

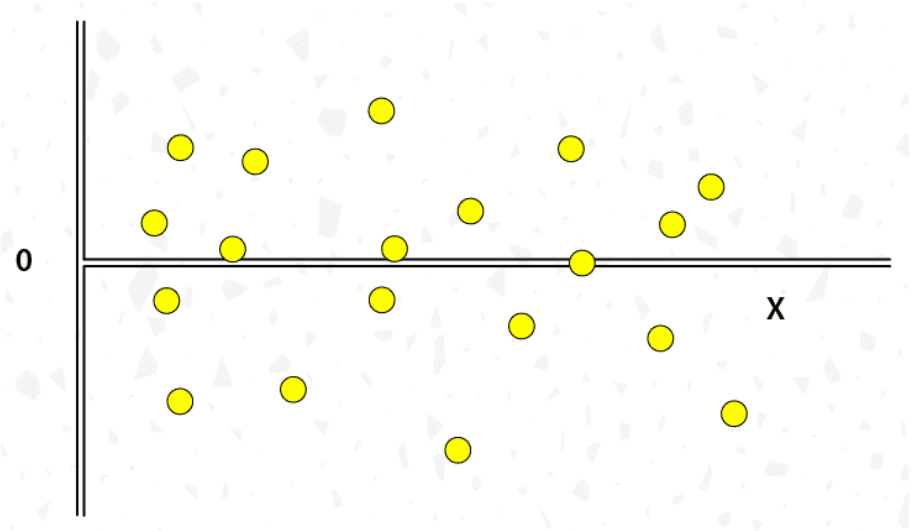


Gráfico dos resíduos com distribuição homocedástica.

Observe que os pontos estão bem distribuídos de ambos os lados do eixo x, não formando um padrão aparente, o que indica a homocedasticidade dos resíduos. A averiguação dos resíduos é feita basicamente por meio de uma inspeção visual, e os programas estatísticos fornecem esses gráficos para avaliação.

Os testes que possuem os pressupostos de normalidade dos resíduos e homocedasticidade da variância são bastante robustos. Assim, mesmo que uma variável não passe no teste de normalidade, o modelo gerado ainda é válido.

2.3 Modelos lineares generalizados (GLMs)

Conforme estudado, para construirmos modelos lineares, é necessário que os dados tenham os seguintes requisitos: distribuição normal, independência dos dados (quando os dados coletados em uma unidade amostral não interferem nos dados coletados da outra unidade) e homocedasticidade da variância. No entanto, muitas vezes, os dados coletados não apresentam todos esses requisitos. Assim, quando os erros não têm distribuição normal e/ou a variância não é constante, aplicamos os modelos lineares generalizados (em inglês General Linear Model - GLM). Vamos conhecê-los brevemente.

Propostos em 1972, os GLMs são uma expansão do espectro de atuação dos modelos lineares clássicos. Eles buscam unificar os modelos estatísticos lineares, de modo que regressões, Anova, covariância e outros são considerados casos particulares de GLMs.

Com a utilização dos GLMs, evita-se a realização de transformações nos dados para atender aos pressupostos clássicos paramétricos ou a aplicação de análises não paramétricas.

Os modelos lineares generalizados consistem na utilização de uma função, conhecida como função de ligação, para linearizar a relação entre as variáveis preditora e resposta. Para cada tipo de variável resposta, espera-se uma determinada distribuição dos erros e, para cada distribuição, existe uma função de ligação padrão sugerida, também conhecida como função canônica.

Distribuição dos erros	Função de ligação
Normal	Identidade/Log
Poisson	Log
Gama	Inversa
Binomial	Logit

Distribuições dos erros e funções de ligação padrão.

Com base na natureza da variável resposta, é possível encontrar a função de ligação. Vamos conhecer algumas dessas distribuições:

- **Distribuições gama e exponencial**
  Nessas distribuições, os dados ficam agrupados em torno de zero ou de valores muito pequenos. Isso é comum quando estamos lidando com dados de distribuição espacial de grandes predadores, como a onça pintada, uma vez que esses grupos ocorrem em baixas densidades e precisam de uma ampla área de caça e, normalmente, percorrem grandes distâncias.

- **Distribuição de Poisson**
  É uma distribuição para dados de contagem. Essa distribuição refere-se ao número de vezes que determinado evento ocorre por unidade de tempo, medida ou contagem. Exemplos: número de frutas por planta, número de ovos por fêmea de peixe, número de larvas por ponto de amostragem, número de ocorrência das espécies em determinada unidade de área.

- **Distribuição binomial ou logística**
  É uma distribuição discreta que admite dois resultados, como 0 ou 1, sim ou não, verdadeiro ou falso, presença ou ausência. Um bom exemplo dessa aplicação é quando queremos modelar os dados de presença e ausência de uma espécie ao longo do gradiente de uma variável ambiental.

Conforme Turkman e Silva (2000), a modelagem dos dados por um GLM tem três etapas: a formulação do modelo, os ajustes e a seleção e validação.

Etapas da modelagem de dados GLM

- **Formulação do modelo**
  Na etapa de formulação do modelo, são escolhidas as variáveis resposta e preditoras e identifica-se a distribuição dos erros e a função de ligação.

- **Ajustes**
  Na segunda etapa, estimam-se os coeficientes e os parâmetros que irão indicar a adequabilidade e a capacidade explicativa do modelo. Os programas estatísticos entram nesta etapa. Após a adição das variáveis e da função de ligação, a análise é realizada e produz como resultado uma tabela semelhante à da Anova e da regressão. Essa tabela tem as estimativas da inclinação da reta (slope) e dos demais coeficientes do modelo, bem como o valor do teste para avaliar a significância da estimativa dos coeficientes, o valor de P associado a esses testes e os parâmetros que indicam o quanto da variação é explicada pelo modelo.

- **Seleção e validação**
  Por fim, com todos os dados em mãos, é realizada a seleção e a validação do modelo que melhor explica os dados, comparando os modelos entre si e levando em conta os critérios de seleção escolhidos.

Referências

Unidade 1

CALLEGARI-JACQUES, S. M. **Bioestatística**: princípios e aplicações. Porto Alegre: Artmed, 2003.

VAN EMDEN, H. **Statistics for terrified biologists.** Washington: Blackwell Publishing, 2008.

Unidade 2

GOTELLI, N. J.; ELLISON, A. M. **A Primer of Ecological Statistics**. Cary, NC: Sinauer Associates, 2011.

TURKMAN, M. A. A.; SILVA, G. L. **Modelos lineares generalizados:** da teoria à prática. Lisboa: SPE, 2000.