Inferência Estatística

Modelos Lineares Clássicos

Fabio Cop (fabiocopf@gmail.com)

Última atualização em 15 de julho de 2021

Nos exercícios abaixo, escreva explicitamente quais são as hipóteses (H_0 e H_a) em teste o qual o nível de significância α escolhido. Não de esqueça de verificar os pressupostos dos modelos.

As tabelas apresentadas no documento devem ser criadas por você no formato apropriado. Outros exercícios pedem para que você importe um conjunto de dados que pode ser acessado em:

https://github.com/FCopf/probest-exercicios/tree/master/datasets.

Caso o nome das colunas seja muito longo, é interessante renomeá-los mantendo nomes curtos. É interessante também evitar o uso de *espaços*, *acentos* e *cedilhas* nos nomes das colunas.

ANOVA de um fator

1. Um pesquisador estava testando se 4 tipos de dietas tinham efeitos distintos no ganho em peso de coelhos. Deste modo, foram selecionados 20 coelhos aleatoriamente, separando-os em 4 grupos que receberam diferentes rações. O ganho em peso foi medido após duas semanas.

A	B	C	D
9.9	5.7	2.8	9.1
10.3	3.5	8.4	3.3
8.5	4.7	4.9	8.1
7.0	5.3	5.5	8.9
5.1	7.7	7.6	8.8

Existem evidências de que as rações têm efeitos diferentes no ganho em peso dos animais? Se sim, quais tratamentos tiveram os maiores ganhos?

2. O dados abaixo apresentam a quantidade (kg) de alimento consumido por cervos adultos em diferentes épocas do ano (Zar, 2010). O consumo médio entre os meses é diferente? Descreva os padrões observados e suas conclusões com base no resultado teste estatístico em gráficos que justificam sua resposta.

Fevereiro	Maio	Agosto	Novembro
4.7	4.6	4.8	4.9
4.9	4.4	4.7	5.2
5.0	4.3	4.6	5.4
4.8	4.4	4.4	5.1
4.7	4.1	4.7	5.6
NA	4.2	4.8	NA

3. Neste exemplo, é descrito um experimento inteiramente aleatorizado com o objetivo de comparar a produtividade de quatro variedades de milhos. Os resultados do experimento estão na tabela abaixo, onde cada medida representa representa a produção de milho em $kg/100m^2$ obtida em cada parcela:

A	B	C	D
25	31	22	33
26	25	26	29
20	28	28	31
23	27	25	34
21	24	29	28

a. Teste a hipótese de que os valores médios diferem entre as variedades (A, B, C e D). Em caso de rejeição de H_0, faça o teste *a posteriori* adequado e explique os resultados.

O experimento e a ANOVA são descritos no livro *Estatística Experimental* (Sônia Vieira, 1999).

4. O arquivo `Diversidade_Peixes.csv` mostra a diversidade de peixes em riachos amostrados em quatro tipos de entorno (Mata Preservada, Plantio de Eucalipto, Cana e Pastagem).

a. Represente em um boxplot, a diversidade em função do tipo de entorno. Monte uma tabela com um resumo descritivo para cada tipo de entorno utilizando a função `summary` do R. Compare os valores deste resumo com o boxplot e responda o que representam as linhas centrais, os limites das caixas e os limites das linhas em cada grupo?

b. Teste a hipótese de que a diversidade média seja diferente para ao menos um tipo de entorno.

c. Verifique os pressupostos da ANOVA (normalidade, homogeneidade de variâncias). Mostre o gráfico de resíduos. O que você conclui da avaliação dos pressupostos desta ANOVA?

d. Caso você rejeite H_0 e aceite os pressupostos, faça o teste a posteriori de Tukey para verificar quais grupos são diferentes. Descreva verbalmente suas conclusões.

ANOVA em blocos e fatorial

1. Os dados abaixo apresentam a concentração do aminoácido alanina ($mg/100ml$) na hemolinfa de machos e fêmeas de 4 espécies de diplópodes (Zar, 2010).

Sexo	Espécie 1	Espécie 2	Espécie 3
Macho	21.5	14.5	16.0
Macho	19.6	17.4	20.3

Sexo	Espécie 1	Espécie 2	Espécie 3
Macho	20.9	15.0	18.5
Macho	22.8	17.8	19.3
Fêmea	14.8	12.1	14.4
Fêmea	15.6	11.4	14.7
Fêmea	13.5	12.7	13.8
Fêmea	16.4	14.5	12.0

a. Existe diferença na concentração média de alanina entre os sexos. Se sim, explique esta diferença. Faça o mesmo para as espécies.

b. Existe interação entre os fatores Sexo e Espécie?

c. Se houver diferença entre as espécies, faça o teste de Tukey para indicar como de dão estas diferenças.

2. Para o arquivo `Custaceos.csv`, faça uma ANOVA de dois fatores para testar as hipóteses de que o comprimento da carapaça varia em função do sexo e do ambiente. Não de esqueça de testar os pressupostos da ANOVA.

a. Represente estas relações por meio de um único boxplot. Crie um gráfico de interação. Você consegue entender a relação entre este gráfico e o boxplot?

b. Descreva verbalmente o resultado dos efeitos principais e da interação nesta ANOVA.

3. Utilizando o arquivo `NPK.csv`, compare o efeito dos níveis de adubo (N, P e K) sobre o peso de plantas. Verifique antes se existem observações para todas as combinações dos níveis de tratamento. Como realizar a análise caso não haja? *Observe que as colunas N, P e K então configuradas como variáveis numéricas. Transforme-as em fatores para realizar a ANOVA.*

4. Para o arquivo `Moluscos.csv`, teste os efeitos da Espécie, Sexo e Temperatura sobre o metabolismo dos animais. Descreva verbalmente os resultados utilizando gráficos e tabelas apropriados.

Regressão Linear Simples

1. Importe o arquivo `Pressao_pulso.csv`. Faça um gráfico de dispersão entre pressão diastólica e idade. Utilize um modelo de regressão linear para testar a hipótese de que a pressão diastólica varia em função da idade. Responda:

a. Qual foi sua conclusão? Explique.

b. Quais foram os coeficientes da regressão. Interprete o valor numérico associado ao coeficiente de inclinação da reta.

c. Interprete o valor do coeficiente de determinação R^2.

d. Teste os pressupostos do modelo e responda: a regressão linear foi um modelo apropriado?

2. Importe o arquivo `Mananciais_resumido.csv` e faça um gráfico de dispersão entre *Volume_reservatorio* (variável resposta - Y) e *Chuva* (variável preditora - X).

2.1. Para o **Sistema Guarapiranga** responda.

a. O gráfico de dispersão sugere uma relação entre Y e X do tipo:

- Negativa;
- Nula;
- Positiva.

b. Ajuste uma regressão linear e calcule os coeficientes estimados da regressão $\hat{\beta}_0$ e $\hat{\beta}_1$ e a variância do resíduo s^2.

c. Com nível de significância $\alpha = 0,05$, teste as hipóteses:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

d. Para uma quantidade de Chuva de 110 mm anuais, calcule os intervalos de *confiância* e de *predição* para o volume do reservatório.

2.2. Para o **Sistema Rio Grande** responda.

a. O gráfico de dispersão sugere uma relação entre Y e X do tipo:

- Negativa;
- Nula;
- Positiva.

b. Ajuste uma regressão linear e calcule os coeficientes estimados da regressão $\hat{\beta}_0$ e $\hat{\beta}_1$ e a variância do resíduo s^2.

c. Com nível de significância $\alpha = 0,05$, teste as hipóteses:

$H_0 : \beta_1 = 0$

$H_a : \beta_1 \neq 0$

d. Para uma quantidade de Chuva de 110 mm anuais, calcule os intervalos de *confiância* e de *predição* para o volume do reservatório.

e. Calcule o coeficiente de determinação R^2.

3. Importe o conjunto de dados `galapagos_alr3.csv`. Faça um gráfico de dispersão entre riqueza de espécies (NS) e área das ilhas ($Area$). Note que a relação é não-linear. Em geral, a relação riqueza de espécies vs área pode ser descrita por:

$$NS = \beta_0 \times Area^{\beta_1}$$

a. Como você utilizaria um modelo de regressão linear simples para testar se a riqueza de espécies aumenta com a área das ilhas?

b. Após responder ao item anterior, aplique uma regressão linear, teste os pressupostos do modelo e apresente os resultados.

4. A relação entre o peso e o comprimento em diversos animais é não-linear. Como muitas outras relações biológicas, esta relação pode ser ajustada por *funções potência*. Veja por exemplo (Xiao et al. 2011). Funções potência têm o formato:

$$Y = \beta_0 \times X^{\beta_1}$$

Esta relação pode ser denominada também de *relação alométrica*, em que β_1 é denominado de *coeficiente de alometria*. O coeficiente é dito **alométrico negativo** (< 3), **isométrico** ($= 3$) ou **alométrico positivo** (> 3). Analise a relação entre peso e comprimento do lambari *Hollandichthys multifasciatus* (arquivo `Hollandichthys_multifasciatus.csv`) e teste a hipótese de que o coeficiente de alometria seja isométrico. Apresente o gráfico de dispersão.

Análise de Covariância (ANCOVA) e extensões

1. A partir do arquivo `Remoção_fruto.csv`, faça o teste t comparando o efeito do tratamento (com e sem remoção) sobre o peso médio dos frutos. Qual foi sua conclusão? Em seguida faça uma análise de covariância para testar o efeito do tratamento sobre o peso dos frutos porém, controlando para o efeito da covariável *peso da raiz*. Qual sua conclusão agora? Como você explica as diferenças entre os resultados do teste t e da ANCOVA?

2. Utilizando o conjunto de dados `Pressao_pulso.csv`, faça uma ANCOVA para testar se homens e mulheres têm diferentes pressões Diastólicas, controlando o efeito da idade. Quais são as médias ajustadas de homens e mulheres?

3. Importe o arquivo `Acaros.csv`. Este arquivo mostra a riqueza de espécies de ácaros (S) para 70 parcelas. Em cada parcela, a porção coberta por vegetação foi classificada em `None`, `Few` e `Many`. Foram medidas também concentração de água no solo ($ConcAgua$) e densidade do substrato ($SubsDens$).

a. Faça gráficos de dispersão entre $S \sim ConcAgua$, entre $S \sim SubsDens$, faça as regressões lineares e descreva os resultados.

b. Faça agora um modelo de regressão múltipla para modelar a riqueza de ácaros em função da concentração de água e da densidade do substrato.

   i. Interprete os coeficientes de inclinação da reta para os efeitos da concentração de água e da densidade do substrato no modelo de regressão múltipla.

   ii. Como você explica as diferenças entre este modelo e os resultados do item a.?

4. Calcule a riqueza de espécies de peixes a partir do arquivo `doubs_species.csv`. Faça um modelo de regressão múltipla entre a riqueza de espécies em função da distância da cabeceira (distance from source = dfs), altitude (alt), pH e oxigênio dissolvido (oxy). As variáveis ambientais podem ser encontradas no arquivo `doubs_environment.csv`. *Para obter o número de espécies, pense em como utilizar a função* ***`apply`****, ou use a função* ***`specnumber`*** *do pacote vegan.*

a. Mostre que as variáveis distância da cabeceira e altitude são altamente colineares.

b. Qual o efeito desta colinearidade no modelo de regressão múltipla ao inserir as duas variáveis?

c. Diante disto, monte um modelo mais apropriado e interprete os valores dos coeficientes de regressão, do R^2 e do R^2 ajustado.

5. Utilizando o arquivo `Peso_crianca.csv`, rode um modelo para testar qual(is) fatores (Idade da mae, tempo de gestação, hábito de fumar) afetam o peso de recém nascidos. Interprete os resultados e apresente gráficos e tabelas apropriados.

6. Com os arquivos `doubs_species.csv` e `doubs_environment.csv` rode um modelo de regressão múltipla para testar quais variáveis afetam a diversidade de peixes. Teste o pressuposto de colinearidade entre as variáveis preditoras antes de decidir quais devem ou não ser incorporadas ao modelo. A partir desta seleção inicial, faça uma regressão *stepwise.* Quais variáveis se mantiveram no modelo final? Apresente os resultados em uma tabela.

7. Faça o mesmo para o arquivo `Biomassa_vegetal.csv` e monte um modelo para explicar quais fatores (% de área exposta, tipo de vegetação, luminosidade, temperatura e umidade) afetam a biomassa vegetal. Houve variáveis colineares? Se sim, como você decidiu qual excluir? Após rodar uma regressão *stepwise* (sem as variáveis colineares), quais variáveis se mantiveram no modelo final e quais foram significativas?