Base

Testes Clássicos

Os testes clássicos frequentistas podem ser divididos em dois grandes grupos: paramétrico e não paramétrico[1)]. Os paramétricos estão baseados em uma função probabilística teórica. As funções probabilísticas são definidas por parâmetros, a Gaussiana, por exemplo, é definida pelos parâmetros média (μ) e desvio padrão (σ). Os testes não paramétricos independem de uma distribuição teórica e por isso não há necessidade de estimar parâmetros. Além disso, não há pressupostos associados à variabilidade nos dados. Um conjunto de testes importantes nesse contexto são os chamados testes de postos (*rank*), nos quais os dados são ordenados e as posições de diferentes tratamentos são testadas para ver se há correlação entre as posições. Não iremos tratar os testes não paramétricos aqui, mas lembrem-se que são uma alternativa para teste de hipóteses clássicos, quando os pressupostos desses não são atendidos.

Os testes paramétricos clássicos foram desenvolvidos independentemente para a solução de problemas distintos. Por isso, não há uma unificação analítica, que só aconteceu posteriormente com a integração dos modelos lineares, como veremos nas próximas aulas. A aplicação é definida basicamente pela natureza das variáveis resposta (dependente) e preditora (independente) e pela hipótese estatística subjacente.

Principais testes clássicos frequentistas

A tabela abaixo apresenta os principais testes clássicos frequentistas e sua aplicação com relação às variáveis preditoras e resposta, e à hipótese estatística subjacente.

Tipo de Variável		Estatística Clássica	
Resposta	Preditora	Teste	Hipótese
Categórica	Categórica	Qui-quadrado	independência
Contínua	Categórica (2 niveis)	Teste t	$\mu_1 = \mu_2$
Contínua	Categórica	Anova	$\mu_1 = \mu_2 = \mu_n$
Contínua	1 Contínua	Regressão	$\beta_1 = 0$
Contínua	>1 Contínua	Reg. múltipla	$\beta_1 = 0 ; \beta_n = 0$
Contínua	Cont + Categ	Ancova	$\beta_1 = \beta_2 ; \alpha_1 = \alpha 2$
Proporção	Contínua	Reg. Logística	$logit(\beta_1) = 1$

Anova

Na aula sobre teste de hipótese utilizamos técnicas de Monte Carlo para testar a hipótese de que duas médias são distintas, ou que uma é maior/menor que outra, tanto no exemplo do tutorial Tutorial Árvores do Mangue, quanto no exercício altura dos alunos. Em ambos os casos estávamos comparando médias de dois grupos distintos, por exemplo dois tipos de solos no mangue ou gênero dos alunos. O nosso procedimento simulou o teste frequentista ***t*** de Student, que utiliza uma distribuição estatística ***t*** [2] e dessa forma podemos comparar o valor observado em nossos dados com a distribuição estatística e testar a hipótese das médias serem diferentes, sem a necessidade de simular o cenário nulo, como apresentamos na aula teste de hipótese.

Caso não esteja confortável com o procedimento de simulação do cenário nulo e consequente obtenção do ***p-valor*** , refaça o tutorial teste de hipótese. No procedimento apresentado está a lógica básica por trás de todos os testes de hipótese clássicos.

A *Análise de Variância* (**ANOVA**) é uma generalização do ***teste-t***, desenvolvida por Ronald Fisher 100 anos atrás (1918). Apesar de idoso, é um teste muito popular, talvez o mais utilizado em ciências naturais. A hipótese subjacente da ANOVA é de diferença entre as média de 2 ou mais grupos. O procedimento para o cálculo da estatística da ANOVA, chamada de ***F***, está associado à partição da variância dos dados, por isso o nome. Uma maneira clássica de apresentar o resultado do teste de **ANOVA** é a a chamada **tabela de ANOVA**. Essa tabela será utilizada para avaliarmos outros modelos também, por isso é importante entender o que ela nos diz.

Tabela de ANOVA

Vamos montar nossa tabela de ANOVA a partir dos dados de colheita de um cultivar em diferentes tipos de solos, apresentado no livro de Robert Crawley, The R Book, como segue abaixo:

Tradução livre da descrição do livro "*The R Book*" (Crawley, 2007)

Robert Crawley

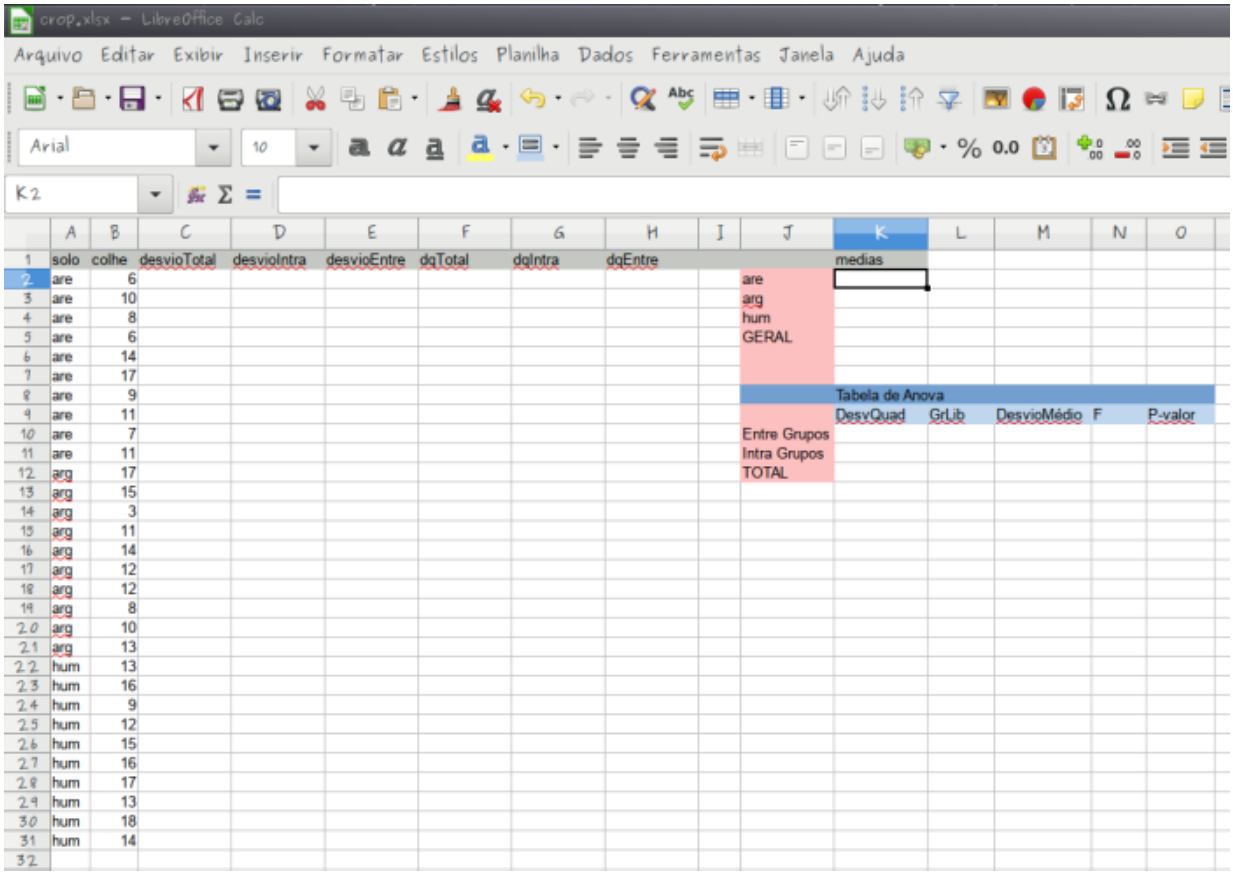
"... a melhor forma de entender o que está acontecendo é trabalharmos um exemplo. Temos um experimento em que a produção agrícola por unidade de área é medida em 10 campos de cultivo selecionados aleatoriamente de cada um de três tipos diferentes de solo. Todos os campos foram semeados com a mesma variedade de semente e manejados com

as mesmas técnicas (fertilizantes, controle de pragas). O objetivo é verificar se o tipo de solo afeta significativamente o rendimento de culturas, e caso afete, quanto.”[3)]

- baixe o arquivo

  crop.xlsx

  ;
- abra em uma planilha eletrônica;



- calcule a média geral e de cada grupo e guarde nas células correspondentes à direita;
- calcule o quanto cada observação desvia da média geral e guarde na coluna “desvioTotal”;
- faça o mesmo para a observação e a média do seu grupo e guarde na coluna “desvioIntra”;
- para cada observação, calcule o quanto a média do seu grupo desvia da média geral e guarde na coluna “desvioEntre”;
- para cada um dos desvios calculados anteriormente, vamos elevar ao quadrado e guardar na coluna correspondente de desvio quadrático (***dq****)

- O que representam as somas das colunas (***dq****)?
- Compare os valores obtidos com as formulas abaixo:

Desvios quadráticos total

$$ SS_{total} = \sum_{i=1}^k\sum_{j=1}^n (y_{ij} - \bar{\bar{y}})^2 $$

Desvios quadráticos internos ao grupo

$$ SS_{in} = \sum_{i=1}^k\sum_{j=1}^n (y_{i,j} - \bar{y}_{i})^2 $$

Desvio quadrático entre os grupos

$$ SS_{en} = \sum_{i=1}^k\sum_{j=1}^n (\bar{y}_{i} - \bar{\bar{y}})^2 $$

- complete a tabela;
- interprete o resultado;
- repita o teste no Rcmdr e compare os resultados;
- faça um gráfico que represente bem os dados;

Como calcular o p-valor a partir do F

- A função *DIST.F* no Excel ou LibreOffice calcula o p-valor a partir da estatística **F** e graus de liberdade;
- usualmente a função recebe o valor de **F**, seguido dos graus de liberdade entre e intra grupos;
- o resultado da função *DIST.F* é a probabilidade cumulativa;
- o p-valor é igual a 1 menos essa probabilidade.

ANOVA no Rcmdr

- importe os dados apenas com as colunas de dados brutos;
- o menu *Estatísticas* está separado em tipos de estatísticas e qual o parâmetro associado ao teste de hipótese estatístico;
- o nosso teste é sobre médias, portanto no sub-menu *Médias*;
- nele há a opção *ANOVA para um fator (one way)...*
- o resultado aparecerá na janela *Output*.

Entregar aos monitores até o início da próxima aula:

- a tabela de ANOVA completa gerada na planilha;
- a tabela de ANOVA resultante do teste no Rcmdr;
- um gráfico para representar os resultados;
- a interpretação dos resultados.

Regressão Linear Simples

Análise de Resíduos de Regressão Linear

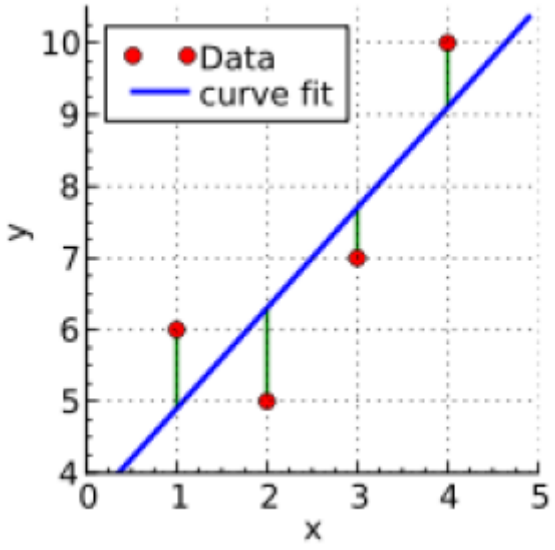
Quando realizamos uma análise mais aprofundada sobre a relação entre duas variáveis numéricas contínuas podemos ajustar uma reta que represente o melhor ajuste entre os dados e que possa nos ajudar a prever valores da variável resposta (eixo Y) a partir de valores da variável preditora (eixo X). Se o ajuste é uma reta, esse tipo de análise é chamado de **Análise de Regressão Linear**. A explicação detalhada sobre como funciona essa análise foi apresentada na aula sobre Análise de Regressão Linear. Alguns aspectos importantes que precisam ser lembrados para entendermos esse tutorial:

- A reta ajustada (também denominada “linha de regressão”) passa obrigatoriamente pelo ponto que representa a média da variável Y e a média da variável X.
- Os pontos das observações estarão distribuídos em torno dessa reta.
- A distância vertical (projetada no eixo Y) de cada ponto até a reta é chamada de **resíduo** ou **erro** daquele ponto.
- A linha de regressão é aquela que minimiza os resíduos (na verdade, **a soma dos quadrados dos resíduos**)

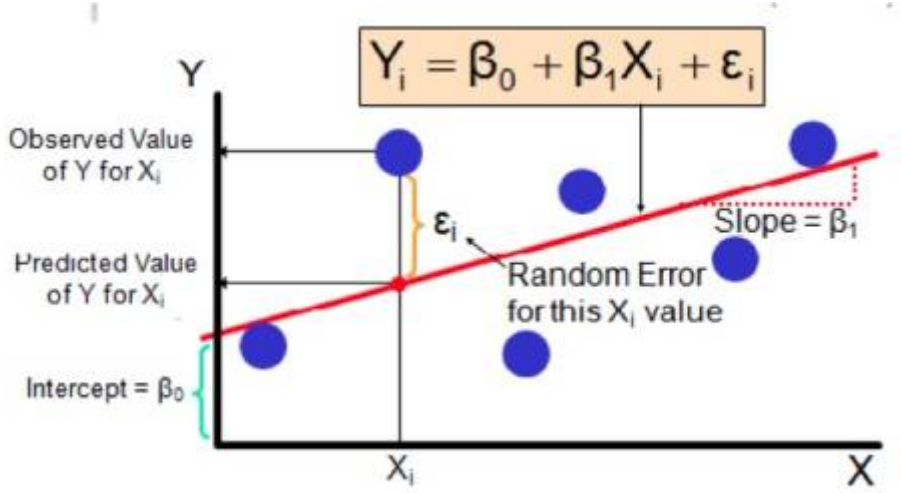
O objetivo desse tutorial é aprender a fazer uma análise de regressão linear e a avaliar a distribuição dos resíduos/erros, pois os modelos de regressão linear possuem importantes premissas relacionadas a eles.

O que são os erros/resíduos e como calcular?

Os erros/resíduos indicam o quão longe os valores de Y observados estão dos valores de Y estimados pela linha de regressão ajustada. Eles estão representados em verde na figura abaixo:



Os erros/resíduos de cada observação são calculados projetando-se no eixo Y o valor de Y observado e o valor de Y estimado (ou predito) pela reta e calculando-se a diferença entre esses dois valores.



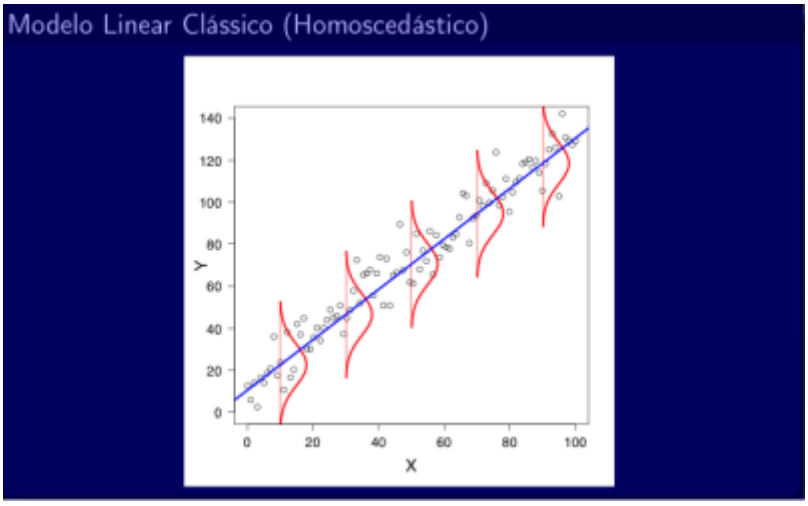
Premissas de uma Análise de Regressão Linear

- **LINEARIDADE** - Uma reta representa o melhor ajuste aos dados

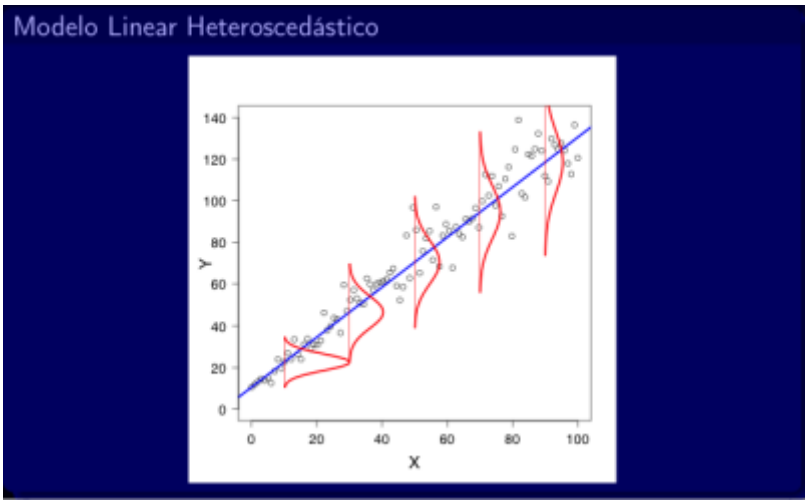
- **DISTRIBUIÇÃO NORMAL DOS ERROS/RESÍDUOS** - Para cada valor de X, os erros devem seguir uma distribuição normal. Se fossem feitas muitas réplicas para cada um dos valores de X, a distribuição dos erros referentes aos vários valores obtidos para Y nas muitas réplicas seguiria uma distribuição normal (Veja as curvas em vermelho na figura de homoscedasticidade abaixo). Porém, em geral, não são feitas réplicas e é necessário assumir que os resíduos seguem essa distribuição.

- **VARIÂNCIA DOS ERROS/RESÍDUOS CONSTANTE** - Para qualquer valor de X, a variância dos erros é a mesma. Se fossem feitas muitas réplicas para cada um dos valores de X, a distribuição dos erros referentes aos vários valores obtidos para Y nas muitas réplicas apresentaria uma mesma variância para qualquer valor de X. Porém, em geral, não são feitas réplicas e é necessário assumir que essas variâncias são iguais.

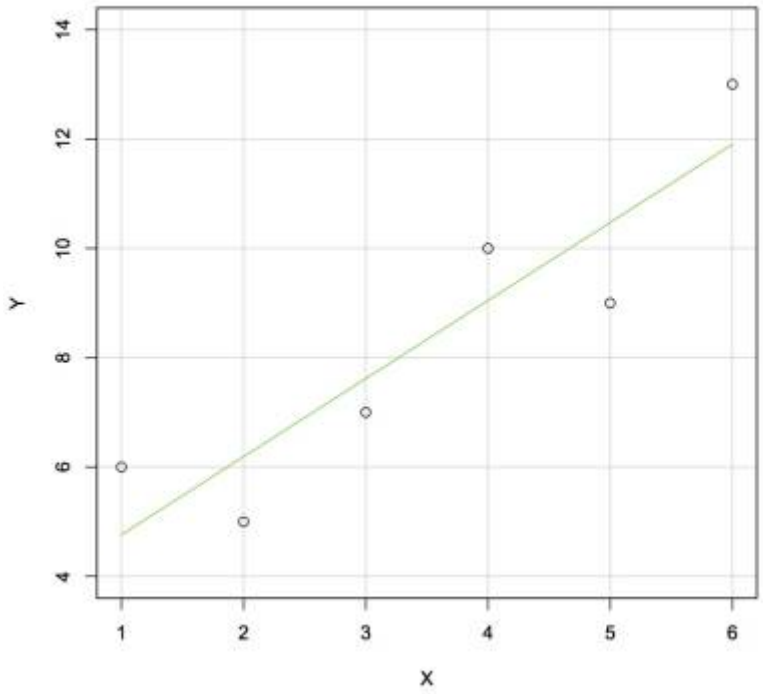
Quando essa premissa é cumprida, temos o que chamamos de **homoscedasticidade**:



Quando ela não é cumprida, observamos uma **heteroscedasticidade**. Note na figura abaixo que para valores pequenos de X a variância é menor (distribuição estreita) e que para valores maiores de X temos uma variância grande (distribuição larga):



Agora vamos estimar os valores dos resíduos para um exemplo hipotético abaixo:



Faça uma tabela como essa e anote os valores aproximados que você consegue obter por esse gráfico:

X	Yobservado	Yestimado	Resíduo
1			
2			
3			
4			
5			

X	Yobservado	Yestimado	Resíduo
6			

Checando as premissas

Ok, agora que você entendeu como são calculados os erros/resíduos, vamos trabalhar com conjuntos de dados maiores para podermos entender como checar as premissas da análise de regressão linear de uma maneira um pouco mais realista:

Baixe os arquivos de dados para o seu diretório:

- algas_peixes.csv
- algas_peixes2.csv
- insetos_peixes.csv
- vol_inds.csv

Descrição dos conjuntos de dados:

Atenção: esses conjuntos de dados não são reais, são simulações produzidas com o objetivo de inserir nos dados alguns padrões que frequentemente encontramos em estudos reais

Imagine que um grupo de pesquisadores vem trabalhando há muito tempo com peixes da família Rivulidae que ocorrem em lagos temporários. Esses peixes crescem e se reproduzem nesses lagos temporários durante o período de chuvas e seus ovos ficam dormentes durante o período de seca. Os pesquisadores estão interessados em compreender as relações tróficas e as possíveis limitações de espaço nas áreas de ocorrência desses peixes.

Do total de lagos temporários existentes, foram sorteados 20 lagos e na época chuvosa os seguintes dados foram coletados:

- - Biomassa de algas
- - Biomassa de insetos aquáticos
- - Volume do lago
- - Biomassa de peixes herbívoros
- - Biomassa de peixes insetívoros
- - Número de indivíduos adultos da espécie mais abundante (*Austrolebias charrua*)

- O primeiro conjunto de dados (algas_peixes.csv) foi obtido com o objetivo de analisar se a biomassa de algas existente nos lagos influencia a biomassa de peixes herbívoros e se essa relação é linear.

- O segundo conjunto de dados (algas_peixes2.csv) foi obtido com o mesmo objetivo anterior, mas as medidas foram tomadas no ano seguinte (ano 2).

- O terceiro conjunto de dados (insetos_peixes.csv) foi obtido com o objetivo de analisar se a biomassa de insetos existente nos lagos influencia a biomassa de peixes insetívoros e se

essa relação é linear.

- O quarto conjunto de dados (vol_inds.csv) foi obtido com o objetivo de analisar se o volume de água de cada lago afeta o número de indivíduos da espécie *Austrolebias charrua*, que é uma das espécies dominantes nesses lagos, e se essa relação é linear.

O primeiro passo é ajustar um modelo de regressão linear aos dados obtidos.

Inicialmente vamos trabalhar com o conjunto de dados *algas_peixes.csv*

Como saber se os erros/resíduos seguem uma distribuição normal?

Lembre dos métodos usados no tutorial de ANÁLISES EXPLORATÓRIAS DE DADOS. Você tem várias opções para avaliar visualmente a distribuição de um conjunto de valores. Basta aplicar a mesma lógica para a distribuição dos erros/resíduos.

Como saber se a variância dos erros/resíduos é constante?

Considerando que a reta obtida pelo modelo linear separa os pontos observados de Y de modo que eles fiquem distribuídos da melhor forma possível acima e abaixo da reta, teremos tanto valores positivos quanto valores negativos de resíduos para os diferentes valores de X.

Para um dado valor de X, teremos um valor de *Y_estimado* (que aparece na planilha de dados como "*fitted.RegModel.**"). Relembrando, os *Y_estimados* são os valores projetados em Y quando o valor de X cruza a reta de regressão.

Se esperamos que a variância dos resíduos seja constante ao longo dos valores de X, deveríamos também esperar que o espalhamento dos valores dos resíduos (positivos ou negativos) sejam similares para os diferentes valores de *Y_estimado*.

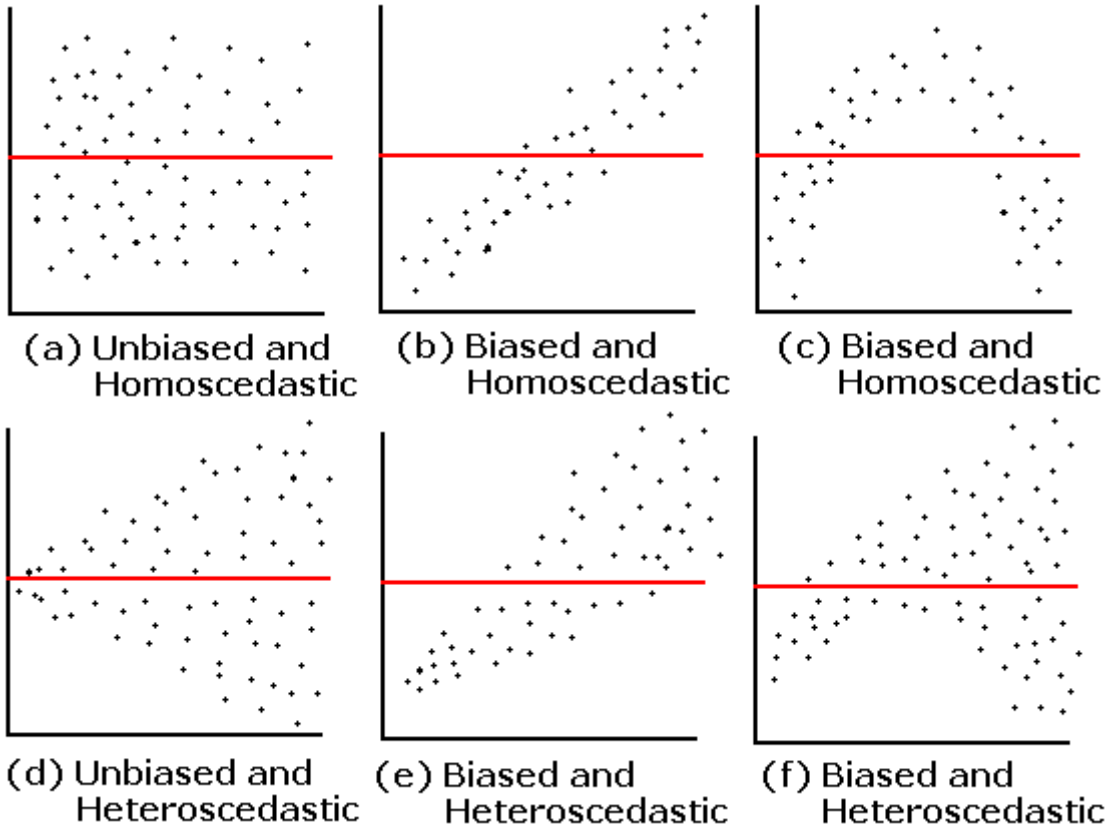
Então, podemos fazer um gráfico em que relacionamos os valores de *Y_estimado* ("*fitted.RegModel.**") e os valores dos *Resíduos* ("*residuals.RegModel.**") para cada *Y_estimado*. Com esse gráfico podemos avaliar se a distribuição dos resíduos é similar ou se há um maior ou um menor espalhamento dos valores de resíduos para alguns valores de *Y_estimado*.

residuo2

Como você interpreta esse gráfico? Você nota algum padrão na distribuição dos erros/resíduos?

Esse mesmo gráfico que é utilizado para avaliar se a variância é constante (homoscedasticidade), também pode ser utilizado para checar se existe alguma assimetria, algum viés (positivo ou negativo) ou alguma tendência de que a relação seja melhor definida por uma curva do que por uma reta.

A figura abaixo mostra vários exemplos desse gráfico entre *Resíduos* e *Y_estimado*, com ou sem homoscedasticidade e com ou sem vieses (*Biased* ou *Unbiased*):



Ao interpretar esses gráficos, lembre-se sempre que aqui não estão sendo representados os seus dados brutos, e sim os resíduos e os valores de *Y_estimado*!

Como saber se alguma observação está influenciando demais os parâmetros da regressão?

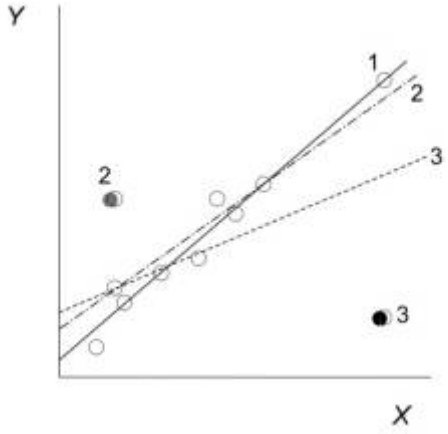
Além de testar as premissas, também é importante fazer um diagnóstico para verificar se existem *outliers* e se eles influenciam muito o resultado da análise de regressão.

Para medir a influência que uma dada observação tem sobre a inclinação da reta estimada pelo modelo de regressão linear, usamos uma medida denominada **"Distância de Cook"** que é calculada para cada observação e avalia a relação entre o erro/resíduo (***e***) e a *leverage* (***hii***) da observação. A *leverage* (que pode ser traduzida como "alavancagem") indica o quanto um dado valor de X influencia o valor de *Y_estimado*. Repare, pela equação abaixo, que quanto maior for o erro/resíduo (***e***) e a *leverage* (***hii***) de uma dada observação, maior será a distância de Cook referente a ela. Porém, se a *leverage* for alta para uma dada observação, mas o erro/resíduo for pequeno, essa

observação não terá um valor alto de Distância de Cook e, possivelmente, não terá tão grande influência sobre a inclinação da reta.

$$D_i = \frac{e_i^2}{(p+1)QME} \frac{h_{ii}}{(1-h_{ii})^2}.$$

Valores altos de Distância de Cook para uma dada observação indicam que se ela fosse retirada das análises, a inclinação da reta de regressão poderia mudar muito. Veja o exemplo abaixo, do livro de Quinn & Keough (2008), mostrando o efeito de três diferentes observações sobre a inclinação da reta. Os números das retas (1, 2 e 3) indicam como seria a reta se aquela determinada observação (1, 2 ou 3, respectivamente) fosse retirada.



Se você não entendeu essa figura, peça ajuda!

Então, podemos fazer um gráfico em que plotamos o valor dos *Resíduos* em relação aos valores de *leverage* e nesse gráfico os pontos que possuírem as maiores *leverage* e os maiores erros/resíduos (positivos ou negativos) serão as observações com maiores **Distâncias de Cook** e consequentemente, com maiores **influências** sobre os parâmetros da reta.

Devido ao tempo escasso, não vamos construir esse gráfico passo-a-passo. Vamos usar uma opção mágica que vai mostrar 4 gráficos de diagnóstico de uma só vez e incluirá esse gráfico para que você possa analisar.

O primeiro passo é ajustar um modelo de regressão linear aos dados obtidos. Para o primeiro conjunto de dados (algas_peixes.csv), nós já fizemos isso, então, vamos apenas recuperar o resumo do modelo para depois fazermos os gráficos sobre esse modelo:

Final

Entendendo essa figura:

- Os dois gráficos à esquerda relacionam os resíduos aos valores de *Y_estimado*. Dentre esses, o gráfico inferior utiliza os resíduos padronizados[4] para diminuir eventuais problemas com assimetria (*skewness*) nos dados. Em geral, basta checar um deles e já será possível identificar problemas de

heteroscedasticidade e de viés nos resíduos.

- O gráfico superior à direita é o gráfico quantil-quantil, que aprenderemos na aula de análises exploratórias de dados (AED). Ele nos ajuda a identificar se os resíduos [5)] se ajustam bem a uma distribuição normal (checagem da normalidade dos resíduos). Se os pontos desse gráfico estiverem bem próximos da linha diagonal (observem principalmente as extremidades), isso indica que os valores dos resíduos estão bem ajustados a uma distribuição normal. Se nas extremidades os pontos estiverem distantes da linha, a distribuição dos resíduos é assimétrica, apresentando caudas mais longas ou mais curtas, a depender da posição em que ocorrem esses pontos distanciados

.

- O gráfico inferior à direita é o gráfico que mostra a relação entre resíduos (padronizados) e a *leverage* das observações. É nesse gráfico que podemos também conferir a Distância de Cook. As linhas tracejadas indicam os limites para valores de distância de Cook que são considerados altos (acima de 0,5). Pontos localizados fora dessa linha tracejada são observações com alta Distância de Cook e que devem, portanto, ser analisados cuidadosamente. Repare que os pontos com as maiores Distâncias de Cook têm números que ajudam você a identificar a qual observação o ponto se refere.

Salve esse conjunto de gráficos como .pdf e identifique-o com o nome do arquivo de dados

Repita o mesmo procedimento para os outros três conjuntos de dados e avalie quais premissas estão sendo atendidas ou não para cada um.

Exercício para entregar até a próxima aula:

1) Os gráficos de diagnóstico dos outros três conjuntos de dados (algas_peixes2.csv; insetos_peixes.csv; vol_inds.csv)

2) Para cada conjunto de dados, faça sua interpretação sobre a distribuição dos resíduos, incluindo avaliação de:

- 2.1 - normalidade;
- 2.2 - homoscedasticidade;
- 2.3 - viés;
- 2.4 - influência dos pontos.

1)
Há pelo menos duas definições distintas para o termo estatística não paramétrica. A definição aqui é circunscrita aos testes frequentistas que não estimam parâmetros de uma distribuição probabilística.
2)
essa distribuição foi desenvolvida por William Gosset
3)
The best way to see what is happening is to work through a simple example. We have an experiment in which crop yields per unit area were measured from 10 randomly selected fields on each of three soil types. All fields were sown with the same variety of seed and provided with the same fertilizer and pest control inputs. The question is whether soil type significantly affects crop yield, and if so, to what extent.
4)
se tiver interesse em entender como é feita essa padronização, utilize a ajuda do Rcommander ou do R, mas não precisa fazer isso nesse momento

5)

note que ele também está usando resíduos padronizados

From:
http://labtrop.ib.usp.br/ - **Laboratório de Ecologia de Florestas Tropicais**

Permanent link:
http://labtrop.ib.usp.br/doku.php?id=cursos:planeco2019:roteiro:07-class_base

Last update: **2019/12/11 14:31**