Testes Clássicos

Os testes clássicos estatísticos estão inseridos no escopo da estatística frequentista ou inferência frequentista. Nessa abordagem a inferência é baseada na frequência ou proporção dos dados amostrados. A maior parte dos testes frequentistas clássicos foi desenvolvida independentemente para a solução de problemas distintos. Por isso, não há uma unificação analítica completa, que só aconteceu posteriormente com a integração oferecida pelos modelos lineares, como veremos nas próximas aulas. Nos testes clássicos 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	$logit(\beta_1)=1$

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

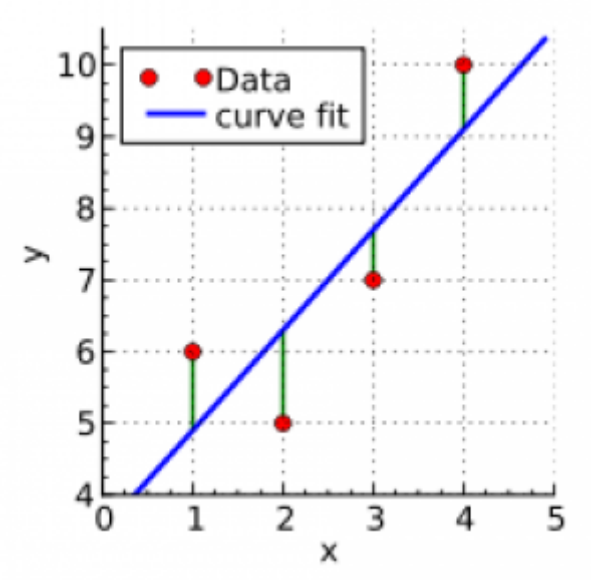
ajudar a prever valores da variável resposta (eixo Y) a partir de valores da variável preditora (eixo X). Se o ajuste é uma reta, esse tipo de análise é chamado de **Análise de Regressão Linear**. A explicação detalhada sobre como funciona essa análise foi apresentada na aula sobre Análise de Regressão 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**, **desvio** ou **erro** daquele ponto.
- A linha de regressão é aquela que minimiza os resíduos (na verdade, **a soma dos quadrados dos resíduos**)

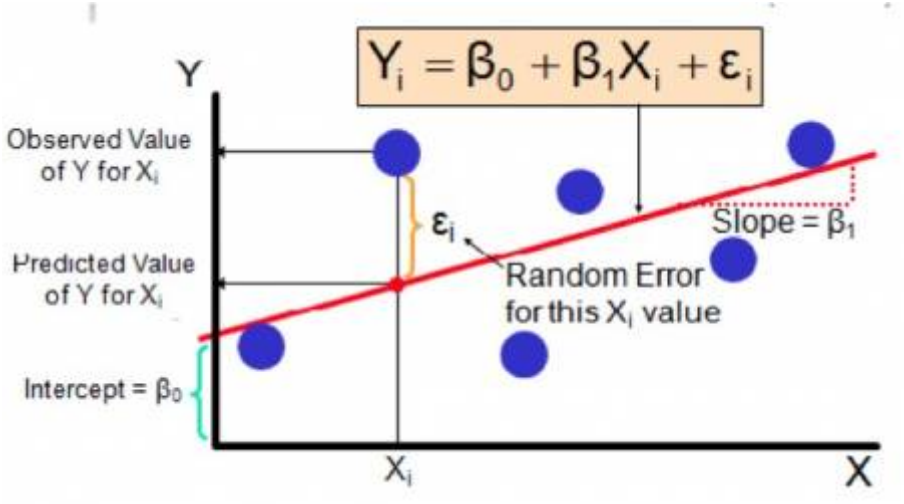
O objetivo desse tutorial é aprender a fazer uma análise de regressão linear e a avaliar a distribuição dos erros/resíduos, pois os modelos de regressão linear possuem importantes premissas relacionadas a eles.

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.



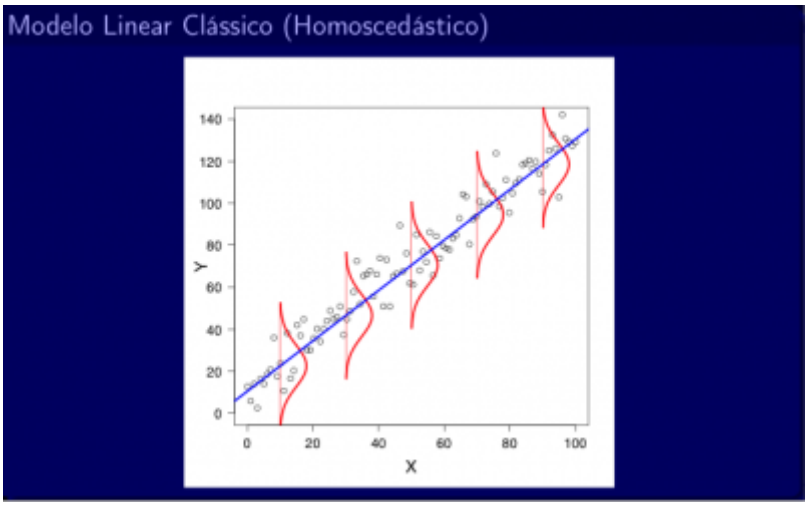
Premissas de uma Análise de Regressão Linear

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

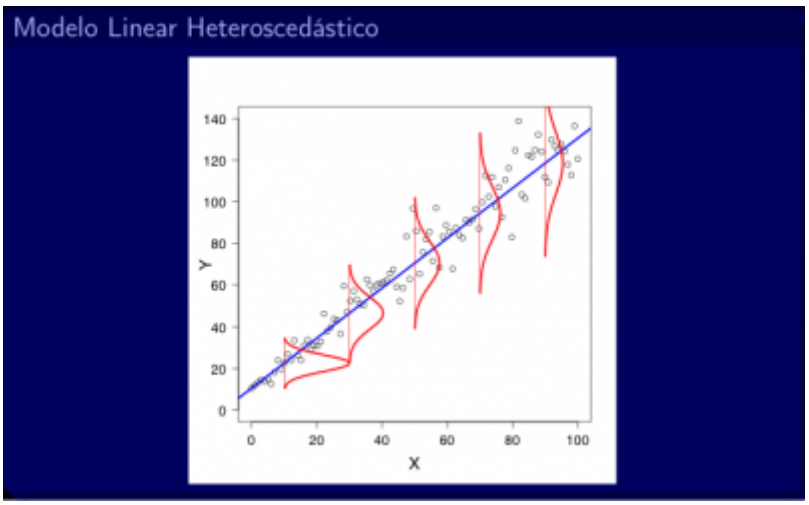
- **DISTRIBUIÇÃO NORMAL DOS ERROS/RESÍDUOS** - Para cada valor de X, os erros devem seguir uma distribuição normal. Se fossem feitas muitas réplicas para cada um dos valores de X, a distribuição dos erros referentes aos vários valores obtidos para Y nas muitas réplicas seguiria uma distribuição normal (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.

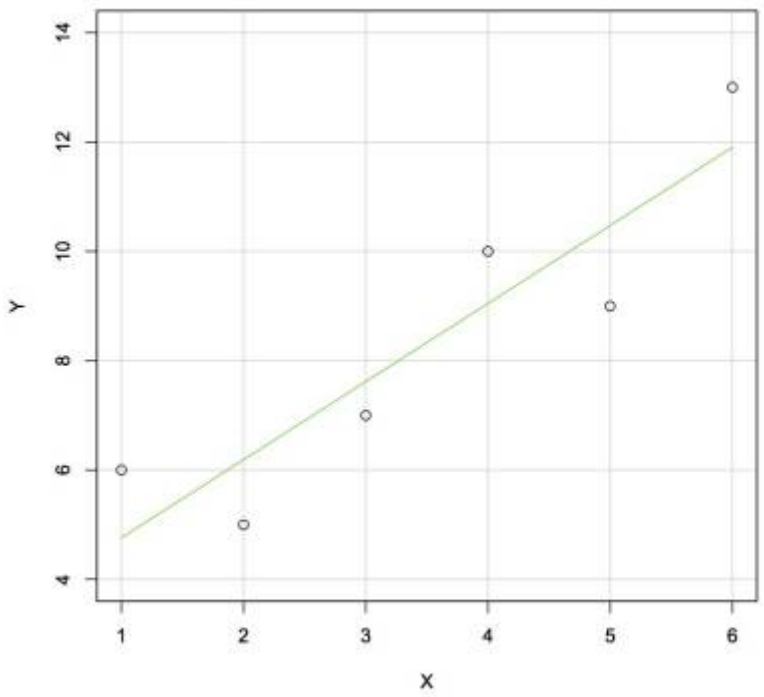
Quando essa premissa é cumprida, temos o que chamamos de **homoscedasticidade**:



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:



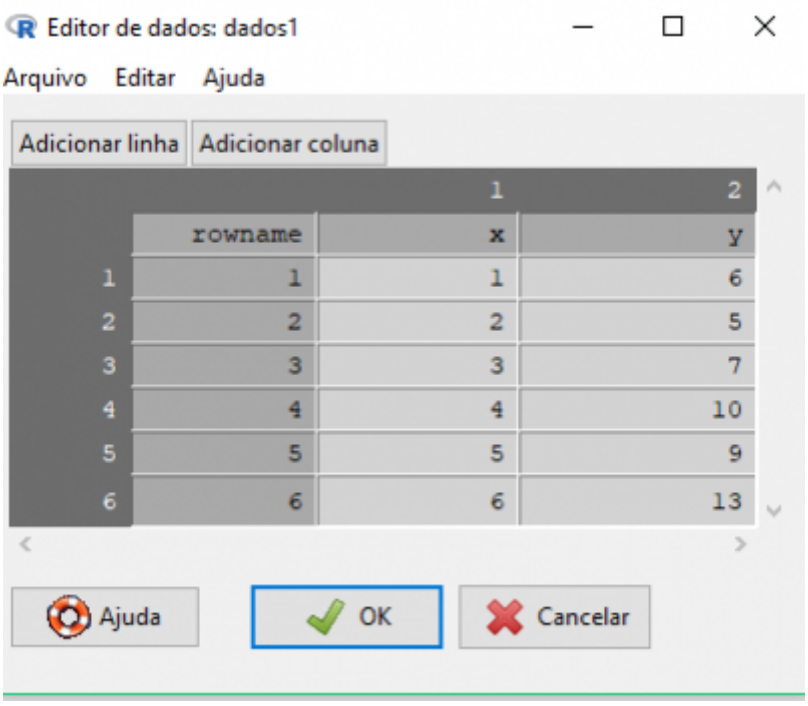
Faça uma tabela como essa e anote os valores aproximados que você consegue obter por esse gráfico:

X	Yobservado	Yestimado	Resíduo
1			
2			
3			
4			
5			
6			

Agora vamos checar no Rcommander esses mesmos dados:

1) Abra o Rcommander (se você não sabe como fazer isso, veja aqui)

2) Crie um novo conjunto de dados no menu **Dados > Novo conjunto de dados**. Defina o nome como *dados1*. Preencha a nova planilha de dados com as informações da figura abaixo e clique OK. Agora você já tem os dados para analisar!



	rowname	x	y
1	1	1	6
2	2	2	5
3	3	3	7
4	4	4	10
5	5	5	9
6	6	6	13

3) Para ajustar um modelo de regressão linear simples vá ao menu **Estatística > Ajuste de Modelos > Regressão Linear...**. Na janela que se abre, escolha a coluna Y dos dados como Variável resposta e a coluna X dos dados como Variável Explicativa e clique OK.

4) Se tudo correu bem, aparecerá na janela */Outputs* o resumo dos resultados da regressão. Caso não apareça, vá em **Modelos > Resumir modelo** e clique em OK[1)].

5) Para visualizar os Y estimados e os resíduos no Rcommander vá em **Modelos > Adicionar estatísticas calculadas aos dados**. Na janela que se abrir, selecione apenas as opções **Valores ajustados** e **Resíduos**. Agora clique no botão **Ver conjunto de dados** e veja as colunas adicionadas. Compare os resultados da regressão feita no Rcommander com os que você calculou a partir do gráfico.

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*

Importe o arquivo para o Rcommander (**Dados > Importar arquivos de dados > de arquivo texto , clipboard, URL...**) e importe os dados *algas.peixes*. Atenção, pois o Separador de Campos que deve ser selecionado para essa planilha de dados é **semicolons [;]**.

Conheça os dados, clicando no botão **Ver conjunto de dados** e também em **Estatísticas > Resumos > Conjunto de dados ativo…**.

Avalie visualmente a relação entre as variáveis com o gráfico de dispersão em: **Gráficos > Diagramas de dispersão**. Na aba de Opções marque **Boxplots marginais**, **Smooth line** e **Mostre espalhamento (spread)**. Como o objetivo dos pesquisadores é analisar o efeito da biomassa de algas sobre a biomassa de peixes, faça o gráfico selecionando biomassa de peixes no eixo Y e biomassa de algas no eixo X.

Ajuste um modelo de regressão linear da biomassa de peixes em função da biomassa de algas. Para isso, vá em **Estatística > Ajuste de Modelos > Regressão linear**. Escolha a biomassa de peixe como Variável resposta e biomassa de algas como Variável Explicativa.

No menu **Modelos** podemos olhar o resumo dos resultados do modelo clicando em **Resumir modelo**, olhando os valores dos coeficientes dos modelos. Como vimos anteriormente, podemos também obter os resíduos e os valores ajustados do modelo clicando em **Adicionar estatísticas calculadas aos dados** e selecionando **Valores ajustados** e **Resíduos**. Esses valores serão colocados como colunas novas na planilha de dados e para visualizá-los, basta clicar no botão **Ver conjunto de dados**.

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

Para isso vamos usar os resíduos da regressão que foram incluídos como uma coluna na sua planilha de dados e aparecem com o nome *“residuals.RegModel.*”* (o “*” será um número que vai depender de quantos modelos você já fez até aqui. Por exemplo, se esse é o segundo modelo que você está calculando desde que abriu o Rcommander, a variável vai se chamar “residuals.RegModel.2”. Mas não se preocupe com esse número).

A partir do menu **Gráficos**, escolha **Histograma** e selecione a variável “residuals.RegModel.*”. **Essa figura se assemelha a uma distribuição normal?**. Se sim, isso é um bom indício de que seus resíduos têm uma distribuição normal. Se não, será necessário repensar se a regressão linear simples é a análise mais adequada para esses dados e/ou se é necessário fazer alguma transformação de variáveis [2)].

Essa é uma análise muito simplista e mais para frente nesse roteiro vamos conhecer outros métodos para avaliar a distribuição dos 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*.

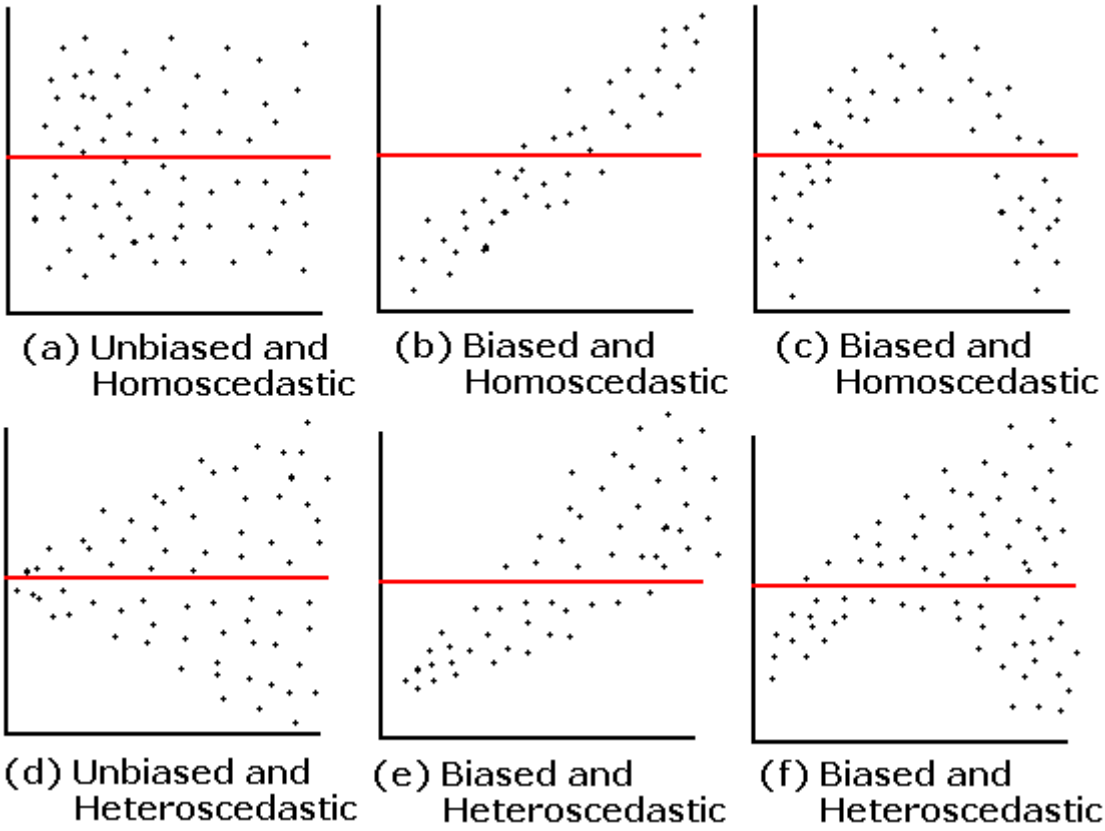
Então, podemos fazer um gráfico em que relacionamos os valores de *Y_estimado* ("*fitted.RegModel.**") e os valores dos *Resíduos* ("*residuals.RegModel.**") para cada *Y_estimado*. Com esse gráfico podemos avaliar se a distribuição dos resíduos é similar ou se há um maior ou um menor espalhamento dos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

Para fazer esse gráfico, vá para o menu **Gráficos > Diagrama de dispersão**, escolha para o eixo Y os resíduos (que foram incluídos na sua planilha de dados como *residuals.RegModel.**) e para o eixo X os valores estimados de Y (que também foram incluídos na sua planilha de dados, como *fitted.RegModel.**). Antes de dar "OK", vá até a aba **Opções** e deixe selecionada apenas a caixa *"Smooth line"*.

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*!

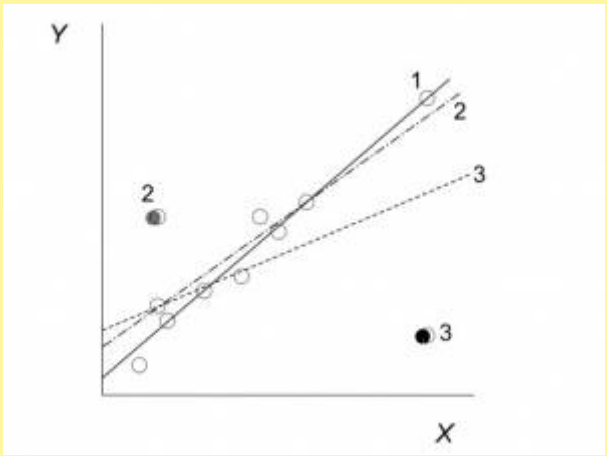
Como saber se alguma observação está influenciando demais os parâmetros da regressão?

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 é extremo considerando a amplitude dos demais valores de X. 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, ou seja, sua influência sobre a estimativa dos parâmetros da distribuição. 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, ou seja, 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 mantida no conjunto de dados. A reta 1 indica como ficaria a reta sem as observações 2 e 3.



Se você não entendeu essa figura, peça ajuda!

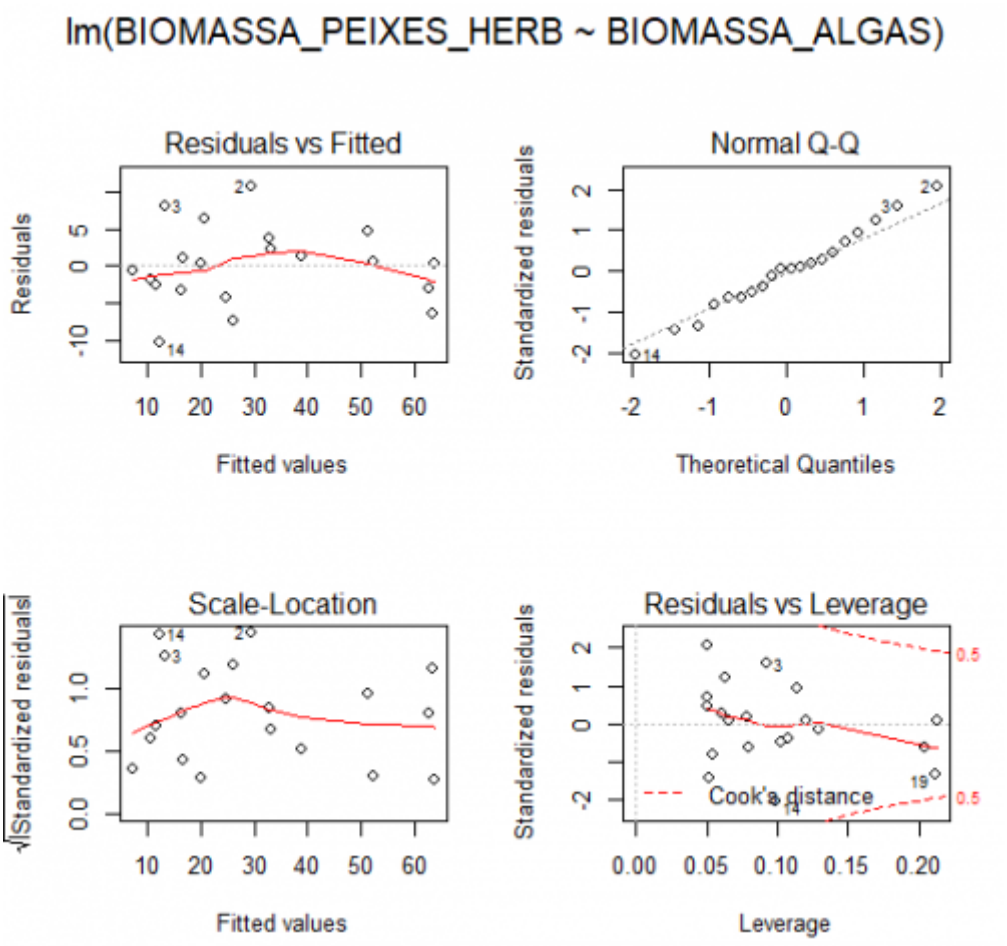
Então, podemos fazer um gráfico em que plotamos o valor dos *Resíduos* em relação aos valores de *leverage* e nesse gráfico os pontos que possuírem as maiores *leverage* e os maiores erros/resíduos (positivos ou negativos) serão as observações com maiores **Distâncias de Cook** e consequentemente, com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

Vá ao menu **Modelos > Resumir modelos**.

Agora, vamos definir que sejam construídos os 4 gráficos de diagnóstico: Vá ao menu **Modelos > Gráficos > Diagnósticos gráficos básicos**.



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[3] 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. Ele nos ajuda a identificar se os resíduos [4] 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

.}}preservefilenames::QQPlot_CaudaLonga_CaudaCurta.jpg

- 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 vermelhas 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

PARA ENTREGAR ANTES DO INÍCIO DA PRÓXIMA AULA

Repita o mesmo procedimento realizado acima para os conjuntos de dados **“algas_peixes2.csv”** e **“vol_ind.csv”** e avalie quais premissas estão sendo atendidas ou não para cada um. Envie pelo formulário abaixo ou pelo link do formulário.

1) Os gráficos de diagnóstico dos dois conjuntos de dados;

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 - influência dos pontos.

1)
não se preocupe com as opções que aparecem na janela que se abrirá
2)
posteriormente falaremos disso
3)
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
4)
note que ele também está usando resíduos padronizados