Princípios da Estatística Frequentista

Os testes clássicos estatísticos estão inseridos no escopo da estatística frequentista ou inferência frequentista. Nessa abordagem a probabilidade é considerada uma frequência e a inferencia está baseada na frequência com que eventos ocorrem nos dados coletados. 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

Conceitos importantes

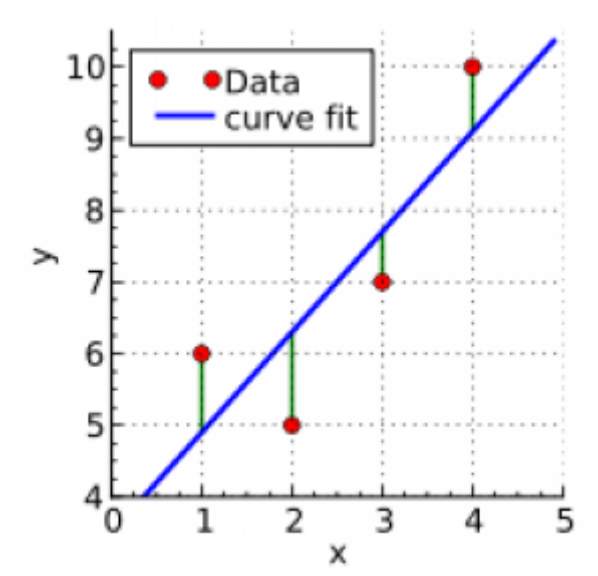
Quando realizamos uma análise mais aprofundada sobre a relação entre duas variáveis numéricas contínuas podemos ajustar uma reta que represente o melhor ajuste entre os dados e que possa nos ajudar a prever valores da variável resposta (eixo Y) a partir de valores da variável preditora (eixo X). Se o 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 relação (reta) entre a variável preditora (x) e a variável resposta (y) é descrita por uma equação simples (**y=a+bx**) com apenas dois parâmetros: a (intercepto) e b (inclinação da reta).
- O **intercepto (a)** representa o valor estimado da variável resposta (y) quando a variável preditora (x) é igual à zero.
- A **inclinação (b)** representa o efeito da variável preditora (x) sobre a variável resposta (y), ou seja, o quanto a variável y aumenta (ou diminui) à cada unidade da variável preditora (x).
- Quando usamos regressão linear para testar hipóteses científicas em ecologia, majoritariamente estamos interessados em saber o efeito da variável preditora (x) sobre a variável resposta (y), ou seja, se a inclinação da reta (b) é significativamente diferente de zero. Em geral, não temos hipóteses científicas acerca do intercepto do modelo.
- 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**)

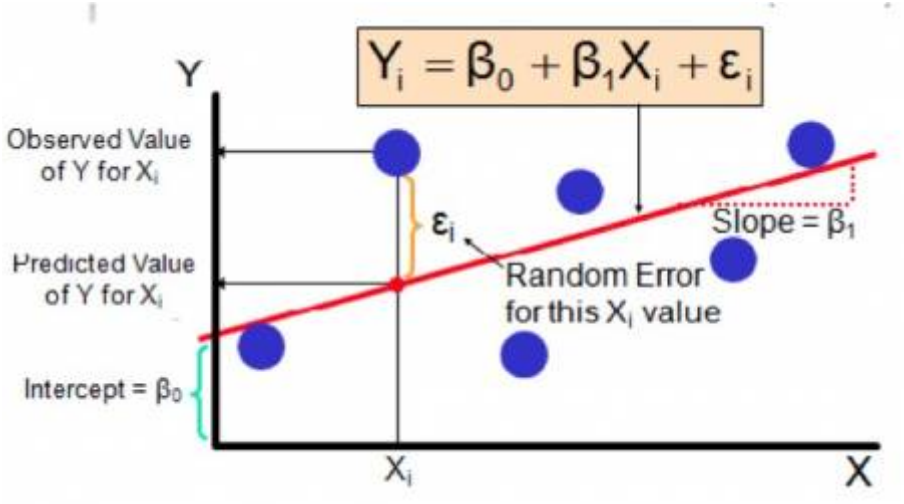
O objetivo desse tutorial é fazer e interpretar os resultados de uma análise de regressão linear, assim como avaliar os pressupostos do modelo.

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.



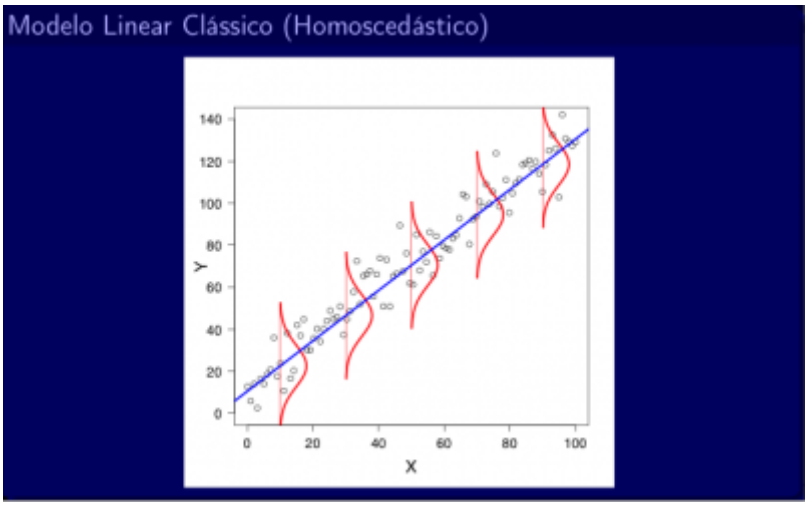
Premissas de uma Análise de Regressão Linear

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

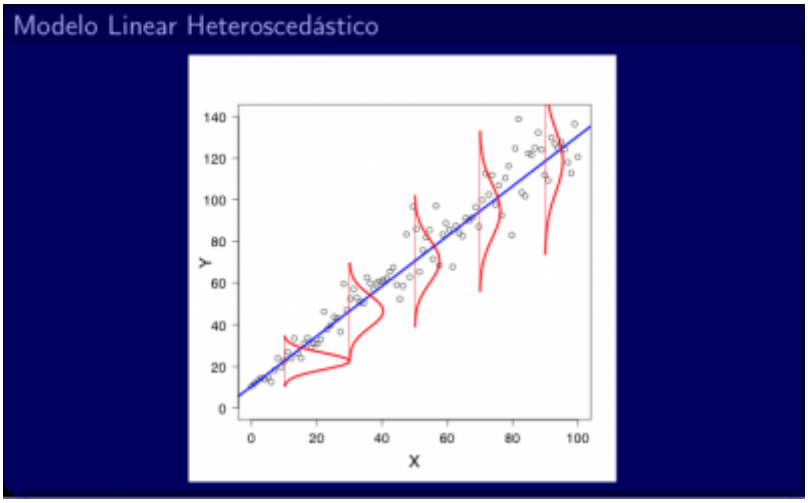
- **DISTRIBUIÇÃO NORMAL DOS ERROS/RESÍDUOS** - Para cada valor de X, os erros devem seguir uma distribuição normal. Se fossem feitas muitas réplicas para cada um dos valores de X, a distribuição dos erros referentes aos vários valores obtidos para Y nas muitas réplicas seguiria uma distribuição normal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



Regressão linear na prática

Agora que você entendeu como funciona a regressão linear, vamos para um exemplo prático.

Baixe o arquivo de dados para o seu diretório:

- produtividade_chuva.txt

Exemplo hipotético

Pesquisadores interessados em entender o efeito da precipitação sobre a produtividade primária liquída em ecossistemas terrestres, selecionaram 30 áreas naturais distribuídas por todo o globo. Dada a importância da água para a fotossíntese, a hipótese dos pesquisadores é que quanto maior a precipitação, maior será a produtividade primária líquida dos ecossistemas. Em cada área, os pequisadores coletaram duas informações: a precipitação anual média (mm) e a produtividade primária líquida (Mg/ha/ano).

Hipóteses estatísticas

Considerando que os pesquisadores estão interessados no efeito da precipitação sobre a produtividade, podemos assumir que a precipitação é a variável preditora (x) e a produtividade é a variável resposta (y). Como ambas as variáveis são contínuas, podemos aplicar uma regressão linear simples para testar a hipótese científica. Neste caso, o efeito de precipitação sobre a produtividade será descrito pela inclinação da reta (b). Sendo assim, as hipóteses estatísticas serão:

- H0: B=0
- H1: B≠0

Como fazer a regressão linear simples

1) Abra o RCommander. Caso vc não tenha instalado o pacote no R, acesse o tutorial que explica

passo à passo como instalar e abrir o RCommander.

2) Importe o arquivo para o Rcommander (**Dados > Importar arquivos de dados > de arquivo texto , clipboard, URL...**) e importe os dados *produtividade_chuva*. Atenção, pois o Separador de Campos que deve ser selecionado para essa planilha de dados é **Tabs**.

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

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

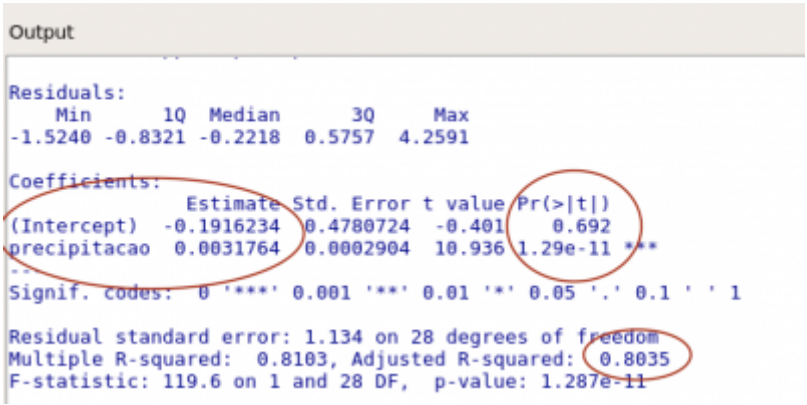
5) Ajuste um modelo de regressão linear da produtividade em função da precipitação. Para isso, vá em **Estatística > Ajuste de Modelos > Regressão linear**. Escolha a produtividade como Variável resposta e precipitação como Variável Explicativa.

6) No menu **Modelos** podemos olhar o resumo dos resultados do modelo clicando em **Resumir modelo**, olhando os valores dos coeficientes dos modelos.

7) Também é possível obter os resíduos e os valores ajustados do modelo clicando no menu **Modelos** 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**.

A hipótese científica foi corroborada?

Para entender os resultados obtidos, primeiramente devemos examinar o resumo dos resultados. Vcs verão algumas informações relacionadas aos dois parâmetros do modelo: o intercepto e a inclinação (no resumo chamado de "precipitação"). Para cada um desses parâmetros há uma estimativa, um erro padrão, um valor de t e um valor de P. Por agora vamos focar apenas na estimativa e em seu p associado.



Quanto ao **intercepto**, vemos que a estimativa feita pelo modelo é de -0.19 e o valor de p é 0.69. Isso significa que quando a precipitação é igual à 0 nosso modelo de regressão estima que a produtividade média dos ecossistemas terretres será de -0.19 MgC/ha/ano. Mas notem que esse valor não é significativamente diferente de zero (como o valor de P=0.69 é maior do que o alfa crítico de

0.05, nós aceitamos H0, a hipótese de que o intercepto é igual a zero). Embora esse resultado possa ser explorado do ponto de vista biológico, lembrem-se que a hipótese científica dos pesquisadores ancora-se na estimativa de b (a inclinação da reta). Então vamos à ela.

Quanto à **inclinação**, o modelo estimou um valor de 0.003 associado a um p extremamente pequeno (P<0.00001). Esses resultados indicam que:

- a precipitação afeta de maneira significativa a produtividade primária em Ecossistemas Terrestres (com este valor de P<0.00001 nós falhamos em aceitar H0 e ficamos com H1, que diz que o B populacional é diferente de 0); Sendo assim, a hipótese científica foi corroborada.
- e com qual magnitude ocorre tal efeito da chuva sobre a produtividade das plantas? Para cada aumento de 1mm na quantidade de chuva média anual de uma localidade observa-se, em média, um aumento em **0.003 MgC/ha/ano** na produtividade primária líquida dos ecossistemas.

Adicionalmente temos o valor de **R2 ajustado** para nos ajudar na interpretação do modelo. O R2 ajustado é de 0.80. Isso significa que a variação na precipitação explica aproximadamente 80% da variação observada na produtividade das diferentes localidades. Os demais 20% são explicados por fatores desconhecidos. Mas, lembre-se que R2 de 80% é muito alto e muito raro de ser encontrado na biologia (efeitos da simulação)!

As premissas do modelo foram atendidas?

Para que as conclusões descritas acima sejam confiáveis, é preciso checar se as premissas do modelo estão sendo atendidas

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 [1)].

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*.

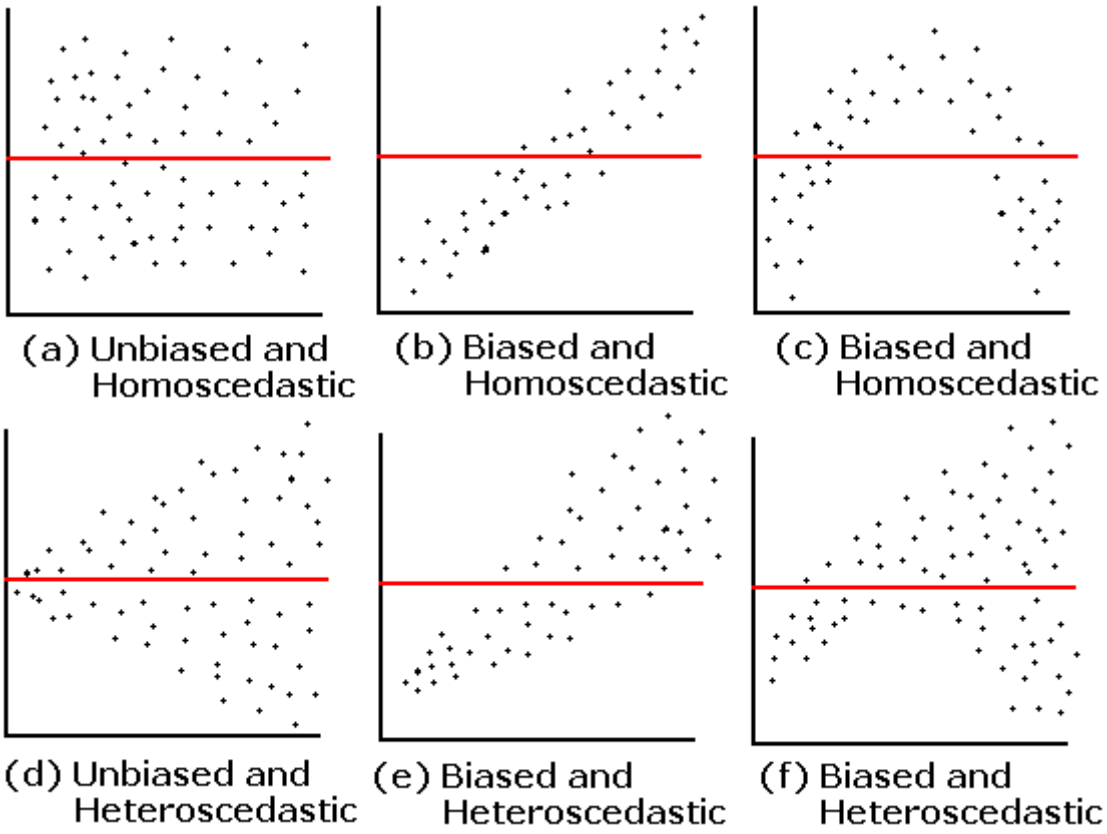
Então, podemos fazer um gráfico em que relacionamos os valores de *Y_estimado* ("*fitted.RegModel.**") e os valores dos *Resíduos* ("*residuals.RegModel.**") para cada *Y_estimado*. Com esse gráfico podemos avaliar se a distribuição dos resíduos é similar ou se há um maior ou um menor espalhamento dos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

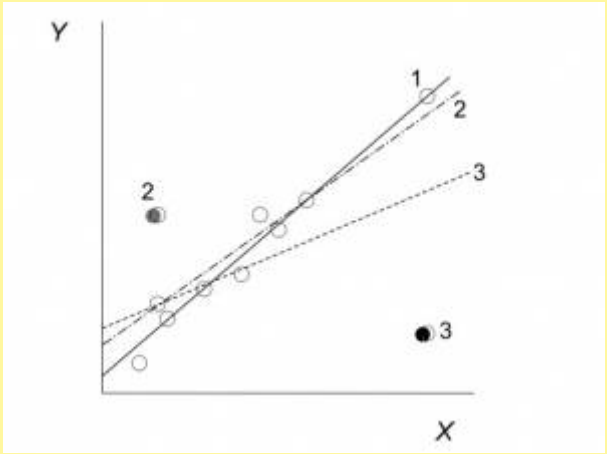
Como saber se alguma observação está influenciando demais os parâmetros da regressão?

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.



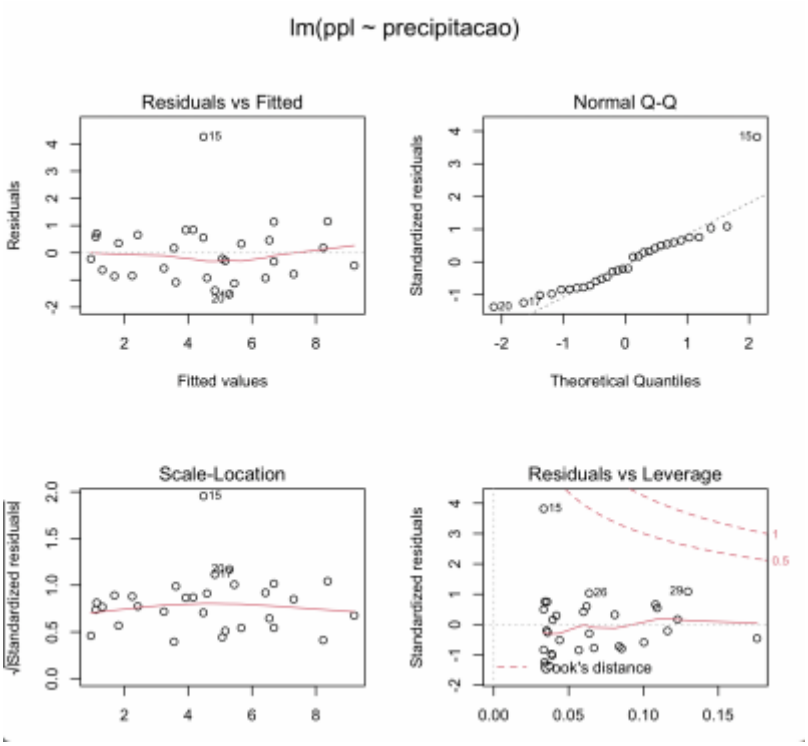
Se você não entendeu essa figura, peça ajuda!

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.

Gráficos Diagnósticos Sintéticos

Para elaborar o conjunto de gráficos diagnósticos do nosso modelo, no RCommander vá em **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[2)] 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 que vemos em nossos dados? De maneira geral, o dados tem variância homogênea e não apresentam viés. No entanto, há um único dado (número 15) que possui resíduo muito maior do que as demais observações (provavelmente um outlier).

- O gráfico superior à direita é o gráfico quantil-quantil. Ele nos ajuda a identificar se os resíduos [3)] 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
. Este gráfico também indica que os resíduos se ajustam bem a uma distribuição normal, com exceção do dado 15 que apresenta uma valor mauito alto e “puxa” a cauda da distribuição, deixando-a mais longa à direita.

- 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. Conforme já indicado nos gráficos anteriores, aqui fica claro que o dado 15 é um outlier pois apresenta resíduo muito elevado. No entanto, esse gráfico indica que seu papel não é muito preocupante já que ele não é um dado influente (está dentro do limite aceito para a distância

de Cook), provavelmente porque possui baixa alavancagem. Sendo assim, não é necessário removê-lo e refazer a análise para avaliar seu papel nos resultados.

PARA ENTREGAR ANTES DO INÍCIO DA PRÓXIMA AULA

Vcs agora continuarão a pesquisa sobre o efeito das condições ambientais na produtividade primária de ecossistemas terrestres. Suponha que vcs estão interessados em entender o efeito da temperatura média anual (°C) sobre a produtividade primária líquida (MgC/ha/ano). Dado que a temperatura média anual dos ecossistemas terrestres está diretamente relacionada ao comprimento da estação de crescimento, a hipótese científica a ser testada é se a temperatura tem efeito positivo sobre a produtividade. Utilize o conjuntos de dados

produtividade_temp.txt

[4], faça a regressão linear simples, interprete os resultados e avalie o atendimento das premissas do modelo. Preencha o seguinte formulário para registrar suas respostas.

1)
posteriormente falaremos disso
2)
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
3)
note que ele também está usando resíduos padronizados
4)
Caso os dados abram em uma aba do navegador, clique com o botão direito do mouse e utilize o menu “Salvar link como...” ou algo parecido para salvar o arquivo em um diretório do seu computador.