

## REGRESSÃO LINEAR

A análise de regressão linear é possivelmente a técnica mais difundida de toda a estatística aplicada. O método não tem apenas um número enorme de aplicações em si mesmo, mas também forma a base conceitual de muitas das técnicas mais avançadas, como análise multivariada, análise de sobrevivência e suavização não paramétrica. Portanto, *um entendimento do que estamos fazendo ao usar uma regressão linear é essencial* para aqueles que necessitam fazer análises estatísticas em quase todos os campos de aplicação.

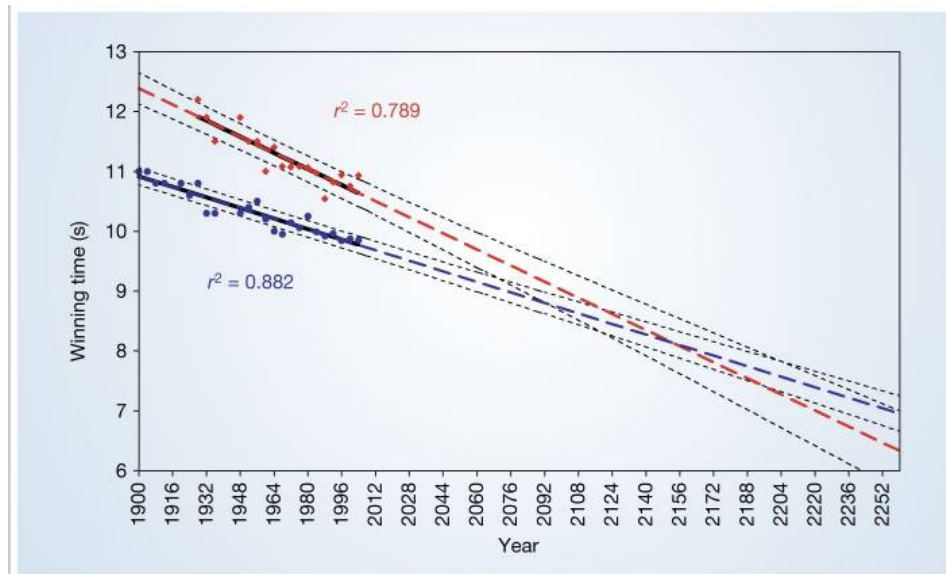
O objetivo da modelagem é **simplificar** o mundo complexo a nossa volta e nos concentrar na essência do problema. *Um modelo não precisa conter todos os detalhes do mundo complexo para ser útil.* Podemos quebrar os problemas em pequenas partes para ajudar na compreensão, isto é, podemos tentar manter constantes algumas das variáveis envolvidas em nosso estudo a fim de nos concentrar nas demais. *Entretanto, devemos estar atentos aos problemas de variável omitida e causalidade reversa para minorar erros nas nossas conclusões sobre causa e efeito.*

Exemplo (causalidade reversa): *em cidades com mais policiais há um número maior de crimes. Logo, o policiamento aumenta a violência urbana!?* O número de policiais determina a criminalidade ou a criminalidade determina o número de policiais?

Neste texto, explicaremos procedimentos usuais para determinar quais regressores serão utilizados no modelo.

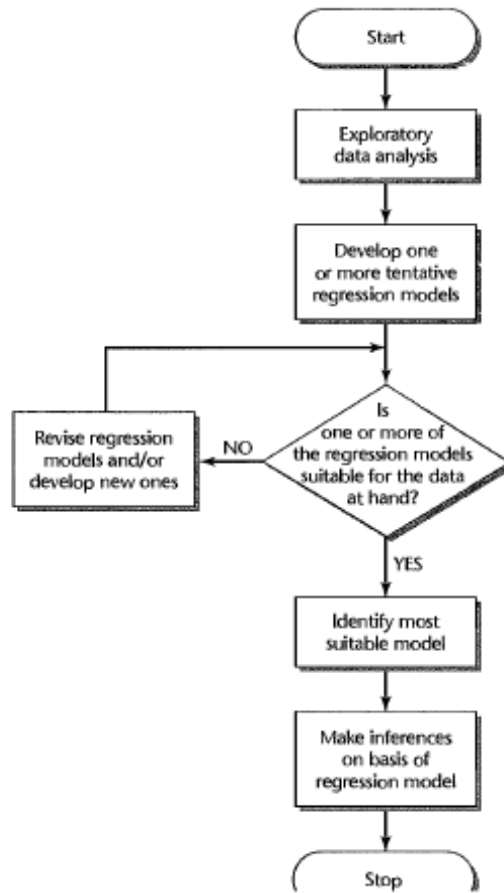
Algumas vezes, é útil transformar ou recodificar os dados para obter um bom modelo. Em economia, costuma-se usar a transformação logaritmo com frequência, mas outras estão disponíveis (inversa, potências) e serão testadas (veja família de transformações Box-Cox, por exemplo).

Exemplo: Momentous sprint at the 2156 Olympics? (Tatem et al, 2004)



Um modelo de regressão é um meio formal de expressar um ingrediente fundamental de uma relação: uma tendência de uma variável resposta (dependente)  $Y$  variar de acordo com a variação de um regressores  $X$  de uma maneira sistemática (em outras palavras, queremos medir o impacto da variação nos níveis dos regressores na superfície de resposta). Postularemos que existe uma distribuição de probabilidade de  $Y$  para cada nível de  $X$  e que esta distribuição de probabilidade varia de uma maneira sistemática conforme  $X$  varia.

A estrutura típica de uma análise de regressão está representada no diagrama a seguir:



- (1) Coleta dos dados;
- (2) Primeiramente, costuma-se explorar os dados (fazendo uma análise gráfica, calculando medidas resumo (correlações, por exemplo)).
- (3) *Desenvolver o modelo teórico. Se for preciso, caso não tenhamos um modelo inicial*, costumamos checar as correlações amostrais entre a variável dependente e os possíveis regressores (utilizando as que são significativas posteriormente, por exemplo). No entanto, tome cuidado para não minerar relações espúrias.

O modelo teórico para o fenômeno em estudo é montado com base em relações causais apontadas na área de estudo. Podemos também combinar um modelo teórico com a análise exploratória.

- (4) Com base no modelo mais amplo (inicialmente costumamos incluir to-

das as variáveis potencialmente importantes para explicar o problema) iremos identificar qual o modelo que se ajusta melhor a nossa informação disponível (a amostra). Para isto utilizamos técnicas de seleção do modelo e fazemos os testes das hipóteses clássicas de um modelo de clássico de regressão linear (homocedasticidade, não autocorrelação, não multicolinearidade, variável omitida, etc...).

O modelo final que utilizaremos será aquele que minimiza os critérios de informação dentre aqueles em que não conseguimos refutar as hipóteses de que o resíduo está “bem ajustado”.

Obs: Este procedimento (*data driven*) recebe críticas.

(5) Previsões e Validação do Modelo.

Iremos expandir estas ideias/passos na sequência do texto.

Um bom modelo é aquele em que as variações da variável dependente (regressando)  $Y$  são razoavelmente explicadas pelas variações das independentes (regressores)  $X$  e **as variações não explicadas (que estão contidas no termo aleatório residual  $u$ ) são aleatórias.**

O procedimento científico pressupõe a validação das hipóteses (para ser corretamente utilizado e para que as inferências sejam adequadas).

Quando minimizamos a soma de quadrados dos resíduos com respeito aos parâmetros (betas) estamos fazendo uma **projeção ortogonal** de  $Y$  sobre o subespaço das colunas geradas por  $X$ . Encontraremos  $\hat{Y} = P_X Y$ , onde  $P_X$  é a matriz de projeção de  $Y$  sobre as colunas de  $X$ .

*Uma regressão, quando minimizamos a soma dos quadrados dos resíduos com respeito aos parâmetros, é a **modelagem de expectativas condicionais**. Queremos, com base na informação obtida (amostra) encontrar uma relação entre as potenciais explicativas e a dependente (queremos estimar a superfície de resposta média, com base na amostra, isto é, um estimador para a média condicional de  $Y$  dadas as informações obtidas, assim podendo estimar os impactos médios das relações causais).* Veja um exemplo no slide 7/18 em [http://rpubs.com/fsabino\\_da\\_silva/383453](http://rpubs.com/fsabino_da_silva/383453)

## Hipóteses do Modelo Clássico de Regressão Linear

H1 (Linearidade dos Parâmetros) A relação entre a variável dependente  $Y$  e as explicativas  $X_1, X_2, \dots, X_k$  é linear:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \dots + \beta_k X_{ki} + u_i$$

Importância: Essa hipótese é necessária para termos uma solução única (e fechada). Sem ela não podemos utilizar o procedimento de mínimos quadrados (a projeção ortogonal citada anteriormente).

H2 (Amostragem Aleatória) - Lei dos Grandes Números e Teorema Central do Limite costumam assumir que a amostragem seja aleatória.

H3:  $E(u | X_1, \dots, X_k) = 0$  para qualquer combinação dos valores das variáveis explicativas (na população). Esta hipótese implica que a  $Cov(u, X_i) = 0$  para qualquer  $X_i$ . Quando isto acontece dizemos que os regressores  $X$  são exógenos.

Importância: Se esta hipótese for válida os estimadores de mínimos quadrados serão não-viesados e consistentes.

H4: Não há multicolinearidade perfeita entre as variáveis explicativas, isto é, a matriz  $X$  tem posto completo.

Importância: Esta hipótese é necessária para que o estimador exista, caso contrário não teremos uma solução (inversas de matrizes não estarão bem definidas).

H5:  $Var(u | X_1, \dots, X_k) = \sigma^2$  para qualquer combinação dos valores das variáveis explicativas (na população). Chamamos esta hipótese de homocedasticidade (da distribuição dos resíduos).

A variância condicional do termo aleatório residual é constante para qualquer combinação dos níveis dos regressores na *população*.

Importância: Com base nesta hipótese e na próxima, tudo o mais constante, conseguiremos demonstrar que os estimadores de mínimos quadrados são os mais precisos (com menor variância dentro da classe dos estimadores lineares não-viesados). Veja Teorema de Gauss Markov.

H6:  $Cov(u_i, u_j | X_1, \dots, X_k) = 0, i \neq j$ . A matriz de variâncias e covariâncias é diagonal.

H7: O tamanho da amostra é maior do que o número de regressores. Esta hipótese é necessária para que possamos encontrar os estimadores de MQ.

H8: Normalidade: O processo gerador da sequência aleatória ( $u$ ) segue uma distribuição normal com média dada por H3 (zero) e variância dada por H5, isto é,

$$u | X_1, \dots, X_k \sim N(0, \sigma^2)$$

Pequenos desvios da hipótese de normalidade não nos causam sérios problemas quando queremos utilizar o procedimento que estamos estudando. Grandes desvios, no entanto, devem nos causar problemas na hora de fazer inferências (se a amostra for pequena pode causar sérios danos à análise).

Olhando as hipóteses elencadas podemos dividi-las em duas partes: (a) hipóteses calcadas em álgebra linear (e cálculo) - 1, 4 e 7 – necessárias para termos uma solução única e fechada (que pode ser expressa através de uma fórmula); (b) hipóteses sobre a distribuição de probabilidade da sequência de resíduos (3, 5, 6 e 8): podemos utilizar procedimentos inferenciais para testar estas hipóteses (as demais necessitam ser assumidas para que possamos utilizar o procedimento de mínimos quadrados).

Com base na explicação do parágrafo anterior, fica claro que as análises visam testar as hipóteses associadas ao termo residual. Se elas forem válidas então teremos estimadores com boas propriedades estatísticas (não-viesados, consistentes e eficientes). Caso elas não sejam válidas, devemos trocar as hipóteses por outras que sejam mais realistas com o que encontramos na amostra. Os procedimentos serão descritos nos passos a seguir.

**Ponto Chave:** Quanta estrutura deve ser imposta? A não imposição de alguma estrutura permite qualquer relação possível entre as variáveis, mas poderemos ter um estimador mais impreciso, isto é, um estimador com variância mais alta. No entanto, uma vez que a estrutura imposta não deve refletir exatamente o processo estocástico (aleatório) que gerou os dados (de onde foram retiradas as observações), então é provável que tenhamos algum erro de especificação. **Desta forma, devemos sempre ter em mente**

**quanta estrutura deve ser imposta** para explorar o possível *trade-off* entre precisão (variância) e erro de especificação.

Um dado pode mudar completamente a inclinação de uma curva/superfície. É importante, portanto, encontrar estes possíveis valores influentes (costuma-se categorizar em três tipos: outliers, influentes e de alavancagem). Veja um exemplo onde a análise exploratória é importante aqui: [https://rstudio-pubs-static.s3.amazonaws.com/52381\\_36ec82827e4b476fb968d9143aec7c4f.html](https://rstudio-pubs-static.s3.amazonaws.com/52381_36ec82827e4b476fb968d9143aec7c4f.html)

Antes de prosseguir, vamos retomar o ponto inicial deste passo. Uma análise descritiva (gráfica) é interessante para verificar se os resíduos parecem aderir às hipóteses clássicas do modelo de regressão linear.

Os resíduos amostrais podem ser utilizados para analisar seis tipos de desvios das hipóteses: (i) A função de regressão é não linear; (ii) A sequência de resíduos não apresenta variância constante; (iii) Os resíduos não são independentes; (iv) Existência de outliers; (v) A sequência não é normalmente distribuída; (vi) Uma ou mais variáveis importantes foram omitidas do modelo.

Para isto podemos analisar os seguintes gráficos (muitos podem ser utilizados para mais de uma função – entre parênteses coloquei as possivelmente mais importantes):

- Gráfico dos resíduos amostrais contra um regressor (i,ii);
- Gráfico dos resíduos amostrais ao quadrado (ou em valor absoluto) contra um regressor (ii);
- Gráfico dos resíduos amostrais contra os valores ajustados (i,ii);
- Gráfico dos resíduos amostrais contra o tempo ou outra sequência (iii);
- Gráfico dos resíduos amostrais contra regressores omitidos (vi);
- Box plot dos resíduos (iv,v);
- QQ-plot (iv,v).

Para mais detalhes veja a seção 3.3 de Kutner et. al (2004). Mais adiante veremos outros tipos de diagnósticos (aqui estamos sugerindo apenas uma análise exploratória, isto é, sem fazer testes de hipóteses). Temos que ter em mente que o número de observações deve ser razoavelmente grande para que estes gráficos nos forneçam uma informação mais confiável sobre o formato da distribuição dos resíduos (por isso, usualmente é preferível procedimentos

mais objetivos, baseados em testes das hipóteses, onde podemos quantificar melhor a nossa incerteza acerca das hipóteses).

A análise subjetiva, no entanto, pode revelar dificuldades de maneira mais clara que testes objetivos de antemão. Alguns testes mais objetivos (minimizando nossos possíveis vieses podem/devem ser feitos). Exemplos:

- Teste de aleatoriedade dos resíduos: Runs Test, Durbin-Watson (se tivermos uma série temporal), etc.;
- Testes de que a variância é constante: Breusch-Pagan, White, Koenker, etc...
- Testes para outliers: resíduos “studentizados” com ajuste de Bonferroni (veja `rstandard` e `rstudent` no R). Trataremos mais adiante deste resíduos.
- Testes de Normalidade: Kolmogorov-Smirnov, Jarque-Bera, teste qui-quadrado, etc.
- Teste de endogeneidade: Hausmann (se desconfiamos da endogeneidade de algum regressor).

Visão global das medidas corretivas:

- Se uma regressão linear não é apropriada para o nosso conjunto de dados, existem, basicamente, duas possibilidades: (1) Abandone o modelo e desenvolva um modelo mais apropriado; (2) Faça transformações do modelo (aplicando logaritmo, por exemplo). Veja transformações Box-Cox, por exemplo;
- Cada enfoque possui vantagens e desvantagens. O primeiro enfoque nos leva a usar modelos mais complexos que podem levar a melhores *insights*, mas também pode nos levar a procedimentos também mais complexos de estimação dos parâmetros. O uso bem-sucedido de transformações, por outro lado, nos leva a métodos relativamente simples de estimação que podem envolver menos parâmetros do que nos modelos mais complexos (vantajoso se a nossa amostra for pequena, devido ao limitado número de graus de liberdade). No entanto, as transformações podem obscurecer as interconexões fundamentais entre as variáveis, embora algumas vezes possa acontecer o contrário.

Em resumo, a análise exploratória serve para extrairmos algum *insight* inicial do pedaço de informação disponível a nossa disposição (a amostra).



Algo muito importante antes de descrevermos as técnicas é saber que os testes não são independentes, isto é, ao testarmos uma hipótese não devemos fazer correções sem antes testar as demais, pois uma correção afeta os resultados dos outros testes. O ideal é fazermos todos os testes de uma vez só e com base em todos fazermos as correções. No entanto, costumamos utilizar algum critério para diminuir um pouco a quantidade de cálculos necessárias para chegar a um bom modelo final.

Obs: Estatística é uma ciência (adota um procedimento científico para tomada de decisão). Para utilizar bem a metodologia, precisamos aderir a um conjunto simples de regras: testar ideias, experimentando e observando, desenvolver as ideias que passam pelo teste, ajustar aquelas que não passam, seguir evidências onde quer que nos levem e questionar tudo.

Costumamos, então, inicialmente fazer um procedimento de seleção da melhor combinação dos regressores (dentre todos os iniciais sugeridos como possivelmente relevantes). A melhor combinação é aquela que minimiza algum critério (erro quadrático médio, erro quadrático médio de previsão ou outra medida semelhante – critérios de informação de Akaike e Bayesiano (AIC e BIC), por exemplo). Os critérios de informação são comumente utilizados, pois estudos mostram que eles selecionam o verdadeiro modelo um número maior de vezes do que medidas como o  $R^2$  ajustado.

Alguns testes costumam ser feitos para checar a adequacidade da nossa regressão. Exemplos:

- Multicolinearidade: VIF e Números de Condição;
- Seleção do Modelo: Backward e critérios de informação;
- Heterocedasticidade;
- Diagnósticos para Observações Influentes: Outliers (valores “grandes” em  $Y$ ), Pontos Influentes (valores que afetam as estimativas, previsões, variância dos estimadores, etc.) e de Alavancagem (influentes com valores não usuais na matriz chapéu, isto é, apenas consideramos valores na matriz de delineamento  $X$ ).

Outliers: Resíduos Padronizados Internamente e Externamente;

**Medidas de Influência:** As maiorias das medidas foram descritas em Belsley, Kuh and Welsch (1980) e são baseadas em diagnósticos de exclusão.

As medidas de influência mais utilizadas são:

DFFITS: Mede o quanto o valor ajustado de  $Y_i$  é afetado ao excluir a observação  $i$  do ajuste;

DFBETAS: Ao contrário de DFFITS, esta medida visa medir a influência de uma observação nas estimativas dos parâmetros;

COVRATIO: Mede o efeito da exclusão de uma observação na variância das estimativas dos parâmetros.

- Multicolinearidade: VIF e Números de Condição: A consequência principal da multicolinearidade é que ela afeta variâncias e covariâncias e, consequentemente, intervalos de confiança;

- Não-Linearidade: Gráficos de Resíduo Parcial, etc.;

- Teste de Não-Autocorrelação: Durbin Watson, etc. - caso tenhamos evidências de heterocedasticidade e/ou autocorrelação nos resíduos, devemos usar um estimador que corrija o potencial problema (um estimador de mínimos quadrados ordinários não seria o estimador mais eficiente agora dentro da classe dos estimadores lineares não-viesados). As correções mais comumente utilizadas são: MQP (mínimos quadrados ponderados) quando sabemos o formato da heterocedasticidade (quem causa a heterocedasticidade, basicamente), MQG (mínimos quadrados generalizados), Newey-West (HAC);

- Teste Global de Suposições do Modelo: Outra possibilidade é fazer um teste que avalia a validade de várias hipóteses simultaneamente (importante já que os testes não são independentes como comentado acima). Um exemplo é o teste desenvolvido no seguinte artigo:

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2820257>. Os autores construíram um teste global para testar  $H_0$ :  $(A_1)-(A_4)$  são válidas,  $H_1$ : pelo menos uma não é válida, onde  $A_1$ : Linearidade,  $A_2$ : Homocedasticidade,  $A_3$ : Não Autocorrelação e  $A_4$ : Normalidade.

O teste visa evitar uma grande probabilidade do erro do tipo I (ao fazermos testes separados, cada um irá conter uma probabilidade de erro e no final das contas o erro total poderá/deverá ser muito maior que o nível de significância utilizado em cada teste)). Os autores comentam que podemos usar testes direcionais (individuais) para encontrar evidências de qual ou quais suposições não são válidas na sequência. Isto é importante, pois para encontrarmos um 'remédio' apropriado, precisamos saber qual(is) das violações deve(m) ter acontecido.

## Forma funcional

- Teste RESET: Para a avaliação da forma funcional, e para que se detecte algum problema na especificação desta, foi proposto o teste RESET de Ramsey. Uma das limitações do teste RESET, segundo Wooldridge (2009), é que ele não dá sinais de qual caminho tomar caso a hipótese nula seja recusada.

- Teste LM

Iremos agora explorar um pouco mais sobre a parte de Diagnósticos (outliers, valores influentes e de alavancagem).

### Tipos de Observações Não Usuais (1)

- Outliers: podem causar más interpretações, pois podem ter uma influência forte no modelo. Muitas vezes, quando retiramos um único valor discrepante os resultados ficam completamente diferentes. Veremos alguns gráficos para entender melhor o fenômeno adiante.

- Casos que são outliers com grande alavancagem exercem influência nas inclinações (e intercepto) do modelo

- Outliers podem ser um indicativo de que o nosso modelo está falhando na captura de importantes características dos dados.

- Um outlier (no contexto de regressão) é uma observação que tem um grande (discrepante) resíduo (diferença entre o valor previsto e o ajustado, isto é, na variável dependente  $Y$ ). Este resíduo não afeta necessariamente os coeficientes de inclinação (veja gráficos mais adiante).

### Tipos de Observações Não Usuais (2)

- Alavancagem: Uma observação que tem um valor atípico na matriz  $X$ ;
- Quanto mais longe estiver da “media” de  $X$  (não importa se para mais ou para menos), maior será a alavancagem da observação no ajuste da regressão;

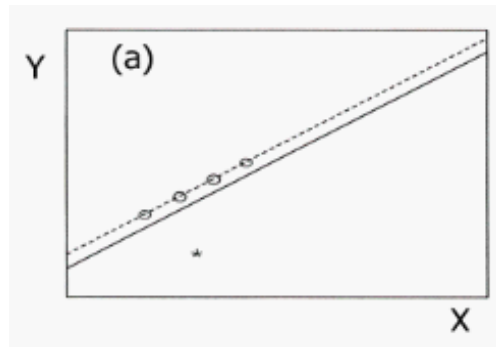
- Grande alavancagem não significa, necessariamente, que a observação influencia os coeficientes da regressão. Isto ocorrerá se ela tiver uma grande alavancagem, mas já seguir o padrão do resto

### Tipos de Observações Não Usuais (3)

- Observações Influentes: Combina grande alavancagem (valor não usual em  $X$ , na matriz de regressores) e “discrepância” (é um outlier em termos do seu valor da variável dependente  $Y$ ). Observações com estas características irão influenciar fortemente a regressão.

- *Nem sempre devemos retirar a observação do banco de dados. A ferramenta não foi desenhada para tomar a decisão*; ela apenas avisa que a observação está fora do padrão de acordo com as regras que elas adotam. *Cabe ao pesquisador verificar e retirar se fizer sentido* (muitas vezes, ocorre um problema de digitação, por exemplo). Vejamos alguns gráficos para entendermos melhor os casos:

Figura (a): Outlier sem influência.



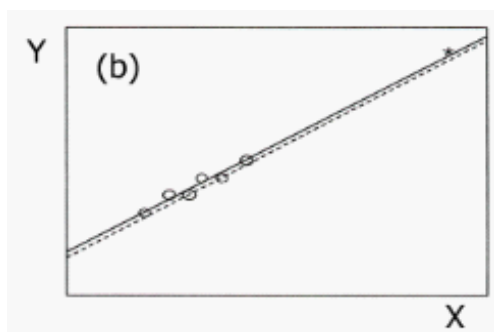
Embora o valor de  $Y$  seja não usual dado o seu valor de  $X$ , ele tem pouca influência na curva de regressão. Isto é devido ao fato da observação estar no “meio” da amplitude de  $X$ . Iremos detectar estas observações usando os resíduos “studentizados” (por exemplo, usando `studres()` no R).

Obs: Resíduos padronizados (por exemplo, `rstandard()` no R) são muitas vezes utilizados, mas o numerador e o denominador da razão não são independentes e, assim, a estatística não irá seguir uma distribuição  $t$ .

Quando os resíduos são “studentizados” esta independência ocorre e costuma-se utilizar como regra usual colocar um “*flag*” na observação quando o valor estiver entre os 2.5% maiores ou menores (fixando a probabilidade do erro do tipo I em 5% - aproximadamente dois desvios-padrão).

Obs: Cuidado ao usar este teste. Se você retirar todas as observações que você ultrapassaram os limites da regra e rodar a regressão de novo, então novos valores irão aparecer como fora do padrão, possivelmente. A ferramenta, novamente, não nos diz para retirar a observação. Ela coloca uma “flag” (atenção) para os usuários analisarem (é uma ajuda na tomada de decisão e não um algoritmo automático).

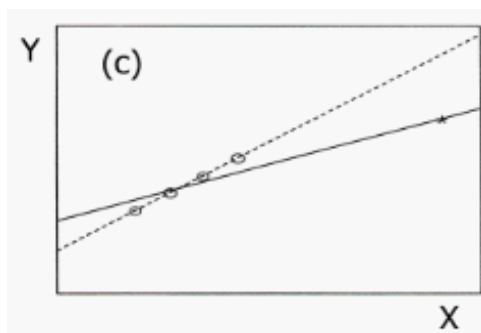
- Figura (b) Grande alavancagem.



A observação tem um grande valor em  $X$  (acima da média das demais observações). Entretanto, o seu valor de  $Y$  está no padrão geral dos demais dados e, então, a observação não tem influência na curva. Detectamos esta observação analisando a matriz chapéu (hat matrix, em inglês).

- Figura (c): Combinação de um valor não usual em  $Y$  (outlier) e em  $X$  (alavancagem) resulta em grande influência. Afeta tanto o intercepto quanto a inclinação.

DFFITS, DFBETAS, COVRATIO e outras irão detectar a observação fora do padrão em (c)



## Testes Formais para Encontrar um Outlier

- *Ajuste de Bonferroni*: Por conta do que comentamos anteriormente, quando estamos selecionando o outlier mais extremo não é uma boa ideia utilizar um simples teste  $t$ . Nós esperaríamos que 5% dos resíduos “studentizados” ultrapassassem os limites, mesmo sem estas observações serem de fato outliers (ultrapassassem aleatoriamente). Para remediar esta situação, costuma-se fazer o ajuste de Bonferroni (por exemplo, podemos usar a função `outlier.test()` da `library(car)` no R).

- Podemos ainda usar gráficos de quantis (*quantile comparison plots*) ou os gráficos de Atkinson (estes últimos baseados em simulações) para ajudar a detectar observações com algum tipo de discrepância.

Ao invés de avaliar desvios de normalidade, podemos, utilizando a mesma ideia (QQ) comparar a distribuição de resíduos “studentizados” do nosso modelo com a distribuição  $t$  apropriada.

### Opções:

(a) Gráfico de influência (*influence plot* ou “*bubble plot*”)

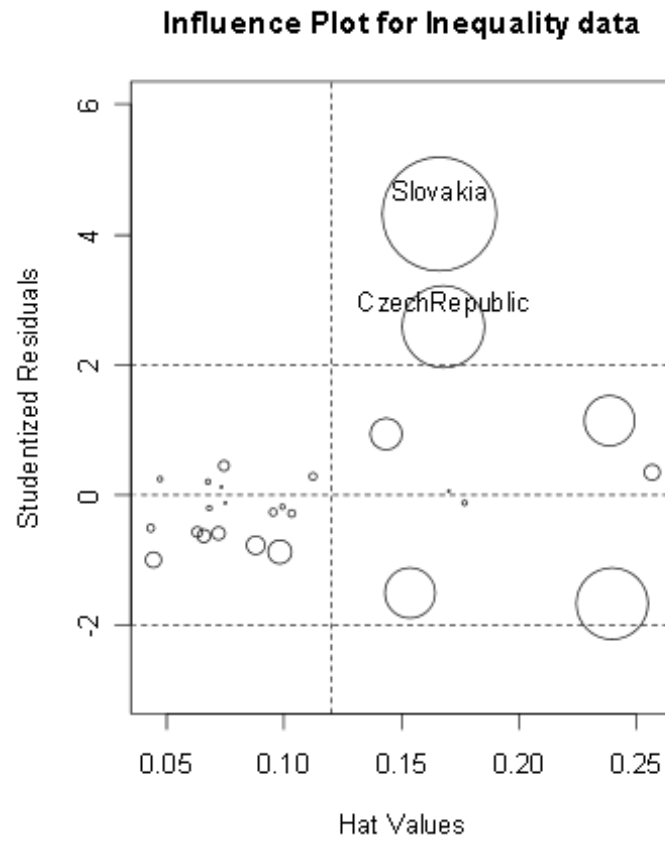
- Mostra ao mesmo tempo
  - resíduos “studentizados”
  - valores da matriz chapéu  $H$
  - D de Cook (Atkinson (1985) definiu uma estatística Cook modificada e que após ajustada por uma constante é equivalente a DFFITS. Logo, não seria necessário considerar esta medida como um diagnóstico separado, desde que todas as informações relevantes podem ser encontradas em DFFITS).

- O eixo horizontal representa os valores da matriz chapéu  $H$ , onde  $H = X(X^T X)^{-1} X^T$  ( a matriz  $H$  é chamada de matriz chapéu, devido ao fato de que  $\hat{Y} = HY$ , isto é, é a matriz que coloca um  $\hat{\phantom{y}}$  em  $Y$ ).

- O eixo vertical representa os resíduos “studentizados”.

- Os círculos representam o tamanho do peso relativo da estatística D de Cook. O raio é proporcional a raiz quadrada da estatística e, portanto, as áreas são proporcionais a estatística.

Exemplo:



(b) Influência conjunta

- A estatística  $D$  de Cook pode nos ajudar a determinar a influência conjunta de mais de uma observação se houver um número pequeno de observações influentes.

- Poderíamos excluir os casos sequencialmente, atualizar o modelo e explorar a estatística  $D$  de novo, mas isto não é recomendável este procedimento como explicamos anteriormente).

- Isto costuma ser impraticável, pois há potencialmente um número grande de subconjuntos para explorar.

- Gráficos de regressão parcial (*Added-variable plots*, também chamados de *partial-regression plots*) nos fornecem um método mais útil de avaliar a influência conjunta.

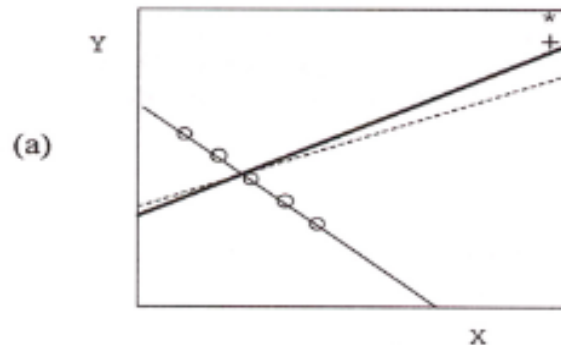
Exemplo:

- A linha sólida abaixo representa a regressão com todos os casos incluídos. A linha "quebrada" é a regressão sem a observação marcada com um asterisco. A linha sólida "mais leve" representa a regressão com ambas as observações (+ e \*) excluídas.

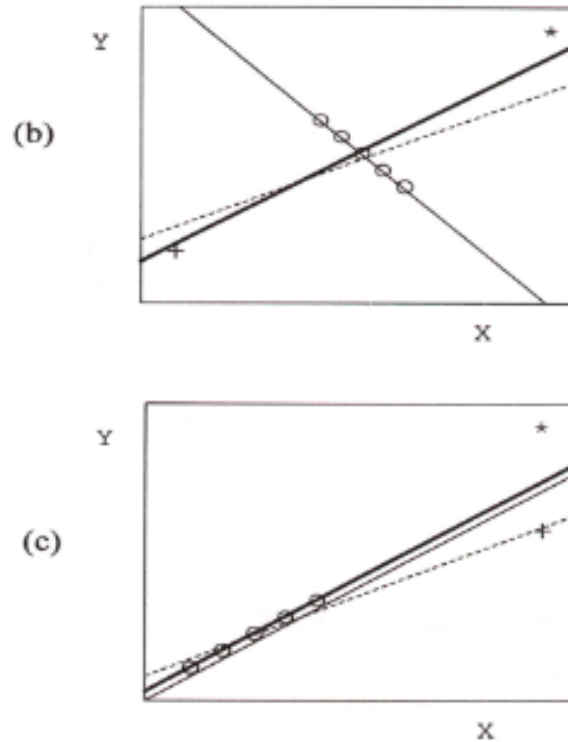
- Dependendo de onde o conjunto de casos influentes cair, eles podem ter um efeito conjunto com efeitos muito diferentes na curva de regressão.

- (a) e (b) mostram exemplo de observações conjuntamente influentes, pois elas mudam as inclinações quando incluídas conjuntamente.

- As observações em (c) quando incluídas conjuntamente se compensam e, assim, têm pouca influência na inclinação da curva (linha de regressão, no caso).







- Outros gráficos: CRPlots, ou *Component Residual Plots* (gráficos residuais dos componentes), auxiliam a identificar se os regressores possuem uma relação linear com a variável dependente dado que os outros regressores estão mantidos no modelo.

- Quando temos apenas um regressor, um *scatterplot* (gráfico de dispersão) da variável dependente contra o regressor fornece uma indicação da natureza da relação. Se há mais que um regressor, a visualização se torna mais complicada, embora ainda seja útil fazer gráficos parciais de dispersão. Entretanto, isto não levará em conta o efeito que outro regressor esteja causando no modelo.

- Estes gráficos são uma extensão dos “gráficos dos resíduos parciais” (*partial residuals plots*). Gráficos dos resíduos parciais, essencialmente, tentam modelar os resíduos de um regressor contra a variável dependente. CRPlots adicionam uma linha indicando onde o melhor ajuste (linear) do modelo se encontra.

- No R podemos utilizar a opção `crPlots` dentro da `library(car)`.

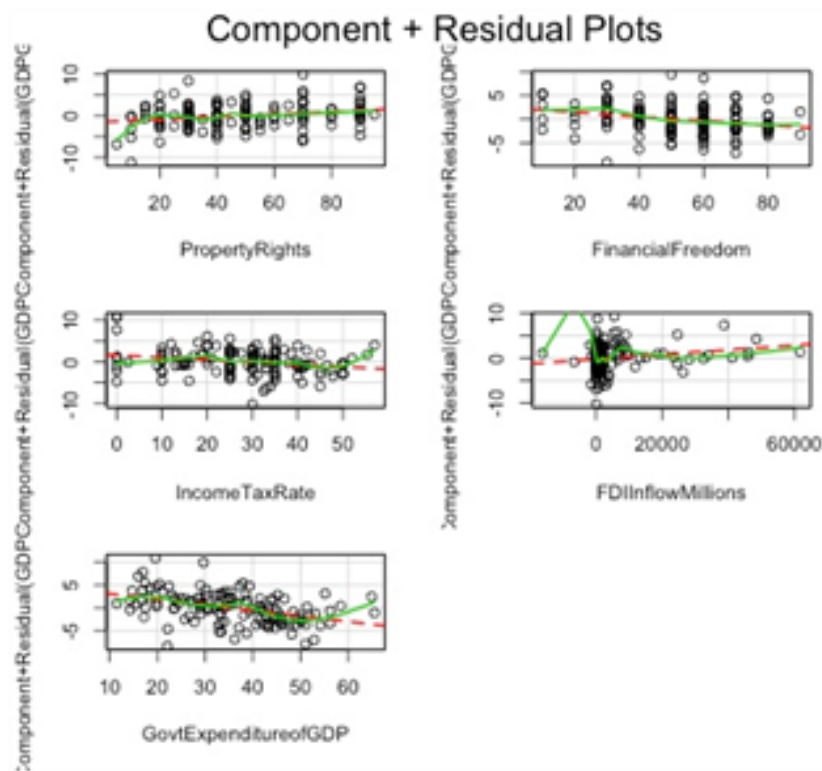
Exemplo:

```
library(car)
lm_fit<-lm(y~x1+x2+x3)
crPlots(lm_fit)
```

Nos exemplos a seguir, o CRPlot diz respeito à linha indicativa do melhor ajuste. Diferenças significativas entre esta linha (em vermelho) e a curva com melhor ajuste não paramétrico (em verde - por *default* o R utiliza o suavizador loess) podem sugerir que não há relação linear com a variável dependente. Neste caso, podem ser feitas alterações para que os resíduos estejam melhor ajustados, dentre as quais, por exemplo,  $x^2$ ,  $1/x$  e  $\log(x)$ .

Obs: Podemos usar uma transformação da classe Box-Cox. Comentaremos isto mais adiante.

Caso nenhuma das transformações funcione, a literatura sugere que se deve considerar a adoção de um modelo não linear.



- Soluções para casos não usuais?

- Observações não usuais podem acontecer devido a uma digitação errada. Neste caso, podemos retificar ou apagar a observação (os métodos descritos irão captar estas observações atípicas).

- Quando tivermos poucas observações atípicas podemos tratá-las separadamente.

- Se tivermos muitas, isto pode ser um indicativo de que o modelo foi mal especificado (por exemplo, uma variável importante que responde pela explicação delas pode ter sido omitida – chamamos isto de “problema da variável omitida”. Isto afeta a hipótese 3 de um modelo clássico de regressão linear e fará com que os estimadores não tenham boas propriedades estatística – serão viesados e inconsistentes).

- Devemos manter estas observações no modelo, a não ser que tenhamos sérias razões para removê-las (digitação errada ou casos que sabemos que dificilmente irão se repetir).

Exemplo: o Brasil adotou um sistema de câmbio flutuante em 15/01/1999 e o índice Ibovespa aumentou 33.4% neste dia - é algo fora do padrão que não deverá se repetir comumente). Podemos utilizar outros métodos alternativos ao estimador de mínimos quadrados ordinários que irão lidar melhor com estas observações (regressões robustas dão um peso menor aos *outliers*).

- Frequentemente, os métodos robustos encontram resultados similares ao método de MQO quando omitimos os casos influentes, porque eles dão um peso pequeno a estas observações, quando comparados ao método tradicional.

- Existem inúmeros métodos de regressão robusta. Um dos mais conhecidos é a regressão onde minimizamos a soma dos desvios absolutos dos resíduos (LAD - *least absolute deviations*, isto é, minimizamos em relação aos parâmetros a soma  $\sum_{i=1}^n |u_i|$ ). O método clássico minimiza a soma dos quadrados dos resíduos, dando assim um peso maior às observações discrepantes do que este método. No entanto, podemos mostrar que agora estaríamos estimando (ao minimizar esta norma) a **mediana condicional** de  $Y$  dado os  $X$ 's (quando minimizamos  $\sum_{i=1}^n u_i^2$  estamos modelando a **média condicional** de  $Y$  dado os  $X$ 's).

### Resumo de Diagnósticos para Observações Influentes

- Amostras pequenas são especialmente vulneráveis a observações atípicas
- é difícil combatê-las neste caso.

- Resultados baseados em “grandes” amostras também podem ser afetados.

- Mesmo no caso em que a amostra é grande e as variáveis estão em uma amplitude limitada, uma digitação errada, por exemplo, pode ainda influenciar o modelo.

- Casos atípicos são somente influentes quando eles são atípicos tanto em  $X$  (matriz de dados - pontos de alavancagem) quanto em  $Y$  dados os valores de  $X$  (*outliers*).

- Podemos testar se uma observação é um *outlier* utilizando resíduos “studentizados” e gráficos QQ.

- Alavancagem é explorada analisando a matriz  $H$ .

- Influência é avaliada usando DFFITS, D de Cook, DFBETAS e COVRATIO, entre outras medidas.

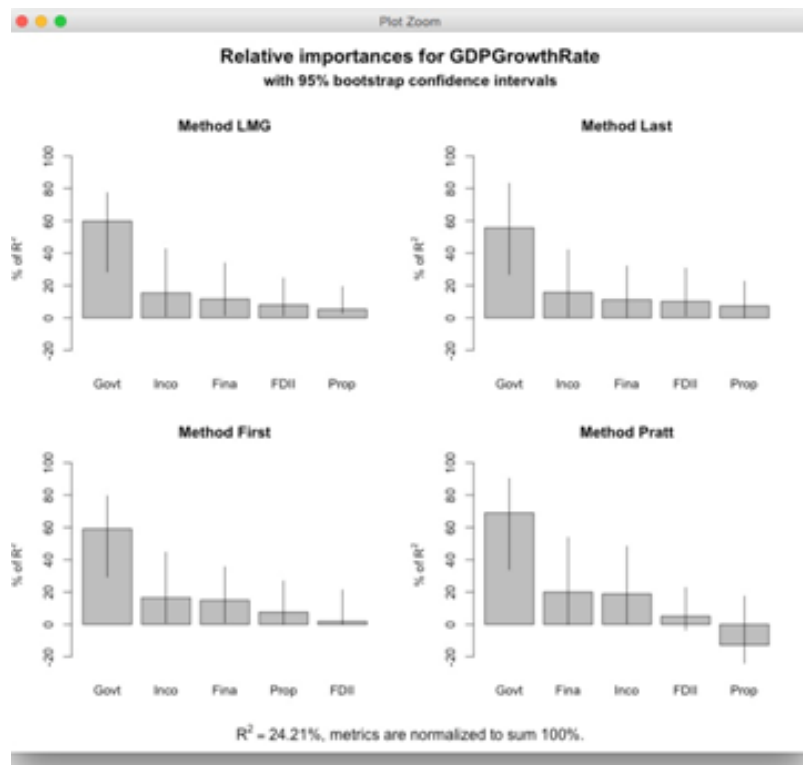
- Gráficos de Influência (*influence plots* ou *bubble plots*) são úteis porque mostram os resíduos “studentizados”, os valores da diagonal da matriz  $H$  e as distâncias de Cook no mesmo gráfico, isto é, uma medida de cada grupo (*outlier*, alavancagem e influência).

- Influência conjunta é melhor avaliada utilizando “gráficos de regressão parcial (*Added-Variable Plots* ou *partial-regression plots*).

(c) Importância Relativa das Variáveis (por exemplo, podemos usar a função `relaimp` no R).

- Fornece uma decomposição da variância explicada pelo modelo (contribuição de cada regressor).

Exemplo:



- Em resumo, precisamos verificar a validade (ou melhor, se não conseguirmos refutar) os pressupostos do modelo. Para tal, precisamos fazer uma análise residual. Caso contrário, nossas interpretações acerca dos resultados podem ser completamente equivocadas.

- A verificação das observações atípicas/influentes é outra parte importante a ser considerada.

- A validade do modelo estrutural também precisa de análise (seleção dos regressores, linearidade, etc.). No entanto, há críticas de que este procedimento é *data driven*, isto é, haverá mudanças na seleção se utilizarmos outros bancos de dados com as mesmas variáveis.

## Validação do Modelo

Esta parte envolve previsão dentro e fora da amostra. A primeira considera toda a informação disponível na amostra para estimar o modelo e, então,

e então analisa o seu poder preditivo em relação às observações dentro da própria amostra. A segunda utiliza informações de uma parte da amostra para estimar o modelo e então realiza previsões para o restante das observações (em economia chamamos de previsão dentro e fora da amostra - em computação/*machine learning*/inteligência artificial é comum chamarmos de amostra de treinamento e amostra de teste quando queremos fazer previsões fora da amostra).

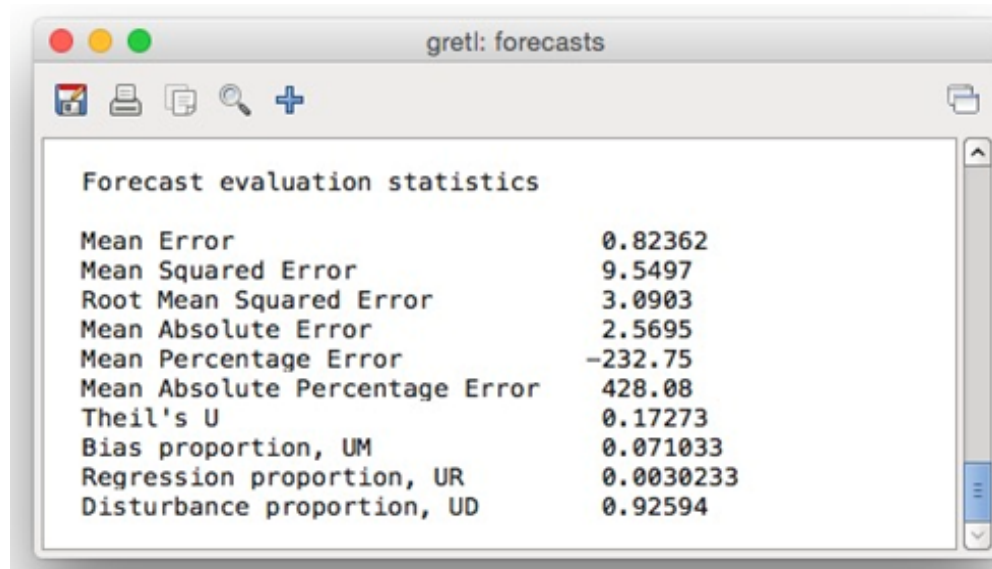
Após a elaboração, adaptações e testes nós partimos para a validação do modelo. É comum particionarmos em duas partes como descrevemos no parágrafo anterior: a quebra usualmente é feita de maneira ad hoc (85, 90% das observações para modelagem dentro da amostra, por exemplo, e 10, 15% para fazer a previsão).

Para avaliar os resultados costuma-se usar uma gama de critérios. Em economia/econometria, são usuais, por exemplo:

- U de Theil
- Proporção da Perturbação (UD)
- MAPE: erro percentual absoluto médio
- MPE: erro percentual médio, etc..

UM (proporção do viés) e UR (proporção da regressão), idealmente, devem ser tão próximos de zero quanto possível e UD, idealmente, o mais próximo de 1 (a soma das 3 componentes é 1).

Exemplo:



The image shows a screenshot of a software window titled "gretl: forecasts". The window has a standard macOS-style title bar with red, yellow, and green buttons. Below the title bar is a toolbar with icons for file operations (save, print, copy, paste, search, and a plus sign) and a scroll bar on the right. The main content area displays "Forecast evaluation statistics" in a monospaced font. The statistics are listed in two columns: the metric name on the left and the numerical value on the right.

Forecast evaluation statistics	
Mean Error	0.82362
Mean Squared Error	9.5497
Root Mean Squared Error	3.0903
Mean Absolute Error	2.5695
Mean Percentage Error	-232.75
Mean Absolute Percentage Error	428.08
Theil's U	0.17273
Bias proportion, UM	0.071033
Regression proportion, UR	0.0030233
Disturbance proportion, UD	0.92594