Econometria com R: Séries Temporais
- Aula 1 -

Prof. Mestre. Omar Barroso

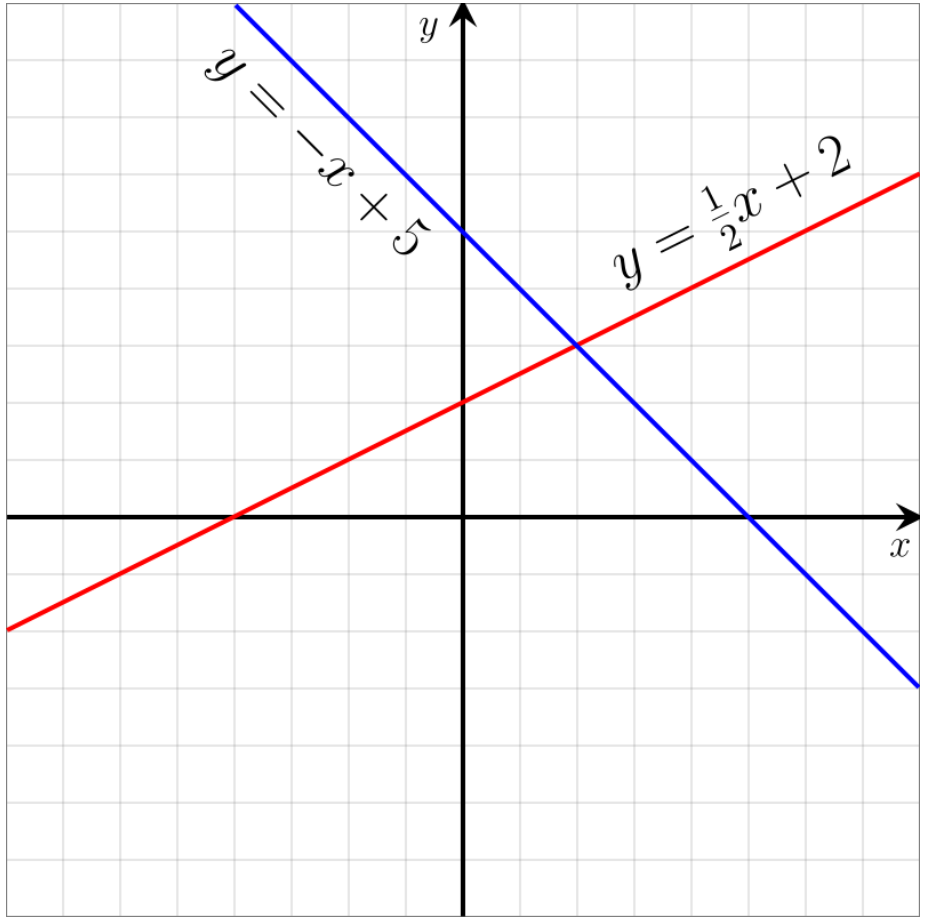
Instituto Brasileiro de Educação, Pesquisa e Desenvolvimento

Terminologia da Regressão Linear Simples (RLS)

- $y = \alpha + \beta_1 x_i + \varepsilon$ (1)
- $i = Número\ de\ observações\ , i = 1, \dots, n$
- $y_i = Váriavel\ \boldsymbol{Dependente}\ (a\ ser\ explicada)$
- $x_i = Váriavel\ \boldsymbol{Independente}\ (a\ que\ explicará\ y)$
- $\alpha = O\ intercepto\ da\ linha\ de\ regressão$
- $\beta_1 = O\ \boldsymbol{coeficiente}$
- $\varepsilon = Termo\ de\ \boldsymbol{Erro}$
- Lembre-se: a equação (1) representa a regressão linear da **população** ou θ

Metodologia da Regressão Linear Simples

- A RLS não deixa de representar uma **função linear** entre a variável **X e Y**. Ou seja, X explica os efeitos que influenciam Y.
- Lembrem-se a RLS é nada mais do que uma **função polinomial** no qual α e β são constantes e números reais.
- Se $\alpha > 0 = inclinação\ positiva; c.c.\alpha < 0 = Inclinação\ negativa$.



Fonte: Symbolab

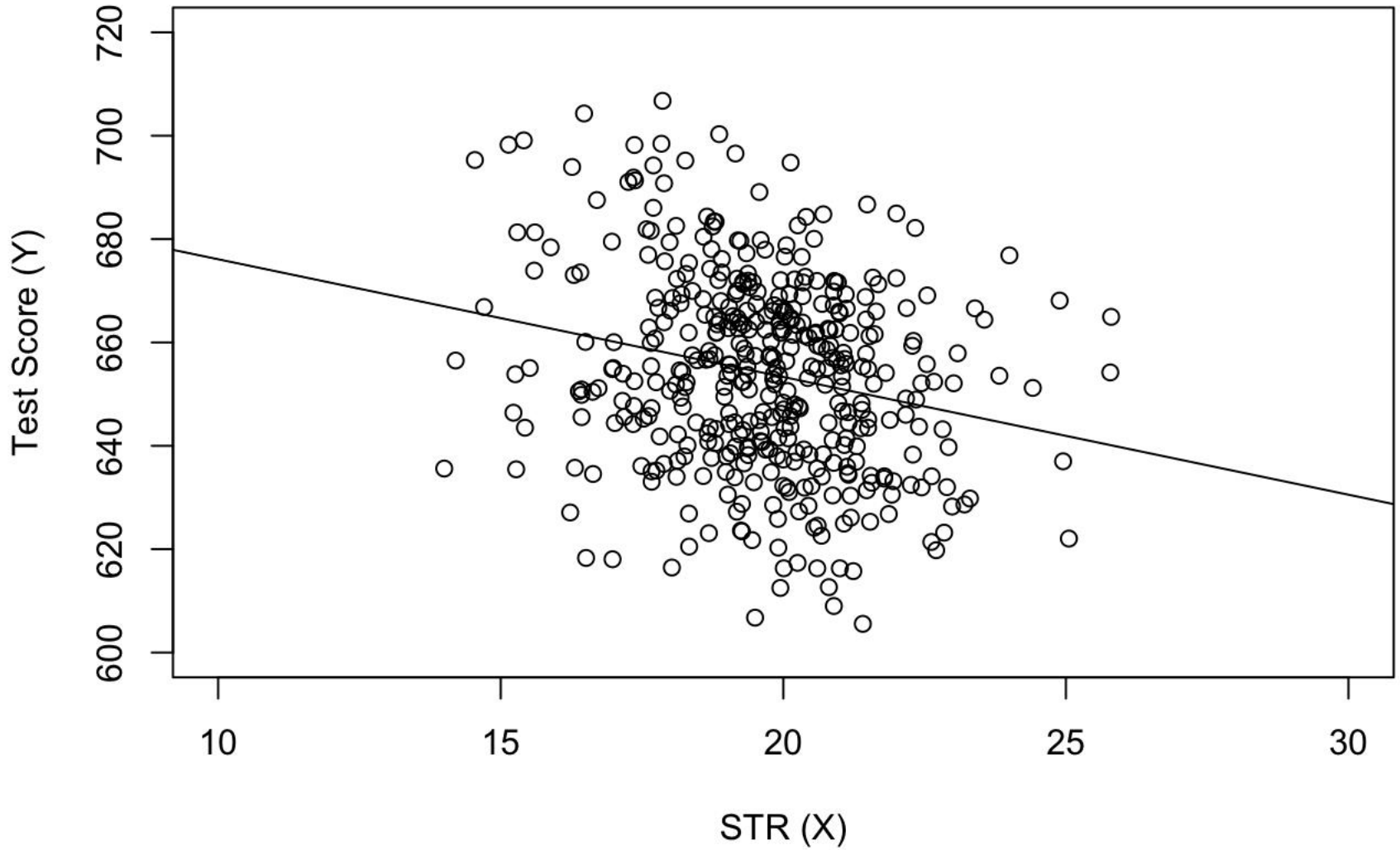
Causalidade (isso existe ?)

- Podemos sugerir que X explica Y na RLS: $y = \alpha + \beta_1 x_i + \varepsilon$.
- Com isso, podemos ASSUMIR que na RLS, **X DETERMINA Y**?
- **NÃO!** Pois o termo de 'erro' ε explica os erros no model, fatores desconhecidos ou simplesmente fatores muito complexos e fora de um contexto matemático. Desta maneira, a regressão **não é** um modelo **determinista** e sim **estocástico**.
- Com isso, como economistas (ou analistas financeiros) o melhor que podemos dizer é que o modelo confere com *Ceteris Paribus*.

Determinismo Vs. Estocástico

- Imagina uma competição de tiro esportivo...
- A regressão **determinista** $y = \alpha + \beta_1 x_i$ seria igual a um **tiro preciso** no meio do alvo.
- Contrariamente, a RLS com o **termo de erro** $(y = \alpha + \beta_1 x_i + \varepsilon)$ seria o equivalente a uma espingarda. No qual, as balas explodem em lugares aleátorios.

Fonte: Daily Mail


Test Score (Y)
600
620
640
660
680
700
720
10
15
20
25
30
STR (X)

Ceteris Paribus

- Na econometria, *ceteris paribus* é um termo latino que significa "t*odo o resto sendo igual*" ou "*mantendo as outras coisas constantes*".
- É um conceito crucial usado para **isolar o efeito** de uma variável de interesse específica, assumindo ao mesmo tempo que todos os outros fatores relevantes permanecem inalterados.
- No contexto de uma análise de regressão, quando dizemos “ceteris paribus”, estamos afirmando que a **relação estimada** entre a variável independente e a variável dependente é verdadeira, assumindo que não existem alterações noutros fatores influentes.
- Devemos notar que a suposição “ceteris paribus” é muitas vezes uma simplificação, e os econometristas/pesquisadores precisam de considerar cuidadosamente potenciais variáveis omitidas ou fatores não observados que possam influenciar as relações estimadas.

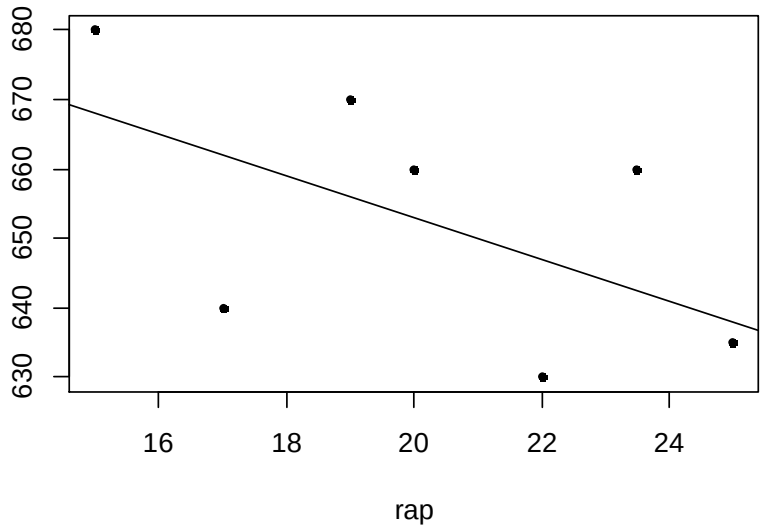
(RLS)Populacional Vs. (RLS) Estimada

- Podemos estimar a RLS baseada na população “inteira”?
- Não... Seria impossível obter dados de uma população inteira. Por exemplo, no CENSO Brasileiro, é inconcebível para o IBGE entrevistar e coletar dados de todos os cidadãos Brasileiros que residem no vasto território nacional.
- Com isso, utilizamos um “*corte*” da população ou uma **amostra**.
- Estimar parâmetros de regressão a partir de uma amostra permite **inferência estatística**. Podemos calcular **intervalos de confiança** e realizar para fazer declarações sobre a variação provável dos parâmetros e **testes de hipóteses** populacionais.

Vamos fazer a nossa regressão e discutir o que foi dito...

Vamos utilizar os seguintes códigos

- *# criando dados amostrais*
- *# razão alunos p/ professores*
- rap <- c(15, 17, 19, 20, 22, 23.5, 25)
- notas <- c(680, 640, 670, 660, 630, 660, 635)
- *# criando um gráfico de dispersão*
- plot(notas ~ rap, ylab="",pch=20)
- # RLS teórica # y = a + b1xi+e#
- #ajustando a linha
- abline(a = 713, b = -3)



Mínimos Quadrados Ordinários (MQO)

- O estimador MQO tem como o objetivo é encontrar os coeficientes $(\beta's)$ de um modelo de regressão linear que **minimizem** a soma dos quadrados das diferenças entre os valores observados e os valores previstos pelo modelo.
- A aproximação é mensurada pela soma do quadrado dos "erros" em prever Y dado X.
- A ideia básica por trás do MQO é encontrar a **'melhor' linha** (ou hiperplano em dimensões superiores) que minimiza a **soma das diferenças quadradas** entre os valores observados (resultados reais) e os valores previstos pelo modelo linear. Os erros, também conhecidos como resíduos, são as diferenças entre os valores observados y e os valores previstos $\hat{y}$.

$$(2) SDQ = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Mínimos Quadrados Ordinários (MQO)

- Os estimadores OLS do coeficiente $\hat{\beta}_1$ e do intercepto α no modelo de regressão linear simples (3) .
- O numerador representa a **covariância**, ou seja, a soma dos produtos cruzados dos desvios de cada ponto de dados em relação às suas respectivas médias (tendência da relação linear das variáveis).
- O denominador representa a variância, ou seja, a soma dos desvios quadrados da variável independente (x) da sua média (a dispersão de x).
- O cálculo da razão entre a covariância e a variância, essencialmente normaliza a covariância pela variabilidade em X.
- Esta normalização é crucial porque **dimensiona a covariação para ser independente** da escala de X. Isso resulta em uma medida que reflete a mudança em **Y por mudança unitária em X**, tornando-o interpretável como a inclinação da reta de **melhor ajuste**.

- $(3)\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=2}^{n}(x_i-\bar{x})^2}$

Coeficiente de Determinação (R^2)

- O R^2 representa a fração **da variância amostral de y_i explicada por x_i**. Em outras palavras, o R^2 pode ser entendido como a razão da **soma explicada dos quadrados** (SEQ) para a **soma total dos quadrados** (STQ).
- A SEQ é a soma dos desvios quadrados dos valores previstos $\hat{y}_i$da média de y_i. Já a soma dos totais quadrados (STQ) é a soma dos desvios quadrados de Yi (sobre a sua média).
- $SEQ = \sum_{i=1}^{n}(\hat{y}_i - \bar{y}_i)^2$
- $STQ = \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$
- $R^2 = \frac{SEQ}{STQ}$

Coeficiente de Determinação (R^2)

- O R^2 fica entre 0 e 1. Sendo 1 (perfeito e útopico) ou seja, nenhum erro cometido ao ajustar a linha de MQO.
- Soma dos Resíduos Quadrados: $SRQ = \sum_{i=1}^{n} \hat{e}_i^2$
- uma medida para os **erros cometidos** ao prever y por x.

Hipóteses Importantes

- H1: A regressão é **linear** sobre os **parâmetros**. Em outras palavras, a relação sobre **Y** e a variável(is) **independente(**s) {x1, x2,...,xn} é linear. Ou seja, $y = \alpha + \beta_1 x_i + \varepsilon$.
- H2: Os erros não devem ser sistemáticamente relacionados (ortogonais) com os valores de x, $Cov(x_i \varepsilon_i) = 0$. Caso existir uma relação $Cov(x_i \varepsilon_i) \neq 0$, teremos um problema de endogeneidade.
- Isso segue o conceito de ortogonalidade: $\sum_{i=1}^{n} x_i \varepsilon_i = 0$.

Hipóteses Importantes

- H3: O Valor Esperado (a média) dos erros de um valor específico de uma variável independente deve ser zero. Ou seja, $E(\varepsilon_i) = 0$.
- Essa hipótese garante que o modelo seja linear e não seja viésado. No qual, pode ser sugerido que o valor esperado do parâmetro seja igual ao valor ‘real’.
- A H3, sugere que os efeitos sistémicos são capturados pelas variáveis independentes.
- E os erros sejam os únicos afetados pela aleatóriedade.

Hipóteses Importantes

- Ou seja, H3 cumpre o teorema de Gauss-Markov e pode ser definido como ‘**Best Linear Unbiased Estimator (BLUE)** ‘.
- De acordo com o teorema de Gauss-Markov, se os erros têm média zero e são homocedásticos (variância constante) e não correlacionados, os estimadores MQO são os ‘melhores’ estimadores lineares não enviesados (variância mínima).

Assumindo Questões do MQO

- Assumindo que H1 e H3 estão presentes, $(X_i Y_i), i = 1, \ldots, n$ são amostras extraídas **independentes e identicamente distribuídas (iid**). de sua distribuição conjunta.
- Com isso, assumimos que valores discrepantes (outliers) serão improváveis ou ‘administrados’.
- A maioria dos esquemas de amostragem usados na coleta de dados de populações produzem amostras **i.i.d**.

Assumindo Questões do MQO

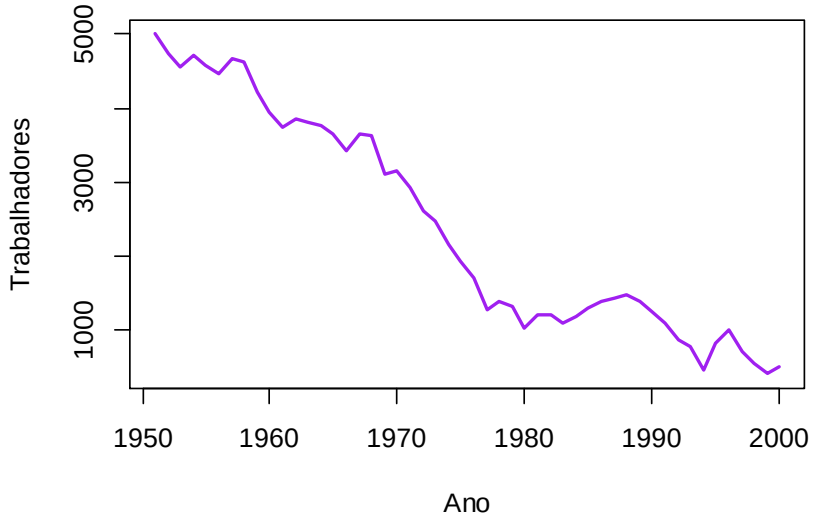
- Refere-se a um conjunto de variáveis aleatórias que são estatisticamente independentes umas das outras e possuem distribuições de probabilidade idênticas.
- Independência significa que o resultado de uma variável aleatória não afeta o resultado de outra.
- Distribuído de forma idêntica significa que cada variável aleatória no conjunto segue a mesma distribuição de probabilidade.

Exemplo IID

- A suposição que a **IID** não é cumprida são os **dados de séries temporais em que temos observações na mesma unidade ao longo do tempo**.
- Por exemplo, considere ‘X’ como o número de trabalhadores numa empresa de produção ao longo do tempo.
- Devido às transformações nos negócios, a empresa corta empregos periodicamente por uma parcela específica, mas também existem algumas **influências não determinísticas** relacionadas ao cenário macroeconômico (crises políticas, guerras, pandemias etc.).
- Usando R podemos facilmente simular tal processo e representá-lo...

Exemplo IID usando R

- $Employment = \alpha + \beta_1 Employment_{t-1} + \varepsilon_t$
- Iniciamos a série com um total de 5.000 trabalhadores e simulamos a redução do emprego com um **processo autorregressivo** que apresenta um movimento descendente no longo prazo e tem erros normalmente distribuídos.
- É evidente que as observações sobre o número de empregados **não podem ser independentes neste exemplo:** o nível de emprego de hoje está **correlacionado** com o nível de emprego de amanhã. Assim, o **i.i.d.** suposição é violada (discutiremos isso mais a diante).



Questões Para Revisão (Não Precisa Entregar)

- Qual é a relevância do termo de erro em relação entre um processo determinista vs. estocástico?
- Qual é a diferença entre uma amostra populacional e estimada? Demonstre e explique a relação as variáveis (dependentes e independentes).
- Qual é o propósito do estimador MQO em uma regressão linear?
- Quais são as hipóteses importantes (H1, H2 e H3)?
- Explique as amostras independentes e identicamente distribuídas (iid).

Referências

- Kleiber, C., & Zeileis, A. (2008). Applied Econometrics with R. Springer.
- Wooldridge, J. M. (2013). Introductory Econometrics: A Modern Approach (5th ed.). South-Western, Cengage Learning. ISBN-13: 978-1-111-53104-1.

Obrigado!