MODELO BÁSICO DE REGRESSÃO LINEAR

stecine.azureedge.net/repositorio/modelo_basico_de_regressao_linear/index.html

OBJETIVOS

Descrever a área de econometria e a abordagem empírica para problemas socioeconômicos;

Introdução

Toda teoria deve gerar previsões, que devem, por sua vez, ser avaliadas empiricamente. Para isso, utilizamos econometria.

Econometria consiste no uso intenso de ferramentas da matemática e da estatística para entender melhor questões sociais e econômicas que nos rodeiam.

Neste tema, introduziremos duas abordagens utilizadas para esse instrumental. Além disso, também introduziremos o **método de estimação dos mínimos quadrados ordinários — ferramenta básica para a análise empírica, útil em todas as áreas do conhecimento**.

Usaremos conceitos de cálculo multivariado e estatística ao longo do tema. Vale revisar alguns resultados básicos, como definições de esperança estatística, variância, covariância, correlação, funcionamento do operador de somatório, lei dos grandes números, teorema central do limite e lei das expectativas iteradas.

Estudo de econometria

Neste módulo, discutiremos um pouco da intuição do estudo de econometria e como estruturar uma análise econométrica.

Relação entre variáveis sócioeconômicas

O objetivo principal da econometria enquanto área de pesquisa é encontrar uma relação causal entre variáveis econômicas ou sociais de interesse, ou seja, o impacto de uma variável sobre outra. Um objetivo secundário, mais fraco, é o de verificar associações (causais ou não) entre variáveis.

O que é uma relação causal?

Há vários exemplos simples no nosso dia a dia, vejamos:

Exemplo

Observe com cuidado a estrutura dessas frases: podemos identificar relações de causa e efeito. Observo dois eventos — **topada com o pé e dor no pé** — e consigo dizer que um foi causado pelo outro: não é mera coincidência.

Dificuldades na busca de relação causal

Nem sempre é fácil identificar relações causais. Não basta que dois eventos pareçam relacionados, é necessário descartar outras causas possíveis.

Se uma pessoa doente toma um remédio e fica boa, será que foi o remédio que a curou ou ela se recuperaria em qualquer hipótese? Será que o remédio pode até ter atrasado essa recuperação? Há vários outros fatores que afetam a saúde de um indivíduo e precisamos isolar o impacto do remédio.

Se o governo implementa uma política pública e o desemprego cai, será que a política foi a responsável pela queda do desemprego?

Vejamos, a seguir, alguns exemplos de perguntas que podem ser respondidas com econometria:

- Um programa que oferece treinamento para funcionários de uma fábrica aumenta a produtividade desses funcionários?
- O uso de drogas lícitas, como álcool ou cigarro, durante a gestação aumenta a probabilidade de que bebês nasçam com problemas de saúde?
- Leis que punem criminosos de maneira mais severa reduzem a taxa de homicídios?
- Uma transferência do governo federal para um município gera aumento de gastos em educação nesse município?

Desafios do pesquisador

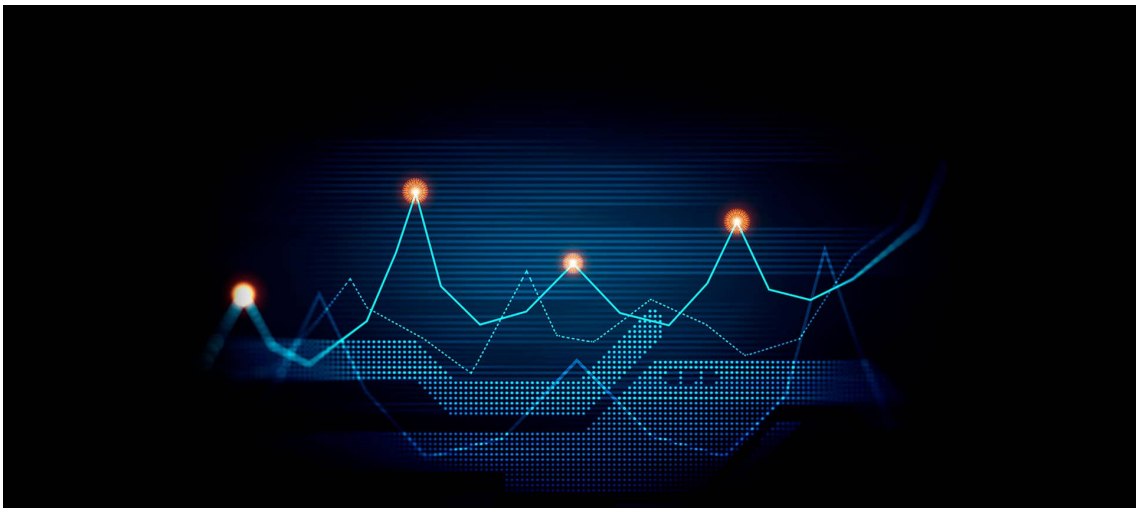
Neste módulo, vamos apresentar quais são as preocupações que um pesquisador que utiliza dados — na academia ou no mercado — deve ter ao iniciar o estudo de relações de interesse.

Vale reforçar que é fundamental distinguir a correlação entre as variáveis da causalidade entre elas. O nosso desafio enquanto pesquisadores que procuram investigar fenômenos sociais com dados será sempre o de encontrar uma relação de causa e efeito entre as variáveis de nosso interesse, usando uma amostra da população.

Neste módulo, cobriremos três tipos distintos de abordagem empírica em econometria:

Na abordagem estrutural, partimos de um modelo e, a partir dele, explicitamos as relações econômicas/sociais que queremos investigar empiricamente. Já a abordagem de forma reduzida com dados experimentais conta com diversas vantagens que facilitam a obtenção de causalidade entre as variáveis, pois o econometrista gera exatamente os dados de que precisa em um experimento específico. Já na abordagem de forma reduzida com dados observacionais (ou simplesmente ‘observados’), os dados não coletados diretamente pelo econometrista para o fim específico, o que impõe alguns desafios adicionais para o pesquisador em comparação à abordagem com dados experimentais.

Vamos estudar essas abordagens.



Abordagem estrutural

Em sua apresentação tradicional, a Econometria consiste na aplicação do instrumental de Matemática e Estatística em dados do mundo real, com o objetivo de responder perguntas econômicas.

Tal definição é, nos dias de hoje, uma referência histórica, pois a Econometria usa métodos frequentes em Ciências Exatas e os aplica a questões em quaisquer ciências — da Biologia à Sociologia, da Psicologia à Engenharia, da Pedagogia à Ciência Política.

Parte das aplicações da Econometria se dão em Ciências Sociais pela característica dos dados disponíveis nessa área. Teremos, no entanto, uma interpretação ampla: Ciências Sociais incluem, por exemplo, a análise de preços de ações na bolsa de valores — objeto também estudado pela Engenharia.

Modelo econométrico

Usaremos, a partir daqui, a apresentação tradicional da Econometria.

Tenha sempre em mente, no entanto, que a econometria se trata de uma ferramenta presente em todas as áreas do saber.

Partimos, nesse caso, de perguntas de natureza socioeconômica. Em seguida, analisamos modelos matemáticos que ilustram a relação entre as variáveis relevantes para essas perguntas. Por fim, usamos dados do mundo real e ferramentas estatísticas para estimar as relações de interesse e encontrar respostas.

A partir desse processo, teremos o nosso modelo econométrico. Para estimar o modelo econométrico, partimos de uma **amostra de dados**.

Amostra de dados

Uma amostra de dados nada mais é do que uma sequência de variáveis aleatórias.

Este fato deixa claro por que precisamos da estatística: o instrumental estatístico permite realizar testes de hipótese sobre os dados a partir dos quais tiraremos conclusões.

Teoria do consumidor

Veremos, agora, a abordagem estrutural com mais de detalhes.

Iniciaremos nossa exposição utilizando um tópico familiar aos economistas: a teoria do consumidor.

Na abordagem utilizada por economistas para entender o comportamento de consumidores, por exemplo, os indivíduos fazem escolhas ótimas de consumo diante dos preços e dada a sua renda.

As escolhas ótimas de consumo são simplesmente as quantidades demandadas (nossa variável dependente), vistas como funções dos preços de mercado e da renda do consumidor (as variáveis explicativas).

Chamamos essas relações de **funções de demanda**

Funções de demanda

Se você já estudou a teoria do consumidor, essa é a função de demanda marshalliana, que depende de preços e renda (em oposição à demanda hicksiana, que depende de preços e utilidade).

Em suma, partimos de um modelo teórico sobre o comportamento do consumidor e derivamos as **equações de demanda**. Temos, então, fórmulas matemáticas que representam a demanda do consumidor e mostram como essa demanda depende de fatores como preços e renda.

Vamos ilustrar isso com um exemplo a seguir:

Equações de demanda – exemplo

Suponha que nosso interesse seja analisar o consumo de arroz no Brasil. A quantidade de arroz demandada pelos consumidores, de acordo com a teoria do consumidor, vai depender:

- Do preço do arroz;
- Do preço dos produtos substitutos e complementares do arroz;
- Da renda daqueles consumidores;
- De outras características desses mesmos consumidores.

Por exemplo, podemos imaginar que o consumo de arroz depende do preço do feijão (que é complementar ao arroz) e do macarrão (que é substituto).

A demanda é dada por:

\

$$q^*arroz(pfeijão,pmacarrão,...,R,\succeq)$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que q*arroz é a quantidade demandada de arroz, que é função:

- Dos preços de todos os bens consumidos por esse indivíduo;
- Da renda R do indivíduo;
- De suas preferências $\succeq$, que representam as características individuais dos consumidores (que podem gostar muito de arroz ou, talvez, prefiram massas).

Nosso modelo econômico parte de uma relação entre a quantidade demandada de arroz pelos consumidores e as variáveis que mencionamos: preço do arroz, preço do macarrão, renda individual e as características do consumidor.

Vamos olhar uma versão particularmente simples. O modelo econométrico linear para a quantidade demandada por arroz, se assumíssemos que sua demanda depende somente do seu próprio preço e da renda, seria, por exemplo:

$$q^*arroz=\beta0+\beta1parroz+\beta2R+u$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

O efeito do preço do arroz sobre a quantidade demandada é dado por β_1. Já o efeito da sua renda é dado por β_2.

Aqui, estamos supondo que o modelo teórico da demanda é verdadeiro.
Esperamos, em nossas estimações, encontrar:

- Um valor de β_1 negativo (maior preço, menor quantidade demandada);

- Um valor de renda positivo (maior renda, maior quantidade demandada).

Atenção

Abordagem estrutural

Vamos, agora, organizar os passos de uma abordagem estrutural.

Passo 01

O primeiro passo nesse processo é a formulação da questão de interesse para o pesquisador, isto é, qual é a variável X que impacta a variável Y.

No exemplo apresentado para a teoria do consumidor, um dos interesses do pesquisador pode ser estimar algum parâmetro com importância dentro dessa teoria. Um exemplo de parâmetro seria a **elasticidade-preço** da demanda, ou seja, quão responsiva a demanda por arroz é a mudanças no preço desse produto.

Elasticidade-preço

A elasticidade-preço nos informa qual é a variação percentual da quantidade consumida quando o preço aumenta 1%.

Passo 02

O segundo passo, após a formulação da pergunta de interesse, é desenvolver o nosso modelo teórico. Na abordagem estrutural, esse modelo vem da teoria do consumidor.

A teoria do consumidor, entretanto, diz que existem outras variáveis que influenciam a quantidade demandada além do preço. Por conta disso, precisamos formular a equação de interesse de maneira completa, incluindo todos os fatores que podem afetar o preço.

Passo 03

O terceiro passo é, a partir desse modelo teórico, construir o modelo econométrico que o representa. No exemplo, podemos ver que Y será a variável dependente — no exemplo do arroz, a quantidade demandada; e X será a variável explicativa (ou independente), a variável cuja relação com Y queremos medir.

Ou seja, **queremos entender como X causa Y**. Essa equação pode ser representada da seguinte maneira:

$Y = \beta_0 + \beta_1 X + \beta_2 Z + u$

No exemplo da demanda de arroz, X seria o preço do arroz. Podemos ter outras variáveis que são expressas pela variável Z incluída no nosso modelo — por exemplo, a renda dos consumidores ou o preço do macarrão.

Note que, na relação econométrica, temos um termo a mais, o termo u. Tal termo será o termo de erro ou **todos os outros fatores que explicam Y e que não estão incluídos em X e Z**.

Já os coeficientes β_1 e β_2 são os **parâmetros da regressão**. Se estivermos particularmente interessados na relação entre preço e demanda, β_1 será o **parâmetro de interesse**.

Passo 04

O quarto passo seria obtermos uma amostra da população que seja útil para estimar nosso modelo, isto é, para mensurar a relação entre Y e X. Isso é necessário para podermos estimar nossos parâmetros discutidos no passo anterior.

Uma vez estimados parâmetros com os dados, podemos partir para o último passo: fazer testes de hipótese sobre eles (mas isso é assunto para outros temas).

Discutiremos, no segundo módulo, como obter os parâmetros de interesse. De todo modo, é crucial ter em mente o tipo de abordagem a escolher, para proceder com a análise em seguida.

Abordagem de forma reduzida

Na abordagem de forma reduzida, o interesse do pesquisador é responder uma pergunta específica sobre como duas variáveis se relacionam.

Exemplo

Na abordagem de forma reduzida, portanto, não temos, necessariamente, um modelo teórico por trás dessa decisão dos gerentes.

Idealmente, o efeito desse treinamento seria calculado pela diferença entre a produtividade do funcionário que teve o treinamento e outro funcionário, semelhante em todos os aspectos, que não foi treinado.

Gostaríamos, na verdade, de observar um mesmo trabalhador que, ao mesmo tempo e nas mesmas condições, recebeu e não recebeu treinamento: qualquer diferença em sua produtividade seria causada somente pelo treinamento. Isso, porém, é uma impossibilidade física. Esse é o problema fundamental da inferência causal, pois sempre observaremos a mesma unidade em apenas uma situação: em um dado momento, essa pessoa foi treinada ou não foi.

A seguir, veremos isso de um modo um pouco mais formal:

Exemplo – produtividade do trabalhador

Sejam y1i e y0i a produtividade do trabalhador i com ou sem treinamento, respectivamente. Esses são conhecidos como os resultados potenciais da intervenção, também chamada de tratamento.

Assim, y1i é a produtividade de i sob tratamento e y0i é a produtividade sem tratamento. Chamaremos o **estado da natureza** onde não há tratamento de **estado de controle**.

Estado da natureza

Trata-se de uma terminologia comum em Economia e Finanças. É simplesmente uma possibilidade entre várias mutuamente excludentes. Por exemplo, se jogamos uma moeda para o alto, há dois estados da natureza possíveis: cara ou coroa.

Cada unidade i tem dois resultados potenciais, mas apenas um resultado observado, isto é, qual foi de fato a realização. O resultado observado é dado por:

$$yi=Diy1i+(1-Di)y0i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que D_i assume valor 1, se a unidade i recebeu a intervenção/tratamento, e 0 caso contrário. O efeito que gostaríamos de obter é o seguinte:

Efeito do treinamento=y1i-y0i

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Como não podemos observar o mesmo trabalhador em duas situações diferentes (ele recebeu treinamento ou não recebeu), **precisamos comparar trabalhadores que receberam e que não receberam treinamento.**

Comparação entre trabalhadores

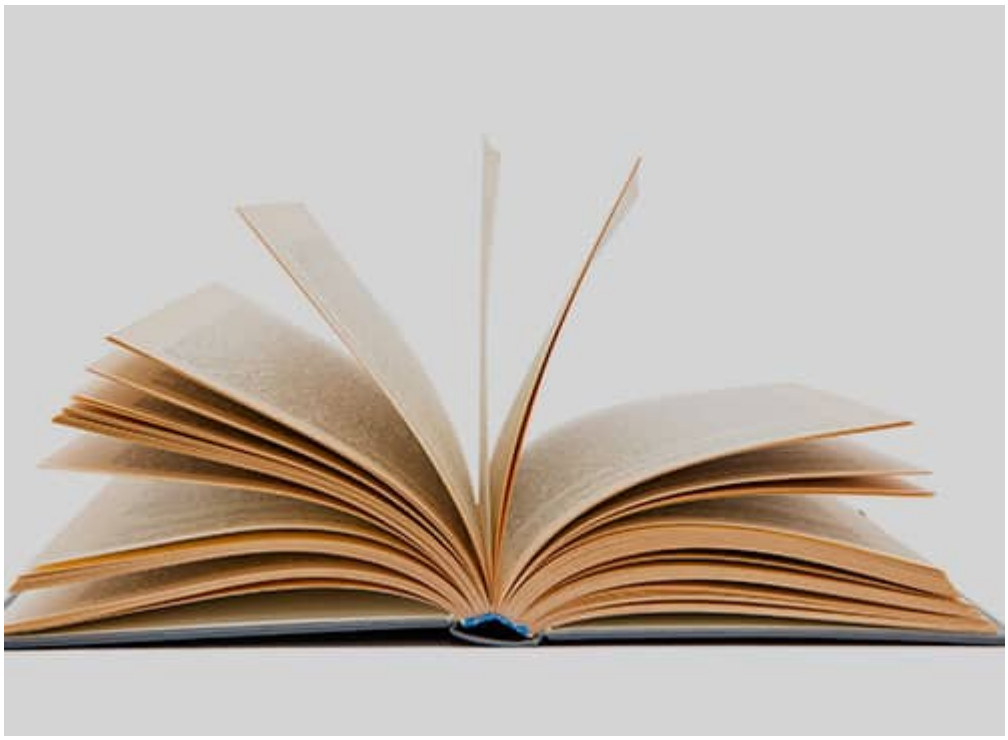
Ao fazer a comparação entre trabalhadores, temos de levar em conta outros fatores que afetam a produtividade desses funcionários. Por exemplo:

Roman Fenton/shutterstock

A temperatura no local onde eles trabalham.

violetkaipa/shutterstock

A qualidade dos equipamentos disponíveis.



Pakhnyushchy/shutterstock

O nível de educação formal deles.

Tais fatores podem ser diferentes entre trabalhadores que receberam treinamento e outros que não receberam. No jargão econométrico, dizemos que temos de **controlar por essas variáveis**, ou seja, filtrar o efeito delas e isolar o efeito da nossa intervenção.

O intuito é comparar trabalhadores que sejam similares e que difiram apenas na exposição ao treinamento. **Isso é o que chamamos de uma análise *ceteris paribus***.

Mantendo todos os demais fatores constantes, qual é o efeito do treinamento?

Resposta

Aleatorização

Ao aleatorizar quem receberá ou não a intervenção, é possível comparar o resultado **médio** desses **grupos** e concluir algo sobre o efeito da intervenção. Em estatística, o resultado médio é obtido com o uso do operador esperança matemática.

Grupos

Os que recebem e os que não recebem treinamento

A lógica disso é que o sorteio, se feito de forma correta, garante que, na média, as variáveis observadas pelas quais controlamos estejam "balanceadas", isto é, são iguais (na média!) entre aqueles que receberam ou não a intervenção.

Podemos, então, escrever:

Efeito médio do treinamento após aleatorização= E[y1i]-E[y0i]

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Como temos aleatoriedade — um sorteio na escolha do tratamento —, ele será independente de Y_i. Veremos, mais adiante, que essa independência entre as variáveis será fundamental para conseguirmos ter o nosso efeito causal.

Na presença de aleatorização, as variáveis y1ie y0i serão independentes entre si. Podemos escrever a esperança condicional em ser tratado ou não como a esperança não condicional.

Dizemos então que há independência na média condicional.

Vamos formalizar essa ideia mais à frente.

A consequência disso é que o efeito médio será obtido pela comparação de médias entre os dois grupos de trabalhadores — os treinados e os não treinados. Esse método permite estimar o impacto do treinamento sobre a produtividade dos funcionários,

independentemente dos outros fatores e insumos que, como vimos, também eram importantes para explicar a produtividade do trabalhador.

Análise de dados observados

Nem sempre é possível “sortear” a intervenção entre as unidades. Pode haver entraves éticos, de custo ou simplesmente queremos avaliar uma intervenção ou relação que já ocorreu no passado e, portanto, está fora do nosso controle desenhar como ela deve ser feita.

Devido às dificuldades de obter dados de natureza experimental, ou seja, obtidos a partir de uma aleatorização, normalmente recorremos a uma análise de **dados observados**.

Os dados observados podem ser:

Dados coletados sem o desenho de experimento por trás;

Dados secundários, coletados por institutos de pesquisa;

Dados administrativos de autarquias ou de algumas empresas públicas ou privadas.

No caso de dados observados, a especificação correta do modelo econométrico é essencial para a obtenção de um efeito que tenha interpretação causal.

Vamos rearranjar a expressão da equação básica de resultados potenciais da seguinte forma:

$$Y_i = D_i y_{1i} + (1 - D_i) y_{0i}$$

$$Y_i = y_{0i} + D_i (y_{1i} - y_{0i})$$

$$Y_i = y_{0i} + \delta D_i$$

$$Y_i = E[y_{0i}] + Y_{0i} + \delta D_i - E[Y_{0i}]$$

$$Y_i = \alpha + \delta D_i + u_i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Vamos explicar passo a passo:

Passo 01

Da primeira para a segunda igualdade, apenas rearranjamos os elementos da equação.

Passo 02

Da segunda para a terceira, escrevemos $\delta = y1i - y0i$

Passo 03

Da terceira para a quarta, apenas “somamos zero” ao adicionar $E[Y0i] - E[Y0i]$ à expressão.

Passo 04

Na última expressão, apenas renomeamos $E[Y0i]$ por α, isto é, o termo constante de intercepto da nossa equação linear, e $Y0i - E[Y0i]$ como o termo de desvio u_i em relação ao valor esperado de $Y0i$ (esse valor esperado é, simplesmente, $E[Y0i]$ — ou seja, estamos usando o operador esperança).

Modelo econométrico simples

Vamos, agora, analisar o modelo econométrico simples da última equação, que estabelece uma relação linear entre a variável de interesse, ou variável dependente (nossa variável Y_i), e a variável explicativa (que é a nossa variável D_i).

Relembrando

Esse foi um exemplo “limpo”, em que apenas ilustramos como podemos sair de uma pergunta para uma especificação na qual cada elemento possui uma interpretação clara.

Vamos ver, a seguir, um exemplo um pouco mais elaborado:

Fatores não observados

Agora, queremos responder a seguinte pergunta:

Mais anos de estudo geram salários maiores no futuro?

Poderíamos usar dados, por exemplo, da Pesquisa Nacional por Amostra de Domicílios, feita pelo Instituto Brasileiro de Geografia e Estatística (IBGE) anualmente, para fazer esse exercício. Nossa equação linear a ser estimada — que chamaremos de regressão linear — será:

$$Salários_i = \beta_0 + \beta_1 Escolaridade_i + u_i$$

(Equação 1)

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Isso significa que, se a escolaridade aumenta em uma unidade (por exemplo, um ano a mais), o salário aumenta, em média, β_1 unidades (por exemplo, quantia adicional recebida por mês no contracheque).

Para que o β_1 seja o efeito causal de uma variação de educação usando anos de escolaridade, como estamos fazendo, sobre os salários dos indivíduos, **teríamos de supor que os demais fatores não observados contidos em u_i, que explicam os salários, se mantêm constantes mesmo após uma variação da educação.**

Em outras palavras, que $\Delta u_i = u'_i - u_i = 0$, , em que u'i é o termo de erro da nossa estimação quando aumentamos a escolaridade em um ano.

É importante entender isso.

Imagine que um aumento de escolaridade seja acompanhado por uma melhora nas condições de saúde — por exemplo, pessoas mais escolarizadas podem aprender sobre hábitos saudáveis e adotá-los.

A melhora nas condições de saúde, por sua vez, pode aumentar a produtividade e, portanto, a renda: pessoas mais saudáveis produzem melhor. Também podemos pensar que algum componente no erro afeta, simultaneamente, a escolaridade e a renda — veremos um exemplo mais adiante.

Na nossa regressão linear, temos apenas a variável **escolaridade**, mas não observamos **condições de saúde** que estão incluídas no termo u_i.

No fim das contas, um aumento da nossa variável escolaridade acaba gerando um aumento em u_i. Ao medir β_1, não estamos capturando apenas o efeito de escolaridade sobre salários, mas o efeito agregado de escolaridade e condições de saúde sobre renda!

Atenção

Vamos fazer mais uma ilustração, a seguir, com um exemplo sobre vacinas.

Exemplo – variável não observada

Imagine que os médicos deem vacinas para as pessoas e apenas meçam o impacto, após algum tempo, sobre as condições de saúde (por exemplo, a resposta imune do organismo). Se as pessoas sabem que foram vacinadas, elas podem pensar: “Estou protegido pela vacina e posso mudar meu comportamento, tornando-me menos cuidadoso”. Esse pensamento afeta a probabilidade de que a pessoa fique doente.

Nesse caso, a vacina tem efeito direto sobre as condições de saúde do paciente — pelo mecanismo biológico que provoca no organismo — e efeito indireto — pela mudança de comportamento que induz.

Por isso, os experimentos médicos são cuidadosos para não informar ao paciente se ele recebeu o medicamento em teste ou um placebo. Nesse exemplo, a variável u_i, não observada, é a mudança de comportamento.

A seguir, vamos ver como isso funciona matematicamente:

Variável u_i

Vamos supor que aumentemos a educação em um ano: chamaremos esse novo nível de educação de Escolaridade'i . Teríamos uma nova estimativa de salários futuros, denotada por Salários'i, dada por:

$$Salários'i = \beta0 + \beta1 Escolaridade'i + u'i$$

(Equação 2)

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Podemos subtrair a equação 1 da equação 2 para obter a associação entre esse aumento e os salários:

$$Salários'i - Salários i = \beta1(Escolaridade'i - Escolaridade i) + u'i - ui$$

$$\Delta Salários i = \beta1 \Delta Escolaridade i + \Delta ui$$

$$\beta1 = \frac{\Delta Salários i}{\Delta Escolaridade i} - \frac{\Delta ui}{\Delta Escolaridade i}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Atenção

Δ é a notação tradicional para variação. Para uma variável X qualquer, $\Delta X = X' - X$. Por exemplo, $\Delta Salários_i = Salários'_i - Salários_i$. Em outras palavras, é a diferença entre $Salários'_i$ (o que se obtém quando o nível de educação é $Escolaridade'_i$) e $Salários_i$ (o que se obtém quando o nível de educação é $Escolaridade_i$).

Desse modo, para que β_1 represente o efeito causal entre anos de escolaridade e salários, devemos ter $\Delta u_i = 0$.

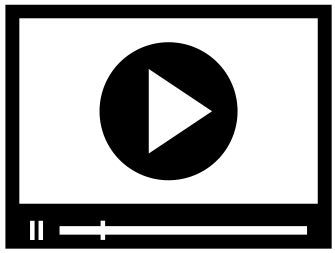
Com isso, todos os demais fatores não observados (isto é, que não conseguimos controlar pela falta de dados) contidos em u_i que explicam $Salários_i$ (a variável dependente) **são considerados constantes** após uma mudança em **$Escolaridade_i$** (a variável explicativa).

Isso significa que não há correlação estatística entre $Escolaridade_i$ e o termo u_i. A mudança observada nos salários futuros, desse modo, deve-se somente à variação nos anos de escolaridade do indivíduo i.

É possível imaginar algumas variáveis importantes que estão presentes em $\mathbf{u_i}$ e podem ser correlacionadas com a educação, como algumas habilidades cognitivas e socioemocionais. Nesse sentido, indivíduos que são mais hábeis obtêm salários maiores — efeito direto de u sobre a variável explicativa Y — e acabam estudando mais — efeito indireto de u sobre a variável explicativa.

Não conseguimos controlar esses fatores a partir dessa regressão. Por conta disso, se estimássemos esse modelo sem garantir que $\mathbf{\Delta u_i = 0}$, estaríamos identificando uma correlação entre educação e salários, e não uma causalidade.

Note que, se os dados foram gerados a partir de um experimento aleatorizado bem-feito, então, $\mathbf{\Delta u_i = 0}$ por construção. Para dados observacionais, no entanto, nem sempre é o caso. Portanto, precisamos ser cuidadosos.



Etapas da análise empírica

Assista, a seguir, a um vídeo sobre as etapas da Análise Empírica:

Verificando o aprendizado

1. O último passo para um desenho de pesquisa utilizando a abordagem estrutural é:

2. Assinale a alternativa correta sobre a abordagem de forma reduzida:

Gabarito
1. O último passo para um desenho de pesquisa utilizando a abordagem estrutural é:

A alternativa **"D "** está correta.

O teste de hipóteses é a última etapa: após obtermos nossas estimativas, precisamos testá-las para saber o grau de confiança que podemos depositar nelas.

2. Assinale a alternativa correta sobre a abordagem de forma reduzida:

A alternativa **"D "** está correta.

A forma reduzida tenta identificar o impacto causal de uma variável explicativa X sobre uma variável dependente Y. A principal hipótese que precisamos fazer para obtermos esse efeito é que os fatores não observados que afetam Y sejam não correlacionados com a variável explicativa.

Avalie este módulo:

Definir conceitos básicos dos estimadores de mínimos quadrados ordinários.

Modelo populacional de regressão linear

Vamos supor que temos uma amostra aleatória qualquer de natureza de **corte transversal**. Essa é uma palavra complicada para dizer que temos observações de vários indivíduos (ou firmas, ou plantas, ou planetas etc.) **no mesmo momento do tempo.**

Corte transversal

Você encontrará com frequência o termo original, em inglês: *cross-section*.

Exemplo

Assuma que temos duas variáveis, x e y, e gostaríamos de verificar como y varia quando mudamos x. Note que, aqui, estamos sendo generalistas e não estamos assumindo nenhuma relação causal entre x e y. Nesse caso, apenas queremos verificar como essas duas variáveis se movimentam juntas.

Existem três perguntas que surgem imediatamente:

1. E se y é afetado por outros fatores que não x?
2. Qual é a forma funcional que relaciona essas duas variáveis?
3. Se estamos interessados no efeito causal de x em y, como podemos distingui-lo de uma mera correlação?

Modelo linear bivariado

Vamos começar com um modelo simples:

$$y=\beta 0+\beta 1x+u$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Assumimos que esse modelo vale para a população de interesse. Esse é um modelo linear bivariado: temos duas variáveis, **y** e **x**, relacionadas por uma forma funcional linear como **y= ax + b**.

Essa equação permite que outros fatores afetem y. Tais fatores são representados pelo termo de erro u, que representa tudo aquilo que afeta y, mas não observamos.

Chamamos o coeficiente β_0 de **parâmetro de intercepto** e chamamos β_1 de **parâmetro de inclinação**. Esses dois parâmetros descrevem uma relação entre **y** e **x** para a população, e queremos estimá-los utilizando dados amostrais.

Atenção

Como estimamos β_0 e β_1?

Utilizando dados e hipóteses. Precisamos ter hipóteses críveis para estimar esses parâmetros de maneira precisa. Em nossa equação, todas as variáveis não observadas estão contidas no termo u, que chamamos de **erro idiossincrático**

Hipóteses e dados

Primeiro, faremos a seguinte hipótese simplificadora sem perda de generalidade:

$$E[u]=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Normalizar a esperança do erro para zero na população é uma hipótese inofensiva. Por quê? Porque a presença de β_0 sempre nos permite isso. Caso a média de u seja diferente de zero (e.g. um valor α_0), apenas adicionamos o valor do intercepto.

Esse ajuste não tem efeito algum sobre o coeficiente β_1:

$$y=(\beta 0+\alpha 0)+\beta 1x+(u-\alpha 0)$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que $\alpha_0=E(u)$. O novo termo de erro é $(u-\alpha_0)$ e o novo intercepto é $(\beta_0 + \alpha_0)$.

A inclinação β_1, porém, não foi alterada.

Qual é a interpretação do intercepto vertical?

É simples: se x=0, o valor de y é, em média, igual a β_0.

Outra hipótese que fazemos envolve a média do erro para cada “fatia” da população determinada pelos valores de x:

$$E[u|x]=E[u]$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Se essa hipótese vale, dizemos que u é **independente na média** de x. Essa é a independência na média condicional de que falamos no primeiro módulo. Isso significa que, quando a nossa variável explicativa (x) muda, o nosso termo de erro (u) não é afetado.

Independente na média

Você pode encontrar algumas vezes o termo em inglês: *mean independent*

Relembrando

Vamos retomar nosso exemplo para facilitar o entendimento do que representa essa hipótese.

Exemplo – efeito de escolaridade em salário

Suponha que estamos estimando o efeito de escolaridade em salário, e **u** é a habilidade inata de cada indivíduo, que não observamos. Independência na média implica que:

$$E[habilidade|x=8]=E[habilidade|x=12]=E[habilidade|x=16]$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Ou seja, a habilidade esperada é a mesma para diferentes grupos da população — com 8, 12 ou 16 anos de educação. Como as pessoas escolhem seu investimento em educação baseado, em parte, em sua habilidade, que é não observada, essa hipótese é

provavelmente violada no exemplo de educação e salário. Precisaremos, portanto, de outros modelos para estimar essa relação (isso também é assunto para outros temas).

Ao combinarmos essa nova hipótese **E[u | x]=E[u]**, que, como vimos, não é facilmente satisfeita nos dados, com **E[u]=0**, que é simplesmente uma normalização dos dados, chegamos à nova hipótese:

$$E[u|x]=0, \text{ para todo } x$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Essa hipótese é conhecida como **hipótese de esperança condicional zero**, sendo uma hipótese-chave para identificarmos efeitos em modelos de regressão linear.

De fato, podemos partir dessa hipótese, como faremos agora. Observe inicialmente que ela implica **E(u)=0**:

$$E(u)=Ex[E(u|x)]=Ex[0]=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

A primeira igualdade é uma aplicação da lei das expectativas iteradas.

Além disso, a independência da média condicional implica que a covariância entre u e x é igual a zero: Cov(u,x)=0. Mas lembre-se de que a definição de covariância é Cov(u,x)= E(ux) - E(u) x E(x). Portanto:

$$E(ux)=E(u)\times E(x)=0\times E(x)=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Usaremos adiante esses dois resultados: E(u) = 0 e E(ux) = 0

Como a esperança condicional é um operador linear, **E[u | x]=0** também nos permite escrever:

$$y=\beta 0+\beta 1x+u$$

$$E[y|x]=E[\beta 0+\beta 1x+u|x]$$

$$E[y|x]= E[\beta 0|x]+E[\beta 1x|x]+E[u|x]$$

$$E[y|x]= \beta 0+\beta 1E[x|x]+E[u|x]$$

$$E[y|x]= \beta 0+\beta 1x+0$$

$$E[y|x]=\beta o+\beta 1x$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Segunda linha

Simplesmente aplicamos a esperança (condicional) em ambos os lados da equação.

Terceira linha

Usamos o fato de que a esperança é um operador linear e, portanto, a esperança da soma é a soma das esperanças.

Quarta linha

Atenção! Usamos o fato de que os parâmetros (β_0,β_1) **não dependem da variável explicativa**. Afinal, esses parâmetros são constantes — desconhecidos, mas constantes — e não dependem do valor da variável explicativa **x**. A esperança (condicional ou incondicional) de uma constante é igual à própria constante ($E[\beta_0 \mid x]=\beta_0$), e usamos novamente o fato de que a esperança é um operador linear para escrever $E[\beta_1 x \mid x]=\beta_1 E[x \mid x]$.

Quinta linha

Apenas observamos que $E[x \mid x]=x$ e usamos nossa hipótese central $E[u \mid x]=0$.

Sexta linha

Temos que a função de regressão populacional é uma função linear de x ou uma função de esperança condicional. Essa relação é crucial para que possamos interpretar β_1, sob certas condições, como um parâmetro causal.

Mínimos Quadrados Ordinários

Vamos avançar, agora, para o assunto principal deste módulo: o modelo de mínimos quadrados ordinários (MQO).

Supondo que temos dados sobre x e y, como podemos estimar os parâmetros populacionais **β_0 e β_1?** Seja **$\{(x_i,y_i): i=1,2,\ldots,n\}$** uma **amostra aleatória de tamanho n** (o número de observações) de uma população.

Vamos escrever nossa equação populacional para determinado indivíduo i:

$$y_i=\beta_0+\beta_1 x_i+u_i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que i indica uma observação em particular (ou seja, um indivíduo na nossa amostra).

Conseguimos observar nos dados y_i e x_i, mas não conseguimos observar u_i. Apenas sabemos que ele existe e afeta, de alguma maneira, nossos resultados. Usamos, então, as duas restrições populacionais que discutimos anteriormente para obter as equações

para β_0 e β_1: E[u]=0 e E[ux]=0.

Podemos, agora, reescrever as restrições apresentadas usando **$u=y-\beta_0-\beta_1 x$**:

$u=y-\beta_0-\beta_1 x$

Isso é apenas uma forma de reescrever o nosso modelo original:

$y = \beta_0+ \beta_1 x + u$

$$E[y-\beta_0-\beta_1 x]=0$$

$$E(x[y-\beta_0-\beta_1 x])=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Essas são as duas condições na população que efetivamente determinam β_0 e β_1.

Novamente, observe que a notação aqui é populacional. No entanto, no mundo real de análise estatística, não temos acesso à população, somente a uma amostra dela. As contrapartidas amostrais das duas equações apresentadas anteriormente são:

$$1nn\sum i=1(yi-\hat{\beta}0-\hat{\beta}1xi)=0$$

$$1nn\sum i-1(xi[yi-\hat{\beta}0-\hat{\beta}1xi])=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que ^β0 e ^β1 são as estimativas dos parâmetros populacionais β_0 e β_1, respectivamente.

O leitor mais atento notará que estamos dividindo por **n, e não n-1**. Não há ajuste sendo feito para graus de liberdade quando calculamos médias amostrais, mas existe quando calculamos momentos de ordem maior como a variância amostral.

Essas são duas equações lineares com duas incógnitas. Vamos, agora, obter algumas propriedades amostrais dessas equações.

Desenvolvimento do Estimador de Mínimos Quadrados Ordinários

Para começar, vamos desenvolver o lado esquerdo da primeira equação:

$$1nn\sum i=1(yi-\hat{\beta}0-\hat{\beta}1xi)=1nn\sum i=1yi-1nn\sum i=1\hat{\beta}0-1nn\sum i=1\hat{\beta}1xi$$

$$=1nn\sum i=1yi-\hat{\beta}0-\hat{\beta}1(1n\,n\sum i=1xi)$$

$$=\bar{y}-\hat{\beta}0-\hat{\beta}1\bar{x}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Em que $\bar{y}=1nn\sum i=1yi$ é a **média amostral** para as n observações de y_i (a notação é análoga para $\bar{x}$) Logo, $1nn\sum i=1(yi-\hat{\beta}0-\hat{\beta}1xi)=0$ implica em , $\bar{y}-\hat{\beta}0-\hat{\beta}1\bar{x}=0$ ou $\bar{y}=\hat{\beta}0+\hat{\beta}1\bar{x}$. Usamos esse resultado para explicitar o intercepto em termos da inclinação:

$\hat{\beta}0=\bar{y}-\hat{\beta}1\bar{x}$

Inserimos esse resultado na segunda equação:

$n\sum i=1(xi[yi-\hat{\beta}0-\hat{\beta}1xi])=0$

Desse modo, obtemos o seguinte resultado:

$$n\sum i=1(xi[yi-(\bar{y}-\hat{\beta}1\bar{x})-\hat{\beta}1xi])=0$$

$$n\sum i=1(xi[yi-\bar{y}+\hat{\beta}1\bar{x}-\hat{\beta}1xi])=0$$

$$n\sum i=1[xi(yi-\bar{y})+\hat{\beta}1\bar{x}xi-\hat{\beta}1x2i]=0$$

$$n\sum i=1xi(yi-\bar{y})=\hat{\beta}1n\sum i=1xi(xi-\bar{x})$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Note ainda que:

$n\sum i=1(xi-\bar{x})2=n\sum i=1(x2i-2xi\bar{x}+\bar{x}2)$

$=n\sum i=1x2i-2\bar{x}n\sum i=1xi+n\bar{x}2$

$=n\sum i=1x2i-2n\bar{x}2+n\bar{x}2$

$=n\sum i=1x2i-n\bar{x}2$

$=n\sum i=1xi(xi-\bar{x})$

$=n\sum i=1x2i-2\bar{x}n\sum i=1xi+n\bar{x}2$

Nessa passagem, usamos o fato de que $\bar{x}$ é uma constante.

$=n\sum i=1x2i-2n\bar{x}2+n\bar{x}2$

Nessa passagem, usamos o seguinte fato:

$n\sum i=1xi=n\bar{x}$

Essa é apenas uma forma de reescrever a média

$\bar{x}$.

E também que:

$n\sum i=1(xi-\bar{x})(yi-\bar{y})=n\sum i=1(xiyi-xi\bar{y}-\bar{x}yi+\bar{x}\bar{y})$

$=n\sum i=1xiyi-n\bar{x}\bar{y}-n\bar{x}\bar{y}+n\bar{x}\bar{y}$

$=n\sum i=1xiyi-n\bar{x}\bar{y}$

$=\sum ni=1xi(yi-\bar{y})$

$=n\sum i=1yi(xi-\bar{x})$

$=n\sum i=1yi(xi-\bar{x})$

Essa expressão é obtida a partir da expressão anterior:

$n\sum i=1xiyi-n\bar{x}\bar{y}$

Desse modo, podemos reescrever $n\sum i=1xi(yi-\bar{y})=\hat{}\beta 1n\sum i=1xi(xi-\bar{x})$ como:

$$n\sum i=1(xi-\bar{}x)(yi-\bar{}y)=\hat{}\beta 1[n\sum i=1(xi-\bar{}x)2]$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Se $n\sum i=1(xi-\bar{}x)2\neq 0$, podemos escrever:, podemos escrever:

$$\hat{}\beta 1=n\sum i=1(xi-\bar{}x)(yi-\bar{}y)n\sum i=1(xi-\bar{}x)2$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Podemos fazer, ainda, uma pequena modificação nessa expressão. Para isso, vamos dividir o numerador e o denominador por (n-1):

$$\hat{}\beta 1=[n\sum i=1(xi-\bar{}x)(yi-\bar{}y)]/(n-1)[n\sum i=1(xi-\bar{}x)2]/(n-1)$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Essa divisão não alterou a expressão, mas nos permite interpretar o nosso estimador.

O que encontramos agora?

O numerador é a covariância amostral entre x e y, e o denominador é a variância amostral de x. Ou seja:

$$\hat{}\beta 1=\text{Covariancia amostral}(xi,yi)\text{Variância amostral}(xi)$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

A fórmula apresentada anteriormente para $\hat{}\beta 1$ é muito importante porque nos mostra como usar os dados que temos em mãos para computar a estimativa da inclinação da reta de regressão linear.

Você sabia

É possível computar esse estimador sempre que a variância de x_i é diferente de zero, isto é, quando temos variação suficiente nas variáveis que queremos usar para explicar y_i. Em outras palavras, se x_i não é constante para todos os valores de i, é possível gerar um estimador de MQO ^β1 para β_1

A intuição é que a variação em x é o que permite identificar o seu impacto em y. Isso também significa, no entanto, que não podemos determinar essa inclinação em uma relação na qual observamos uma amostra em que, por exemplo, todos os indivíduos têm o mesmo número de anos de escolaridade ou qualquer outra variável explicativa que estamos interessados em estimar.

Isso é intuitivo. Queremos saber o impacto de uma variável explicativa sobre uma variável dependente: se a variável explicativa não muda, qualquer mudança na variável dependente deve ter outra causa!

Vamos olhar para a expressão de ^β1 com cuidado. Para facilitar o raciocínio, pense, por um instante, que a variância amostral de X é igual a um. Nesse caso, a covariância amostral entre X e Y é exatamente a inclinação da reta de mínimos quadrados que relaciona Y e X. Em outras palavras, se X aumenta em uma unidade, a variação de Y é, em média, igual a essa covariância.

No caso geral, precisamos corrigir a covariância amostral, dividindo-a pela variância amostral de X. Se essa variância for baixa, a inclinação ^β1 será mais alta, para uma dada covariância entre X e Y. É intuitivo: se todos os valores de X são próximos uns dos outros, a reta de regressão será bastante inclinada.

Uma vez que calculamos ^β1 podemos computar o valor do intercepto ^β0 como ^β0=¯y-^β1¯x Essa é a estimativa de MQO do intercepto. Ela é uma estimativa pois é calculada a partir de médias amostrais. Esse resultado para o intercepto é fácil de obter pois ^β0 é linear em ^β1

Para quaisquer estimativas ^β0 e^ β1 que satisfaçam os resultados obtidos anteriormente, definimos os **valores preditos**:

$$\hat{y}_i=\hat{\beta}_0+\hat{\beta}_1 x_i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Lembre-se de que i={1,…,n} e, portanto, temos n equações dessas. Esse é o valor que prevemos para y_i dado que $x=x_i$. Note, porém, que há um erro nessa predição, pois $y \neq y_i$. Chamamos esse erro de **resíduo**, e utilizamos a notação ^ui para ele. O resíduo é dado por:

$$\hat{u}_i = y_i - \hat{y}_i$$

$$= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Suponha que medimos o tamanho desse resíduo, para cada i, elevando-o ao quadrado. Essa operação eliminará todos os valores negativos desse erro de predição.

Esse tipo de transformação é útil quando queremos somar os valores desses resíduos para ter uma dimensão da qualidade do nosso modelo econométrico e não queremos que valores positivos e negativos se cancelem. Nesse caso, devemos fazer:

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

$$= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Essa expressão é conhecida como **soma dos quadrados dos resíduos (SQR)**.

O resíduo é baseado em estimativas da inclinação da reta e do seu intercepto.

É possível imaginar vários valores para essas estimativas, gerando várias retas diferentes. Como escolher um? Um procedimento natural, se queremos um modelo bom, é que ele “acerte” nas predições. Para isso, gostaríamos que ele minimizasse a SQR.

Queremos, portanto, escolher $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimize SQR. **Utilizando ferramentas de otimização das aulas de cálculo**, é possível mostrar que as estimativas de $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam esse valor são justamente as que obtemos nesta seção. Uma vez que obtemos esses coeficientes, obtemos a **reta de regressão linear via MQO:**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

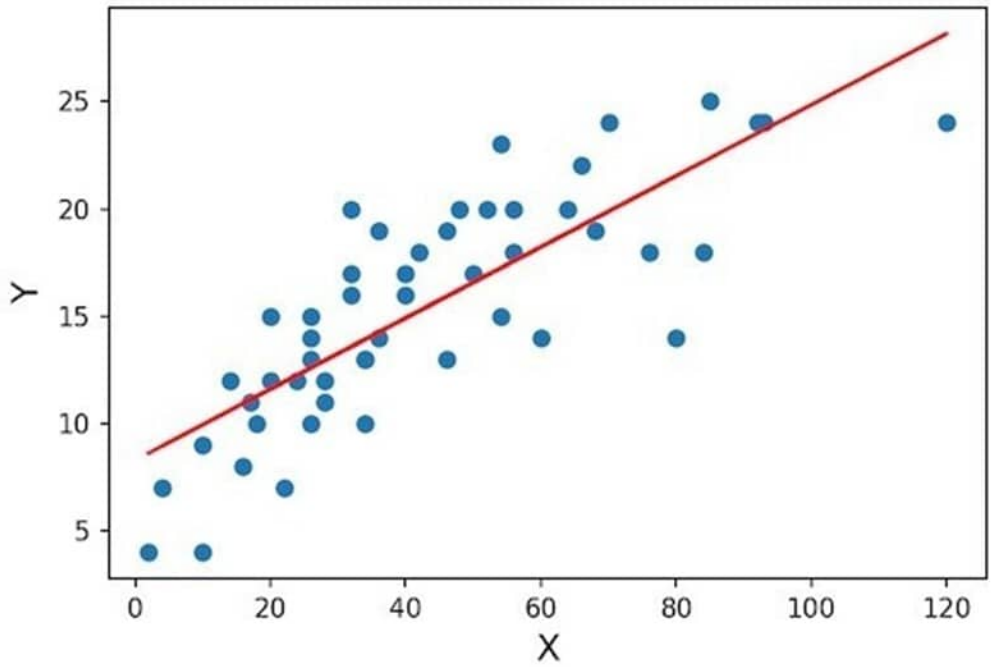
Representação gráfica

Para concluir, vamos fazer uma representação gráfica dos objetos que estamos discutindo. Temos diversas observações para as variáveis x e y: cada ponto azul na imagem abaixo é um par ordenado (x,y) que observamos.

A reta vermelha relaciona essas duas variáveis, mas notamos que ela não passa exatamente por cada um dos pontos azuis. Por quê?

É simples: x não é o único fator que afeta o valor de y — temos também o termo u. A distância vertical entre o um ponto azul e a reta vermelha é exatamente o valor de u para uma observação específica.

Não conhecemos a verdadeira equação da reta que relaciona X e Y para toda população, mas vimos como estimá-la a partir de uma base de dados.



Fonte: YDUQS

Propriedades algébricas de estimadores de MQO

Veremos, agora, algumas propriedades algébricas dos estimadores de MQO obtidos.

Quando temos o intercepto em nossa regressão, podemos escrever:

$$n\sum i=1(yi-\hat{}\beta 0-\hat{}\beta 1xi)=n\sum i=1\hat{}ui=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

A soma dos resíduos da regressão linear obtida via MQO sempre soma zero, por construção. Como yi=^yi+^ui por definição, podemos tirar a média amostral dos dois lados:

$$1nn\sum i=1yi=1nn\sum i=1\hat{}yi+1nn\sum i=1\hat{}ui$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

A média dos valores previstos, no entanto, é igual a zero:

$$1nn\sum i=1\hat{}ui=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

A equação anterior se torna:

$$1nn\sum i=1yi=1nn\sum i=1\hat{}yi$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Ou seja, a média das observações (y_i) é igual à média dos valores previstos ($\hat{}yi$) . De maneira similar, utilizando a equação que usamos para obter as estimativas dos parâmetros, temos que:

$$n\sum i=1(xi[yi-\hat{}\beta 0-\hat{}\beta 1xi])=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

O que implica que a covariância amostral entre as variáveis explicativas x_i e os resíduos é sempre zero:

$$n\sum i=1xi\hat{}ui=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Por fim, como $\hat{}yi$ é função linear de x_i (basta ver a reta de regressão linear via MQO), os valores previstos também não possuem correlação com os resíduos:

$$n\sum i=1\hat{}yi\hat{}ui=0$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

As duas igualdades apresentadas, que determinam que não há correlação entre os resíduos e as variáveis explicativas e os valores previstos, são geradas por construção.

Ou seja, ^β0 e ^β1 são escolhidos de tal maneira que essas igualdades sejam verdadeiras.

Uma terceira propriedade surge se incluirmos a média amostral de x como variável explicativa em nossa regressão. Obteremos a média amostral de y, ou seja, o ponto (¯x,¯y) sempre está na reta de regressão linear obtida via MQO:

$$\bar{y}=\hat{\beta}0+\hat{\beta}1\bar{x}$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Medindo a qualidade da regressão linear

Para cada observação, escrevemos:

$$yi=\hat{y}i+\hat{u}i$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Isso é intuitivo: o valor real de y_i é o valor que prevemos para ele mais uma medida de erro (nosso resíduo).

Vamos, agora, definir medidas para a qualidade de nossa previsão de y_i.

Defina a **soma quadrática total (SQT), a soma quadrática explicada (SQE) e a soma quadrática dos resíduos (SQR)** de tal modo que:

$$SQT=n\sum i=1(yi-\bar{}y)2$$

$$SQE=n\sum i=1(\hat{}yi-\bar{}y)2$$

$$SQR=n\sum i=1(yi-yi)2=n\sum i=1\hat{}ui2$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

Essas métricas de qualidade nada mais são do que variâncias amostrais quando dividimos por **n-1**. Desse modo, SQTn-1 é a variância amostral de yi, SQEn-1 é a variância amostral de ^yi e SQRn-1 é a variância amostral de ^ui . Com uma breve manipulação algébrica, podemos reescrever SQT da seguinte maneira:

$$SQT=n\sum i=1(yi-\bar{}y)2$$

$$=n\sum i=1[(yi-\hat{}yi)-(\hat{}yi-\bar{}y)]\ 2$$

$$=n\sum i=1[\hat{}ui-(\hat{}yi-\bar{}y)]\ 2$$

$$=n\sum i=1(\hat{}ui2-2\hat{}ui(\hat{}yi-\bar{}y)+(\hat{}yi-\bar{}y)2)$$

$$=n\sum i=1\hat{}ui2-2n\sum i=1\hat{}ui\hat{}yi+2\bar{}yn\sum i=1\hat{}ui+n\sum i=1(\hat{}yi-\bar{}y)2$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

$=n\sum i=1[(yi-\hat{}yi)-(\hat{}yi-\bar{}y)]\ 2$

Nessa passagem, somamos e subtraímos ^yi dentro do parênteses da linha anterior.

Finalmente, usando os resultados obtidos na última seção de que n∑i=1^ui=0 e n∑i=1^yi^ui=0 , obtemos:

$$SQT=n\sum i=1\hat{}ui2+n\sum i=1(\hat{}yi-\bar{}y)2$$

$$SQT=SQR+SQE$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

SQT=SQR+SQE

Curiosidade: isso é uma aplicação do Teorema de Pitágoras! O quadrado da hipotenusa é igual à soma dos quadrados dos catetos.

Se assumirmos que SQT > 0, podemos definir a fração da variação total em y_i que é explicada por x_i (ou pela nossa regressão) como:

$$R2=SQESQT=1-SQRSQT$$

Atenção! Para visualização completa da equação utilize a rolagem horizontal

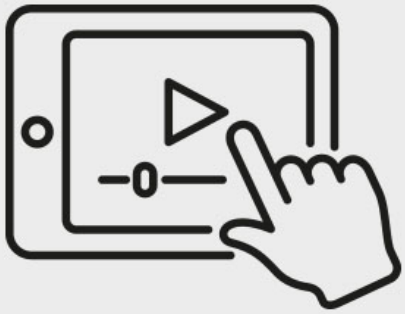
R^2

A medida chamada de “R-quadrado” ou R^2 nos diz a proporção da variância de y_i que pode ser explicada pelas variáveis explicativas que temos em mãos. É possível mostrar que ela é igual ao quadrado da correlação entre y_i e yi e ^yi , e, portanto, está no intervalo [0,1].

Um R^2 igual a zero significa que não há relação linear entre y_i e x_i, e um R^2 igual a 1 significa que há uma relação linear perfeita entre essas variáveis (por exemplo, $y_i=x_i+2$). À medida que o R^2 aumenta, os valores de y_i estão mais próximos da reta de regressão obtida via MQO.

Assista, a seguir, a um vídeo sobre a Medida R^2:

Para assistir a um vídeo sobre o assunto, acesse a versão online deste conteúdo.



Verificando o aprendizado

1. Assinale a alternativa correta sobre a hipótese de independência na média:

2. Assinale a alternativa correta sobre o coeficiente de inclinação da reta de regressão linear:

Gabarito
1. Assinale a alternativa correta sobre a hipótese de independência na média:

A alternativa **"B "** está correta.

Independência na média entre duas variáveis Y e X significa dizer que E[Y|X]=E[Y].

Podemos tirar a esperança em X dos dois lados dessa expressão e obter: E[E[Y|X]]=E[Y]

(**isso é a lei das expectativas iteradas**). Desse modo, temos que:

$$
\begin{aligned}
cov(X,Y) &= E[(X-E(X))(Y-E(Y))] \\
&= E\{E[(X-E(X))(Y-E(Y))|X]\} \\
&= E\{(X-E(X))(E(Y|X)-E(Y))\} \\
&= E\{(X-E(X))(E(Y)-E(Y))\} = 0,
\end{aligned}
$$

Em que, na segunda linha, utilizamos a lei das expectativas iteradas.

2. Assinale a alternativa correta sobre o coeficiente de inclinação da reta de regressão linear:

A alternativa **"D "** está correta.

Essa é a definição do coeficiente de inclinação da reta de regressão linear obtida via MQO. Podemos dizer que a inclinação da reta de regressão é a covariância amostral entre X e Y corrigida pela variância de X.

Avalie este módulo:

Considerações Finais

Neste tema, vimos alguns conceitos-chave para a introdução de econometria. Começamos com uma explicação do que é o método de trabalho em econometria, apresentando as abordagens estrutural e reduzida. Depois, construímos o estimador básico da econometria, pelo método de mínimos quadrados ordinários.

Estudamos, portanto, a base para toda pesquisa econométrica ou de análise de dados. Você encontrará isso a todo momento. Empresas, governos e a academia querem saber qual é o impacto de uma política pública sobre o desenvolvimento econômico, que

fatores incentivam a pesquisa científica, quais são os determinantes dos preços das ações, e assim por diante.

Você deu o primeiro passo para se tornar um cientista de dados!

Para ouvir um *podcast* sobre o assunto, acesse a versão online deste conteúdo.

Avaliação do tema:

REFERÊNCIAS
