Notas de aula para o curso de Econometria I

Nota 2: Regressão simples: motivação, FRP, FRA, interpretação dos parâmetros, estimação, previsões e resíduos

Thiago Fonseca Morello

fonseca.morello@ufabc.edu.br

sala 301, Bloco Delta, SBC

2 Regressão simples

2.1 Motivação

A busca dos determinantes de uma característica de interesse é um exercício recorrente na prática científica. Particularmente, em economia, parte-se da teoria para identificar as variáveis de fundo, ou seja, aquelas em função das quais é possível explicar o comportamento de uma determinada característica socioeconômica tal como a situação do indivíduo perante o mercado de trabalho (estar ou não empregado), investimento de uma empresa em inovação tecnológica e taxa de crescimento do PIB de uma nação, etc. Em outras palavras, seja Y a variável cujo comportamento deseja-se explicar, a teoria postula que existe pelo menos uma variável, X, a qual, a depender do valor por ela assumido, exerce influência sobre o valor assumido por Y.

Um exemplo de particular interesse para o Brasil e para os países não desenvolvidos em geral é o a da relação entre desnutrição infantil e renda familiar. Economistas como Ana Lúcia Kassouf, Rodolfo Hoffman e Antônio Carlos Campino, se dedicaram à investigação desta relação tomando por base, para isso, dados coletados a partir de entrevistas a domicílios brasileiros. O pesquisador Mark Agee, dos Estados Unidos, fez o mesmo, mas, porém, para o caso da Nigéria[1].

A teoria[2] postula, considerando uma sociedade cuja produção e distribuição de alimentos são geridas por mercados, uma relação negativa entre grau de desnutrição infantil, esta a variável a ser explicada, Y, e renda familiar, a qual assumirá a posição de X. A intuição está em que famílias com maior poder de compra têm mais acesso a alimentos e, pois, maior capacidade de manter suas crianças adequadamente nutridas.

O objetivo da análise econométrica não é verificar a consistência lógica ou teórica da relação entre variável explicada, Y, e variável explicativa, X, mas sim sua consistência empírica, entendida esta como a adequação às evidências reveladas pelos dados disponíveis. Ou seja, a partir do momento em que o pesquisador decide qual é

[1] Seguem as referências para os estudos originais dos autores mencionados. Kassouf, A. L. A demanda de saúde infantil no Brasil por região e setor. Pesquisa e Planejamento Econômico, v. 24, n. 2, p. 235-260, ago. Disponível em http://www.memoria.nemesis.org.br/index.php/ppe/article/view/806/745. Hoffman, R. Pobreza, insegurança alimentar e desnutrição no Brasil. Estudos Avançados vol.9 no.24 São Paulo Maio/Agosto 1995. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-40141995000200007.Campino, A. C. C., Aspectos sócio-econômicos da desnutrição no Brasil. Revista de Saúde Pública, São Paulo, 20(1):83-101, 1986. Disponível em http://www.scielo.br/pdf/rsp/v20n1/07.pdf.
Agee, M. Reducing child malnutrition in Nigeria: Combined effects of income growth and provision of information about mothers' access to health care services. Social Science & Medicine 71 (2010) 1973-1980. Disponível em http://www.sciencedirect.com/science/article/pii/S0277953610006696.

[2] Esta afirmação encontra fundamentação mais clara na abordagem das dotações (*entitlement approach*) empregada por Amartya Sen em um dos principais estudos de fenômenos de inanição e fome em massa, a obra "Poverty and Famines: an essay on entitlement and deprivation", tal como se pode comprovar na seção 10.1 do livro.

a relação relevante, cabe à análise econométrica procurar indícios de que tal relação se manifesta ou não nos dados.

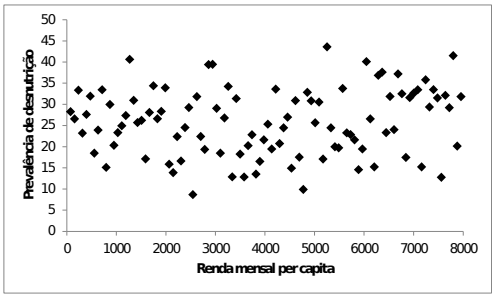
Um primeiro passo neste sentido pode ser dado com a elaboração de um gráfico de dispersão, a partir de um conjunto de dados que contenha informações para X e Y. Para o Brasil, a fonte de dados é a Pesquisa de Orçamentos Familiares de 2008/2009 (POF).

O exame destes dados será postergado. Por enquanto é mais esclarecedor ocupar-se de algumas das possibilidades que os dados podem vir a revelar. O painel a seguir indica três possibilidades. Nenhuma delas contém dados verídicos, mas sim valores gerados artificialmente com uma planilha Excel ®. A medida de grau de desnutrição infantil considerada é a de prevalência, ou seja, porcentagem de crianças de zero a cinco anos com altura consideravelmente inferior ao nível saudável para a idade, de acordo com a Organização Mundial de Saúde (OMS)[3].

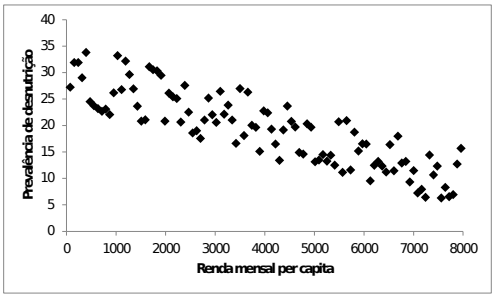
A unidade observacional dos gráficos é o setor censitário, uma região geográfica submunicipal definida pelo IBGE por fins estatísticos[4]. Desta maneira, são observadas, nos gráficos, a renda média dos setores censitários brasileiros e a prevalência de desnutrição em cada um deles. São considerados apenas 100 setores censitários.

Painel 1 Três possibilidades para o gráfico de dispersão

(A)



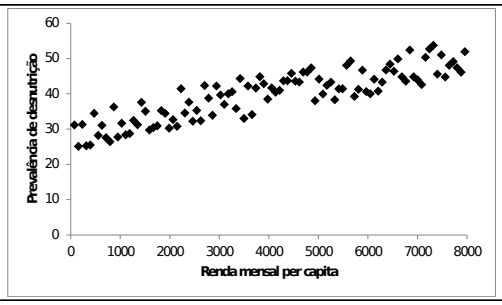
(B)



(C)

[3] Esta medida foi detalhada na primeira aula de laboratório, consultar o arquivo "script_lab_1".

[4] A definição de setor censitário, conforme consta na metodologia do censo demográfico de 2000 é "(...) unidade de controle cadastral formada por área contínua, situada em um único quadro urbano ou rural, com dimensão e número de domicílios ou de estabelecimentos que permitam levantamento das informações por um único Agente Credenciado, segundo cronograma estabelecido (vide página 227 de http://www.ibge.gov.br/home/estatistica/populacao/censo2000/metodologia/metodologiacenso2000.pdf)."



Caso o gráfico de dispersão gerado a partir dos dados coincida com (A), há razão para desconfiar da relação sugerida pela teoria. E isso pois não é possível reconhecer um padrão ou tendência clara. Na verdade, neste caso, os setores censitários se distribuem de maneira praticamente equitativa entre quatro grupos, quais sejam:

1. Grupo (AA): Níveis relativamente altos de renda familiar e níveis relativamente altos de prevalência de desnutrição infantil;
2. Grupo (AB): Níveis relativamente altos de renda familiar e níveis relativamente baixos de prevalência de desnutrição infantil;
3. Grupo (BA): Níveis relativamente baixos de renda familiar e níveis relativamente altos de prevalência de desnutrição infantil;
4. Grupo (BB): Níveis relativamente baixos de renda familiar e níveis relativamente baixos de prevalência de desnutrição infantil.

Entendendo-se por “relativamente alto” e “relativamente baixo”, respectivamente, valores superiores e inferiores à média de cada variável.

Estes quatro grupos correspondem aos quatro quadrantes em que o gráfico de dispersão pode ser dividido, tomando-se como referência as médias amostrais das variáveis. A tabela abaixo apresenta a contagem dos setores censitários em cada um dos quatro grupos possíveis definidos acima para cada uma das três possibilidades de gráficos do painel 1.

Tabela 1 Número de setores censitários em cada grupo para cada uma das três possibilidade de diagramas de dispersão

Grupo	Gráfico		
	A	**B**	**C**
AA	26	9	47
AB	24	41	3
BA	24	44	11
BB	26	6	39

Efetivamente, os quatro grupos têm participação praticamente equivalente para o caso ilustrado pelo gráfico (A), i.e., nenhum grupo predomina. Porém, nos gráficos (B) e (C), tal como a observação deles sugere, há uma tendência à concentração da amostra de setores censitários em grupos específicos. No caso do gráfico (B), setores censitários do grupo AB e do grupo BA predominam (juntos, respondem por 85% da amostra), o que está de acordo com a tendência negativa revelada pelo gráfico (B). Um nível de renda relativamente alto tende a vir acompanhado de uma prevalência relativamente baixa de desnutrição infantil. Já, no caso (C), são os setores

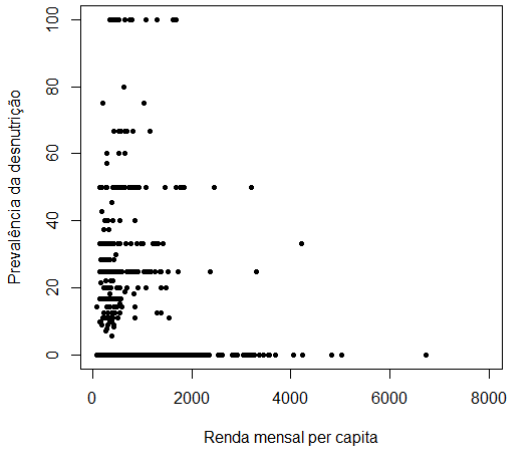
censitários dos grupos AA e BB que se mostram mais recorrentes (88% da amostra), o que está de acordo com a tendência positiva observada.

As duas formas de evidência consideradas, o diagrama de dispersão e a classificação das unidades observacionais em grupos de acordo com os valores das duas variáveis, podem bastar para o pesquisador. I.e., ele pode acreditar que os padrões revelados por estas duas ferramentas são claros o bastante para concluir quanto à validade ou invalidade empírica da relação teórica.

Há, contudo, pelo menos duas razões pelas quais uma abordagem mais precisa se mostra desejável:

1. Dados reais dificilmente seguem tendências claras, conforme **o gráfico abaixo (figura 5)** indica. O gráfico de dispersão e a classificação em grupos podem não revelar claramente uma tendência e nem a total falta de tendência. I.e., os dois instrumentos podem revelar evidências insuficientemente claras, inconclusivas;
2. O pesquisador pode estar interessado em medir a relação quantitativa entre as variáveis X e Y, i.e., determinar em qual magnitude o aumento da renda familiar, via, por exemplo, transferências governamentais de renda, se reverte em redução do grau de desnutrição infantil.

Figura 5 Gráfico de dispersão para a relação entre renda mensal per capita e prevalência de desnutrição, setores censitários brasileiros*



*apenas setores censitários com coeficiente de variação (desvio padrão/média) para a renda mensal per capita inferior à unidade são considerados.

A econometria procura assentar a relação empírica entre X e Y em uma base mais precisa. De fato, a disciplina tem por objetivo fundamental mensurar a relação quantitativa entre duas variáveis X e Y. Para que fique mais claro o termo “relação quantitativa”, cabe atentar para os exemplos de perguntas feitas por estudos econométricos recentes listados a seguir.

1. Em quanto o PIB per capita de um País seria aumentado caso fosse possível reduzir consideravelmente o nível de desigualdade de renda (Barro, 2008)[5]?

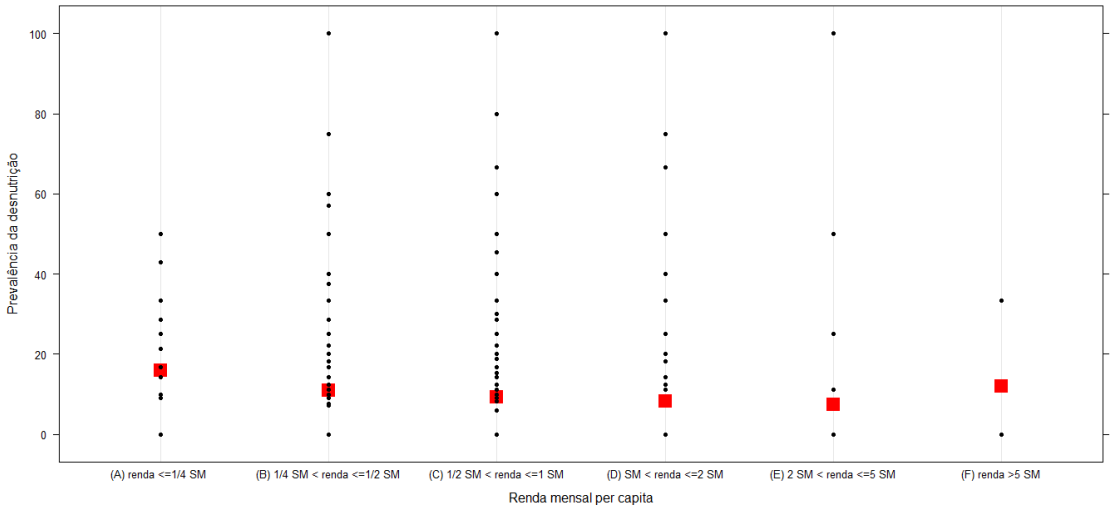
[5] Barro, R.J., Inequality and growth revisited. Working paper series on regional economic integration. Asian Development Bank. Disponível em http://aric.adb.org/pdf/workingpaper/WP11_%20Inequality_and_Growth_Revisited.pdf

2. Qual é o aumento de salário que um trabalhador poderia obter caso seu nível de qualificação fosse ampliado em um ano adicional de estudo (Teixeira e Menezes-Filho, 2012)[6]?
3. Em quanto aumentaria a renda de uma família caso a oferta de microcrédito fosse ampliada (Banerjee et al, 2014[7])?
4. A área de floresta Amazônica desmatada por um produtor agropecuário seria consideravelmente maior caso ele tivesse acesso a mais crédito bancário (Assunção, 2013[8])?

2.2 Função de expectativa condicional

Como apreender a relação quantitativa entre duas variáveis? É possível avançar em tal sentido introduzindo uma pequena sofisticação no gráfico de dispersão. Agora com base nos dados reais da POF 2008/2009, pode-se calcular a média para a prevalência de desnutrição dentro de faixas para a renda familiar, como ilustrado pelos quadrados vermelhos do gráfico abaixo. Os pontos na direção vertical correspondem aos valores que a variável Y assume para as observações cuja renda familiar pertence a uma dada faixa.

Figura 6 Média condicional para a prevalência de desnutrição (quadrados vermelhos) e níveis de prevalência observados na amostra (círculos pretos)*, SM = salário mínimo



*apenas setores censitários com coeficiente de variação (desvio padrão/média) para a renda mensal per capita inferior à unidade são considerados.

O gráfico indica que a média de Y, calculada "dentro" de grupos de observações definidos em função de valores de X, exibe uma tendência aparentemente negativa, ainda que isso não seja muito claro, o que é comum para dados reais. De qualquer maneira, neste estágio do argumento, a atenção deve ser voltada à compreensão do significado das médias representadas pelos quadrados vermelhos. Para isso, é esclarecedor coletar algumas informações do gráfico, tal como segue.

---

[6] Teixeira, W. M., Menezes-filho, N.A. "Estimando o retorno à educação do Brasil considerando a legislação educacional brasileira como um instrumento". Revista de Economia Política, vol. 32, nº 3 (128), pp. 479-496, julho-setembro/2012. Disponível em http://www.scielo.br/pdf/rep/v32n3/08.pdf

[7] Banerjee, A., Duflo, E, Glennester, R., Kinnan, C. "The miracle of microfinance? Evidence from a randomized evaluation." Working paper, http://economics.mit.edu/files/5993

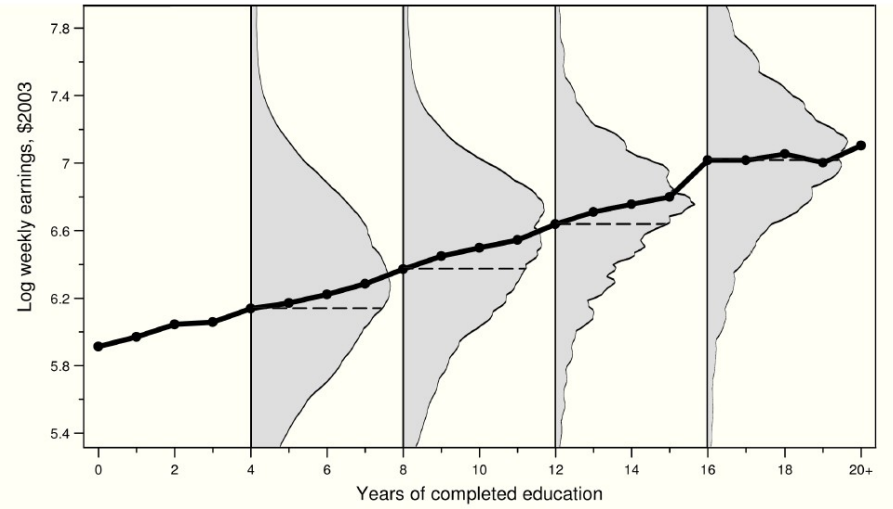
[8] ASSUNÇÃO, J., GANDOUR, C., ROCHA, R., ROCHA, R. 2013. Does credit affect deforestation? Evidence from a rural credit policy in the Brazilian Amazon. Climate Policy Initiative. Disponível em: http://climatepolicyinitiative.org/wp-content/uploads/2012/03/Deforestation-Prices-or-Policies-Working-Paper.pdf

1. Nos setores censitários em que a renda mensal per capita é inferior a ¼ do salário mínimo, a prevalência de desnutrição média é superior a 10%;
2. Esta taxa é inferior a 10% nos setores censitários com renda mensal per capita entre dois e cinco salários mínimos.

A leitura dos dados sugerida pelos “fatos” acima é um pouco mais clara do que o permitido por gráficos e tabelas. E isso pois, dado um determinado nível da variável X, renda familiar, pode-se identificar um único valor correspondente à variável Y, sua média, no caso. Clareza está que resulta do emprego da média para resumir a dispersão da variável Y para cada uma das faixas de X.

De fato, a média amostral de Y para grupos definidos em função de X é análoga ao conceito populacional de expectativa condicional visto em estatística e representado por E[Y|X]. Este conceito propõe que a informação quanto ao valor de X é relevante para determinar qual valor de Y é mais provável, i.e., têm maior probabilidade de ocorrência. O que é o mesmo que dizer que a distribuição probabilística de Y, i.e., a relação que nos diz quais valores de Y são mais prováveis e quais são menos prováveis, varia em função de X. Desta maneira, ao invés de conceber a distribuição probabilística de Y como dada por uma única função de distribuição de probabilidades (FD), é possível pensar que, para cada valor de X, existe uma distribuição probabilística potencialmente distinta para Y. É isso que o gráfico abaixo sugere, tomando como Y o logaritmo do salário semanal e como X os anos de escolaridade, isso para um conjunto de dados referente a uma amostra de trabalhadores.

Distribuição condicional de Y em relação a X (cinza) e Expectativa condicional de Y em relação a X (linha preta)



Fonte: gráfico reproduzido de Angrist, J.D., Pischke, J-S., 2009. Mostly harmless econometrics, an empiricist’s companion. Princeton University Press, New Jersey, US.

Da mesma maneira que existe, para cada valor de X, uma distribuição probabilística potencialmente distinta, existem parâmetros potencialmente distintos que regem tal distribuição. Por exemplo, as distribuições condicionais referentes a valores diferentes de X, podem diferir em função da média populacional, μ. É exatamente esta possibilidade que a notação E[Y|X] indica, uma vez que ela se refere à média populacional de Y para um dado valor de X.

2.3 Função de regressão populacional

Qual é o formato exato de E[Y|X]? Ou seja, como a média populacional de Y varia em função de X? A priori, não é possível saber, uma vez que, como geralmente se parte de dados amostrais, é impossível determinar os valores populacionais dos parâmetros. Porém, é sempre possível afirmar que existe uma relação funcional entre E[Y|X] e X, i.e., E[Y|X] = f(X). Esta relação funcional é denominada por função de expectativa condicional (FEC) ou por função de regressão populacional (FRP).

A função f(X) não necessariamente é linear, ela pode ser quadrática ou exibir qualquer outro comportamento não linear. Porém, é sempre possível tomar uma aproximação linear à f(X), o que pode ser visto, seguindo Gujarati, como uma hipótese de partida, uma primeira aproximação do problema. Ou seja, $E[Y|X] \approx \beta_0 + \beta_1 X$ (1).

O segundo passo crucial para avançar na representação da relação entre X e Y está no fato, demonstrado pela teoria estatística, de que sempre é possível decompor uma variável aleatória em dois elementos[9]. O primeiro deles é a porção da informação contida na variável que é "explicada" por outra variável, o que pode ser representado a partir da expectativa condicional. Tomando Y como a variável "explicada" e X como variável "explicativa", o primeiro elemento em que Y se decompõe é E[Y|X]. O segundo elemento corresponde à porção de Y não "explicada" por X, ou, de maneira mais precisa, não correlacionada com X, porção esta que será denotada por "u". Desta maneira, pode-se escrever Y = E[Y|X] + u (2).

Combinando os resultados (1) e (2) pode-se chegar à função linear abaixo.

$$Y = E[Y|X] + u \approx \beta_0 + \beta_1 X + u$$

Ou, de maneira sintética:

$$Y \approx \beta_0 + \beta_1 X + u$$

O símbolo indicando aproximação linear pode ser substituído, em nome da simplicidade notacional, pelo símbolo de igualdade desde que se tenha em mente que a reta acima é uma aproximação linear para a FRP. Ela também é denominada por reta de regressão linear populacional.

É preciso assinalar a natureza populacional do modelo acima: os coeficientes β_0 e β_1 são parâmetros populacionais, desconhecidos a priori, assim como é o caso da média μ para uma variável aleatória normalmente distribuída.

Outro detalhe fundamental diz respeito à natureza do termo "u". Adotando a nomenclatura de Wooldridge, u será denominado por "termo de perturbação" ou "termo de erro". Ele é equivalente a $Y - \beta_0 - \beta_1 X$, tratando-se, portanto, da porção da variação de Y, ao longo das observações, que permanece não explicada mesmo após a incorporação da informação quanto ao comportamento de X. Gujarati apresenta algumas interpretações para o termo de perturbação. As principais são reproduzidas no que segue.

1. O termo de perturbação capta variáveis que explicam Y, mas são omitidas do modelo pois:
    a. Não são mencionadas pela teoria;
    b. São mencionadas pela teoria, mas não há dados disponíveis para elas;

[9] Este parágrafo segue a interpretação de Angrist & Pischke (2009, p.25-26) para a propriedade de decomposição da função de expectativa condicional.

2. O termo de perturbação capta erros de medida decorrentes do emprego de variáveis proxy. É o que se tem quando as variáveis, tais como definidas pela teoria, não estão disponíveis nos dados, mas há outras variáveis disponíveis, correlacionadas com as primeiras, i.e, que se comportam de maneira parecida. Por exemplo, segundo a teoria do q de Tobin, uma das principais variáveis que explicam o investimento em capital fixo por parte de uma empresa é retorno marginal do capital fixo (medida esta que corresponde ao q de Tobin em si), porém, grandezas marginais dificilmente podem ser calculadas a partir de dados concretos. É praxe utilizar o retorno médio do capital fixo, dado pela razão entre o valor de mercado de uma empresa (retorno medido pelo mercado de ações) e o valor de seu estoque de capital. A diferença entre a medida proposta pela teoria e a medida factível é captada pelo termo de perturbação. Outro exemplo: no artigo "Desigualdade de renda nos Estados Unidos, 1913-1998"[10], os economistas Thomas Piketty e Emmanuel Saez utilizaram declarações de impostos de renda como proxy para a renda individual. Se esta medida fosse utilizada como variável explicativa em uma FRP para a poupança individual, por exemplo, o termo de perturbação captaria a diferença entre a renda efetiva, esta a medida mencionada pela teoria, e a renda declarada no imposto de renda, esta a medida factível incorporada à FRP;
3. O termo de perturbação capta erros de especificação da relação entre X e Y. Muitas vezes a teoria não é precisa o bastante para estabelecer a forma funcional da relação em questão. A aproximação linear pode falhar em captar não-linearidades em tal relação, erro este o que acaba compondo o termo de perturbação.

2.4 Inferência e função de regressão amostral

Os valores populacionais de parâmetros de interesse são geralmente desconhecidos, sendo preciso estimá-los a partir das amostras de dados disponíveis. Não é diferente para o caso da análise de regressão linear, i.e., para os parâmetros β_0 e β_1.

Para atingir o objetivo da análise empírica em economia, o qual é sempre caracterizar a relação entre Y e X, geralmente está disponível apenas uma amostra de valores para as duas variáveis. Por exemplo, para determinar em qual medida a renda familiar explica, no Brasil, o grau de desnutrição infantil, os dados disponíveis mais atualizados correspondem à POF 2008/2009, uma amostra de 55.412 famílias de um total de 57 milhões de famílias brasileiras (apenas 0,1% das famílias foram entrevistadas).

O salto de inferência se mostra inevitável e com base nele se acaba por obter não a FRP, a qual nunca é observada, mas um elemento análogo, cujo conteúdo informacional se resume à amostra, a função de regressão amostral, FRA, representada como segue.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Em que $\hat{\beta}_0$ e $\hat{\beta}_1$são estimadores para β_0 e β_1.

[10] Disponível em http://piketty.pse.ens.fr/fichiers/public/PikettySaez2003.pdf.

2.5 Linearidade e interpretação dos parâmetros

Não necessariamente a relação entre Y e X é linear. Por exemplo, a teoria microeconômica sugere que a relação entre quantidade e custo total médio é quadrática, de modo que a FRP adequada, neste caso, seria dada por $E[CTM|Q] = \beta_0 + \beta_1 Q^2 + u$. Esta especificação é compatível com a definição de regressão linear, mesmo havendo claramente uma relação não-linear entre X e Y.

A afirmação que se acaba de fazer parece contraditória, mas não é. De fato, o significado correto de "linear" no termo "regressão linear" é o de uma relação entre X e Y linear nos parâmetros e não necessariamente linear em X. Para ver a diferença, basta tomar as especificações a seguir.

$E[Y|X] = \beta_0 + \beta_1 X^2$

$E[Y|X] = \beta_0 + \beta_1 \sqrt{X}$

$$E[Y|X] = \frac{1}{\beta_0 + \beta_1 Q}$$

Apenas a última especificação é não-linear nos parâmetros e, portanto, não pode ser estimada com base nas técnicas de regressão linear aqui discutidas. As duas primeiras podem ser estimadas com base nos estimadores de MQO, mas, conforme alerta Wooldridge a intepretação dos coeficientes é distinta quando há não-linearidades em X.

É altamente importante saber interpretar os parâmetros da regressão linear. De início, toma-se uma especificação totalmente linear (na variável explicativa e nos parâmetros), como $Y = \beta_0 + \beta_1 X + u$. Recordando-se que tal expressão a rigor significa $Y = E[Y|X] + u$ e $E[Y|X] = \beta_0 + \beta_1 X$, e, tomando X como uma variável contínua, a interpretação correta de β_1 é que se trata, nas condições até aqui assumidas, do incremento da expectativa condicional de Y em relação à X, resultante de um incremento infinitesimal em X. Ou seja:

$$\frac{d}{dX} E[Y \vee X] = \frac{d}{dX}(\beta_0 + \beta_1 X) = \beta_1$$

Por exemplo, se X é renda em R$/mês e Y a despesa em consumo, em R$/mês, então β_1 corresponde ao número de unidades monetárias em que o consumo mensal do domicílio médio aumenta como o resultado do aumento da renda mensal em exatamente R$1.

Uma segunda especificação altamente recorrente em exercícios econométricos incorpora variável dependente e independente em forma logarítmica – aqui, a base do logaritmo é geralmente o número neperiano. Ou seja:

$$\log(Y) = \beta_0 + \beta_1 \log(X) + u$$

A interpretação do coeficiente β_1 pode ser obtida a partir do raciocínio anterior. Antes é preciso notar que a especificação acima corresponde a, rigorosamente, $E[\log(Y)|X] = \beta_0 + \beta_1 \log(X)$ e $\log(Y) = E[\log(Y)|X] + u$. Com isso:

$$\frac{d}{dX} \log(E[Y \vee X]) = \frac{d}{dX}$$

$$\frac{X}{E[Y \vee X]}\frac{d}{dX}E[Y \vee X]=\beta_1 \leftrightarrow \frac{\frac{dE[Y \vee X]}{E[Y \vee X]}}{\frac{dX}{X}}=\beta_1 \leftrightarrow \varepsilon_X^{E[Y \vee X]}=\beta_1$$

Em que $\varepsilon_X^{E[Y \vee X]}$ representa a elasticidade de E[Y|X] em relação à X. Sendo, novamente, X a renda mensal e Y o consumo mensal, ambos em R$ e, para tornar mais claro o exemplo, $\beta_1 = 0,5$, a interpretação dimensionalmente correta é a de que, para 1% de aumento da renda mensal, tem-se 0,5% (i.e., 0,005) de aumento no consumo mensal.

Quanto à interpretação do intercepto, trata-se, na regressão totalmente linear, do valor assumido por Y quando X é nulo. Ou seja, seria o nível de consumo de domicílios com renda nula, o qual não necessariamente é nulo, pois por conta do acesso a crédito ou, no caso da área rural, da produção para o autoconsumo.

2.6 Estimação

A mera definição da FRA não sugere um caminho para obtê-la. Como é possível chegar a estimativas pontuais para o intercepto e o coeficiente da FRP? Há pelo menos três métodos de estimação que solucionam o problema, por hora basta se ocupar do mais famoso.

Um estimador é, antes de tudo, uma estatística. Estatísticas são usadas com o objetivo de resumir os dados. A média e a variância, por exemplo, resumem a distribuição individual de uma variável. Os estimadores para os parâmetros da FPR também têm de resumir informação, mas, porém, não quanto à distribuição individual de X e Y, mas sim quanto à relação quantitativa entre X e Y.

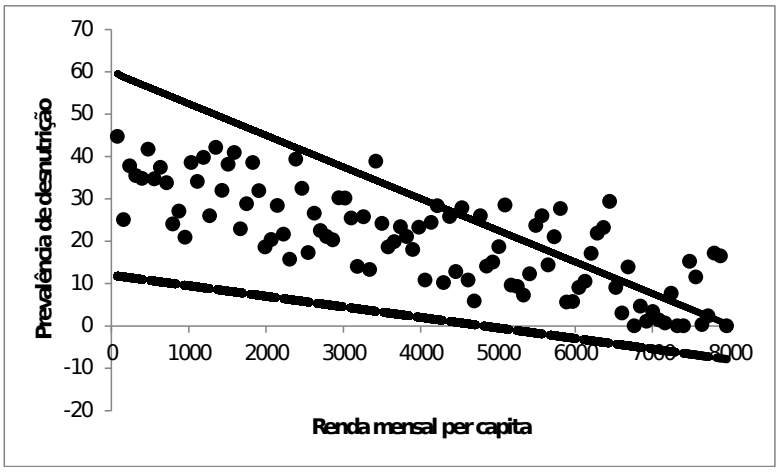
O formato da FRP sugere uma saída para resumir a relação entre X e Y: tomar uma aproximação linear do padrão descrito pelo gráfico de dispersão[11]. Mas, um detalhe crucial, muitas vezes perdido de vista, deve ser assinalado. O gráfico de dispersão em questão não é o construído a partir da amostra, mas sim a partir da população. A razão para isso é de grande importância: o objetivo da análise econométrica não é resumir a relação de X e Y tal como ela se manifesta na amostra, mas sim na população.

Por exemplo, a formulação de uma política nacional de combate à desnutrição infantil deve ser alicerçada na relação que esta variável tem com a renda familiar considerando-se todas as famílias brasileiras. Se for tomado por base apenas um subgrupo de famílias, uma medida de política pública, tal como a transferência de renda, poderá não render o resultado esperado para famílias que não pertencem ao subgrupo considerado.

Deve-se ressaltar, pois, que a imagem de um gráfico de dispersão para a população é puramente uma abstração, pois geralmente não está disponível toda a informação necessária para construí-lo para toda a população-alvo de um estudo econométrico.

Colocada esta ressalva, tomemos, para fins de compreensão, o gráfico abaixo, o qual representa toda a população.

[11] Infelizmente, calcular a média para Y dentro de faixas de X não permite obter uma função que descreva completamente o comportamento da relação entre as variáveis dentro da amostra.



As duas retas observadas no gráfico se mostram pouco adequadas para descrever a relação entre X e Y, dado que se afastam da tendência dominante. O erro cometido ao tentar-se reproduzir, com base nelas, o padrão descrito pelos pontos amostrais, é muito grande. Isso decorre do fato de que elas estão próximas de parte minoritária dos pontos amostrais.

O ideal seria, portanto, que a reta estivesse suficientemente perto de todos os pontos. Com isso, os erros cometidos ao toma-la como base seriam desprezíveis. Obviamente, não é possível traçar uma reta que atenda a esta condição. Mas é possível traçar uma reta que esteja próxima do maior número possível de pontos. O que é equivalente a procurar uma reta que cometa menos e menores erros de aproximação entre todas as retas possíveis.

Para operacionalizar este desiderato é preciso tomar por base uma medida para o total de erros cometidos. Uma possibilidade é tomar a expectativa do valor absoluto do erro de aproximação linear (i.e., a média para o valor absoluto dos erros). A intuição desta medida está em que a expectativa é uma média, e, portanto, contém a soma dos erros. Além disso, como a análise tem por objetivo inferir a distribuição populacional de Y (condicional à X), a atenção, pois, está voltada para a população. Daí porque se toma a expectativa[12].

A medida para os erros de aproximação, portanto, é:

$$E\left[\left|Y-\widetilde{Y}_i\right|\right](1)$$

Em que $\widetilde{Y}$ é o valor de Y que a reta associa a i-ésima observação.

Uma vez que o operador matemático valor absoluto ("| |") não é de fácil manipulação algébrica, toma-se o quadrado dos erros de aproximação linear, ou seja:

$$E\left[\left(Y_i-\widetilde{Y}_i\right)^2\right](1')$$

Ambos operadores, o valor absoluto e o quadrado desempenham a mesma função que é a de eliminar o sinal dos erros.

O próximo passo consiste em retomar a definição da aproximação linear à FRP, $\widetilde{Y}=\beta_0+\beta_1 X$ e a incorporar a (1').

[12] Esta abordagem para obter os estimadores de MQO é uma adaptação da seção 3.1.1 e 3.1.2 de Angrist, J.D., Pischke, J-S., 2009. Mostly harmless econometrics, an empiricist's companion. Princeton University Press, New Jersey, US.

$$E\left[\left(Y_i-\beta_0-\beta_1 X_i\right)^2\right]$$

A reta que corresponde à melhor aproximação linear à FRP é obtida escolhendo-se os valores de β_0 e β_1 que minimizam o quadrado dos erros de aproximação. É o que propõe o método de mínimos quadrados ordinários (MQO). Formalmente, o problema de minimização pode ser escrito como:

$$min_{[\beta_0,\beta_1]}\, E\left[\left(Y_i-\beta_0-\beta_1 X_i\right)^2\right]$$

A resolução deste problema requer o emprego de cálculo diferencial. O que se resume a tomar as derivadas parciais da expressão entre colchetes e igualar as expressões resultantes a zero. Assim fazendo, são obtidas as duas condições de primeira ordem, quais sejam:

$$E\left[\left(Y_i-\beta_0-\beta_1 X_i\right)\right]=0(1)$$

$$E\left[X_i\left(Y_i-\beta_0-\beta_1 X_i\right)\right]=0(2)$$

Ou, alternativamente

$$E\left[Y_i-\beta_0-\beta_1 X_i\right]=0(1)$$

$$E\left[X_i Y_i-X_i\beta_0-\beta_1 X_i^2\right]=0(2)$$

As duas equações acima compõem um sistema determinado de duas equações e duas incógnitas em que as últimas correspondem aos valores dos parâmetros β_0 e β_1. Uma maneira de obter as duas soluções é, em primeiro lugar, distribuindo as expectativas em cada equação e as escrevendo da maneira abaixo:

$$\beta_0+\beta_1 E[X\|i]=E\left[Y_i\right](1')$$

$$\beta_0 E\left[X_i\right]+\beta_1 E\left[X_i^2\right]=E\left[X_i Y_i\right](2')$$

É possível multiplicar a primeira equação por E[Xi] sem alterá-la. Fazendo isso:

$$\beta_0 E[X||i]+\beta_1 E[X\|i]^2=E\left[Y_i\right]E[X\|i](1'')$$

Subtraindo (2) de (1’’):

$$\beta_0 E[X||i]+\beta_1 E[X\|i]^2-\left(\beta_0 E\left[X_i\right]+\beta_1 E\left[X_i^2\right]\right)=E\left[Y_i\right]E[X||i]-E\left[X_i Y_i\right]\leftrightarrow$$

$$\beta_1$$

$$\beta_1=E\left[Y_i\right]\frac{E[X||i]-E\left[X_i Y_i\right]}{E[X\|i]^2-E\left[X_i^2\right]}=\frac{cov(Y,X)}{V(X)}(3)$$

Esta expressão fornece uma fórmula para calcular o coeficiente angular da reta que mais se aproxima da relação entre X e Y. A equação (1’) acima pode ser reescrita como:

$$\beta_0 = E[Y_i] - \beta_1 E[X \| i] \quad (4)$$

Partindo da equação (3), pode-se calcular o valor do intercepto da reta que fornece a melhor aproximação linear a partir de (4).

As fórmulas para os parâmetros, fornecidas pelas equações (3) e (4), contudo, apenas podem ser calculadas a partir de dados que captam toda a população. Porém, geralmente, apenas uma amostra está disponível.

O passo final consiste em aplicar o assim-chamado "princípio da analogia", que estabelece que os estimadores para os parâmetros podem ser obtidos substituindo-se momentos populacionais por momentos amostrais análogos[13]. O operador análogo, na amostra, à expectativa, da população, é a média amostral. Substituindo expectativas por médias nas equações (3) e (4) acima, chega-se a:

$$\hat{\beta}_1 = \frac{N^{-1}\sum_{i=1}^{N}(y_i - \overline{y})(x_i - \overline{x})}{N^{-1}\sum_{i=1}^{N}(x_i - \overline{x})^2} \quad (3A)$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \quad (4A)$$

Eis a fórmula dos estimadores de MQO para os parâmetros da FRP.

2.7 Valores previstos e resíduos

Uma vez obtidas as estimativas pontuais para os parâmetros, intercepto, $\hat{\beta}_0$ e coeficiente, $\hat{\beta}_1$, é possível, com base neles, obter os valores previstos, pela regressão, para a variável dependente. Basta tomar $\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i$, i=1,...,N.

A diferença entre os valores previstos e os valores observados é uma medida para os equívocos cometidos pela regressão. Na maioria dos casos, a reta de regressão (FRA) erra para um número não desprezível de observações.

Isso ocorre por dois motivos. Em primeiro lugar, há o erro de aproximação linear da FRP, uma vez que se toma uma forma linear para essa, mesmo sendo que isso não necessariamente é verdade. Em segundo lugar, há o erro de inferência, oriundo do emprego da informação disponível na amostra para inferir a FRP. Este segundo erro, portanto, diz respeito à discrepância entre a FRA e a FRP e à "qualidade" do salto de inferência.

Uma medida para o tamanho dos erros pode ser calculada como segue:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \text{ (A)}$$

Trata-se do que geralmente se denomina por resíduos da regressão.

[13] A palavra "momento" denota expectativas ou médias de potências de variáveis, o que abrange tanto a média aritmética como a média do quadrado de uma variável.

Um detalhe crucial está na diferença conceitual entre os termos de perturbação da FPR e os resíduos[14]. Os primeiros nunca são observados, exatamente porque representam todas as variáveis explicativas que influenciam a variável dependente mas que **não** são observadas. Já os resíduos são **sempre** observados e é sempre possível os calcular a partir dos dados disponíveis.

A razão apresentada no parágrafo anterior é plenamente suficiente para explicar a diferença conceitual entre erros e resíduos. Mas, para deixar mais claro que se trata de elementos distintos, pode-se recorrer à diferença algébrica, seguindo Wooldridge (p. 56). Aplicando a definição da FRP na equação (A) acima, chega-se a:

$$\hat{u}_i = \beta_0 + \beta_1 x_i + u_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad (A')$$

Após a fatoração, tem-se:

$$\hat{u}_i = u_i + \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i \quad (A'')$$

Ou, de maneira mais clara:

$$\hat{u}_i - u_i = \left(\beta_0 - \hat{\beta}_0\right) + \left(\beta_1 - \hat{\beta}_1\right) x_i \quad (A''')$$

Como o termo do lado direito não é zero, pois os valores estimados para os parâmetros geralmente não são exatamente equivalentes aos valores populacionais[15], fica demonstrando que erros e resíduos são algebricamente distintos.

Apêndice Critério alternativo para obter o estimador de MQO: método dos momentos (Woodridge, seção 2.2)

O método de mínimos quadrados ordinários é apenas um dos métodos a partir do qual é possível obter os estimadores para os parâmetros da FRP. Há dois outros métodos que também permitem chegar a eles, o método de máxima verossimilhança e o método dos momentos. Por hora, será focado o último, uma vez que ele é a base da derivação apresentada por Wooldridge na seção 2.2 de seu livro[16].

O método dos momentos não parte de uma condição de otimização, mas sim de uma hipótese, denominada condição de ortogonalidade. Esta, tal como é o caso do critério de minimização do erro quadrático médio, consiste em uma afirmação que vale para a população. Trata-se de exigir que a covariância entre o termo de perturbação e a variável independente seja nula. Formalmente:

$$\text{cov}[x_i, u_i] = 0 \text{ (MM1)}, \; i = 1, \ldots, N$$

Além disso, assume-se que a expectativa do termo de perturbação é nula.

---

[14] Gujarati comete um ato de imprecisão (ou de incorreção) ao afirmar, na p.49, que o termo de perturbação é conceitualmente análogo aos resíduos. Wooldridge, corretamente, assinala que se trata de elementos conceitualmente distintos em pelo menos três momentos do capítulo 2 de seu livro. Na p. 56 há uma explicação suficientemente clara, a qual é reproduzida no texto.

[15] Isso é verdade mesmo quando, em média, os valores estimados são equivalentes aos valores populacionais, i.e., quando os estimadores são não-viesados.

[16] Segunda edição em inglês.

$$E[u_i] = 0 \text{ (MM2)}, i=1,...,N$$

Da definição de covariância, tem-se $cov[x_i, u_i] = E[x_i u_i] - E[x_i]E[u_i] \Rightarrow cov[x_i, u_i] = E[x_i u_i]$ (*); a última passagem decorre diretamente de MM2. Levando o resultado (*) a MM1, tem-se:

$$E[x_i u_i] = 0 \text{ (MM1')}, i=1,...,N$$

As condições MM1' e MM2 são equivalentes às condições de primeira ordem do problema de minimização do erro quadrático médio, este o critério de obtenção de estimadores fornecido pelo método de mínimos quadrados. O primeiro passo para perceber isso consiste em reescrever MM1' e MM2, explorando a definição do termo de perturbação, tal como segue.

$$E[x_i (y_i - \beta_0 - \beta_1 x_i)] = 0 \text{ (MM1')}, i=1,...,N$$

$$E[y_i - \beta_0 - \beta_1 x_i] = 0 \text{ (MM2)}, i=1,...,N$$

Como segundo passo, recorre-se ao "princípio da analogia", substituindo os momentos populacionais, $E[x_i(y_i - \beta_0 - \beta_1 x_i)]$ e $E[u_i]$, por suas contrapartidas amostrais, $\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_1 x_i\right)$ e $\sum_{i=1}^{N}\left(y_i - \hat{\beta}_1 - \hat{\beta}_1 x_i\right)$, de modo a chegar em:

$$\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \text{ (MM1'')}$$

$$\sum_{i=1}^{N}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \text{ (MM2')}$$

Tem-se, pois, as equações equivalentes às que o método de MQO conduz.

Apêndice Critério alternativo para obter o estimador de MQO: método dos momentos (Woodridge, seção 2.2)

O método de mínimos quadrados ordinários é apenas um dos métodos a partir do qual é possível obter os estimadores para os parâmetros da FRP. Há dois outros métodos que também permitem chegar a eles, o método de máxima verossimilhança e o método dos momentos. Por hora, será focado o último, uma vez que ele é a base da derivação apresentada por Wooldridge na seção 2.2 de seu livro[17].

O método dos momentos não parte de uma condição de otimização, mas sim de uma hipótese, denominada condição de ortogonalidade. Esta, tal como é o caso do critério de minimização do erro quadrático médio, consiste em uma afirmação que vale para a população. Trata-se de exigir que a covariância entre o termo de perturbação e a variável independente seja nula. Formalmente:

$$cov[x_i,u_i] = 0 \text{ (MM1)}, i=1,...,N$$

Além disso, assume-se que a expectativa do termo de perturbação é nula.

[17] Segunda edição em inglês.

$$E[u_i] = 0 \text{ (MM2)}, i=1,...,N$$

Da definição de covariância, tem-se $cov[x_i, u_i] = E[x_i u_i] - E[x_i]E[u_i] \Rightarrow cov[x_i, u_i] = E[x_i u_i]$ (*); a última passagem decorre diretamente de MM2. Levando o resultado (*) a MM1, tem-se:

$$E[x_i u_i] = 0 \text{ (MM1')}, i=1,...,N$$

As condições MM1' e MM2 são equivalentes às condições de primeira ordem do problema de minimização do erro quadrático médio, este o critério de obtenção de estimadores fornecido pelo método de mínimos quadrados. O primeiro passo para perceber isso consiste em reescrever MM1' e MM2, explorando a definição do termo de perturbação, tal como segue.

$$E[x_i (y_i - \beta_0 - \beta_1 x_i)] = 0 \text{ (MM1')}, i=1,...,N$$

$$E[y_i - \beta_0 - \beta_1 x_i] = 0 \text{ (MM2)}, i=1,...,N$$

Como segundo passo, recorre-se ao "princípio da analogia", substituindo os momentos populacionais, $E[x_i(y_i - \beta_0 - \beta_1 x_i)]$ e $E[u_i]$, por suas contrapartidas amostrais, $\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}_1 - \hat{\beta}_1 x_i\right)$ e $\sum_{i=1}^{N}\left(y_i - \hat{\beta}_1 - \hat{\beta}_1 x_i\right)$, de modo a chegar em:

$$\sum_{i=1}^{N} x_i\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \text{ (MM1'')}$$

$$\sum_{i=1}^{N}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \text{ (MM2')}$$

Tem-se, pois, as equações equivalentes às que o método de MQO conduz.