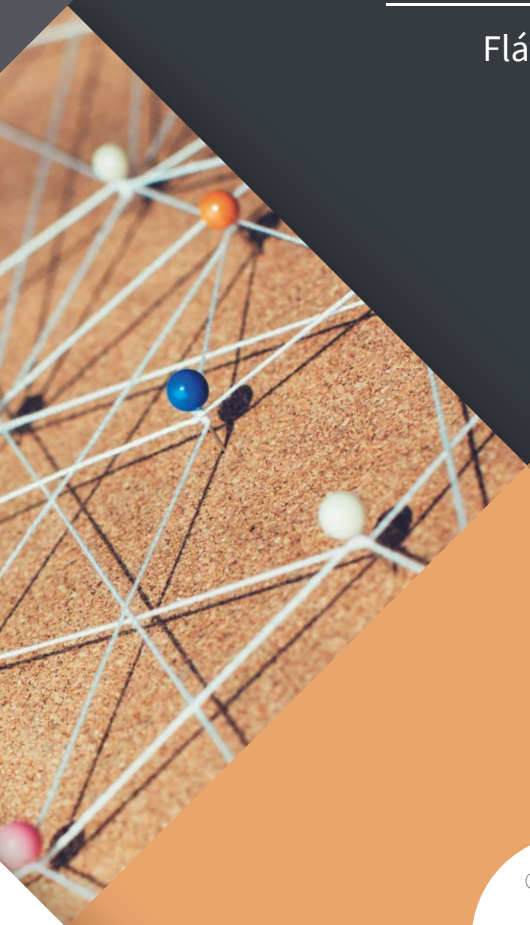
Enap

INTRODUÇÃO AOS MODELOS DE REGRESSÃO LINEAR

Flávia Chein



COLEÇÃO

Metodologias *de Pesquisa*

Introdução aos modelos de regressão linear

Um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas

Enap Escola Nacional de Administração Pública

Presidente
Diogo Godinho Ramos Costa

Diretoria de Seleção e Formação de Carreiras
Diana Magalhães de Souza Coutinho

Diretor de Educação Continuada
Paulo Marques

Diretor de Inovação e Gestão do Conhecimento
Guilherme Alberto Almeida de Almeida

Diretor de Pesquisa e Pós-Graduação
Fernando de Barros Filgueiras

Diretora de Gestão Interna
Camile Sahb Mesquita

Editor: Fernando de Barros Filgueiras. *Revisão*: Luiz Augusto Barros de Matos e Renata Fernandes Mourão. *Projeto gráfico e editoração eletrônica:* Ana Carla Gualberto Cardoso.

Introdução aos modelos de regressão linear

Um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas

Flávia Chein

Brasília – DF
Enap
2019

Ficha catalográfica elaborada pela equipe da Biblioteca Graciliano Ramos da Enap

C5157i Chein, Flávia
Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas / Flávia Chein. -- Brasília: Enap, 2019.

76 p. : il. –

Inclui bibliografia.
ISBN: 978-85-256-0115-5

1. Regressão Linear - Modelos. 2. Estatística. 3. Políticas Públicas - Avaliação 4. Econometria. I. Título.

CDU 519.24

Bibliotecária: Tatiane de Oliveira Dias – CRB1/2230

Enap Fundação Escola Nacional de Administração Pública
SAIS – Área 2-A
70610-900 – Brasília, DF
Telefones: (61) 2020 3096 / 2020 3102 – Fax: (61) 2020 3178
Sítio: www.enap.gov.br

Sumário

Apresentação – Políticas públicas e os modelos de regressão: até onde se pode chegar?

A proposta deste manual surgiu, por meio do convite do Professor Fernando de Barros Gontijo Filgueiras, da Escola Nacional de Administração Pública (Enap), com o intuito principal de levar o conhecimento de ferramentas estatísticas, extremamente úteis para o avanço da avaliação de políticas públicas, ao alcance de gestores públicos com diferentes campos de formação. Dentro dessas ferramentas, inserem-se os modelos de regressão linear.

Os modelos de regressão linear fazem parte de um conjunto de ferramentas comuns entre economistas e estatísticos cujo foco é a realização de inferências, na maior parte das vezes, causais. A inferência consiste em, a partir de evidências encontradas para uma amostra, realizar generalizações de resultados para a população. Ou, de modo mais simples, há um interesse em verificar a correlação entre duas ou mais variáveis e testar o quanto se pode confiar nas estimativas encontradas. Basicamente, os economistas se preocupam com a causalidade na perspectiva de avaliação de políticas. Nesse sentido, a identificação de parâmetros e as inferências causais em economia são motivadas por questões relacionadas a intervenções de políticas públicas (Heckman, 2008). Apenas para ilustrar, um secretário de educação pode estar interessado em estimar se seria interessante contratar mais professores e reduzir o tamanho das turmas, uma vez que, com turmas menores, o professor poderia dar uma atenção mais individualizada ao aluno, ao mesmo tempo em que o aluno teria menos possibilidade de se distrair sem ser repreendido pelo professor, por exemplo.

Portanto, antes de iniciar o estudo dos modelos de regressão linear, é preciso saber que, ao usar dados observacionais, as estimativas de regressão podem ou não ter uma interpretação causal. A proposta deste livro é apresentar a mecânica por trás dos modelos de regressão,

discutindo as chamadas hipóteses de identificação e seus pressupostos estatísticos. Além disso, é preciso estar ciente de que os modelos aqui apresentados são apenas um primeiro passo para adentrar no mundo da econometria, que avança e se transforma constantemente, seja pela introdução de novos estimadores, seja pela formulação de novos testes estatísticos.

Além desta apresentação, este livro está organizado em cinco capítulos. No primeiro capítulo, são introduzidos alguns conceitos para a compreensão dos modelos de regressão linear, como a diferença entre causalidade e associação e alguns fundamentos estatísticos. No segundo capítulo, é apresentado o modelo de regressão linear simples, suas hipóteses básicas e o método de mínimos quadrados ordinários. Já no terceiro capítulo, é discutido o modelo de regressão linear múltipla. No quarto capítulo, são apresentadas as violações das hipóteses básicas do estimador de mínimos quadrados ordinários e suas consequências. E, finalmente, no quinto capítulo, é apresentado o modelo de diferenças em diferenças, com uma ilustração de como é possível avançar na identificação de modelos de regressão linear múltipla em análise de políticas públicas.

Capítulo 1 – Fundamentos do modelo de regressão linear

Conceitos, objetivos, aplicações

Por que utilizamos a análise de regressão? Pode-se utilizar a regressão linear como um instrumento estatístico para, simplesmente, resumir dados, informações.

Na análise de regressão, a preocupação é sempre com a dependência estatística entre variáveis. Trabalha-se com variáveis aleatórias, que têm uma distribuição de probabilidade. Não há nenhum enfoque em relações determinísticas ou funcionais, típicas em ciências como a química (lei de Boyle, lei de Charles) ou física clássica (as três leis de movimento de Newton, a lei da gravidade, as leis da termodinâmica, entre outras).

De acordo com Angrist e Pischke (2009), os modelos de regressão podem ser vistos como um dispositivo computacional para estimação de diferenças entre um grupo de tratados e um grupo de controle, com ou sem covariadas. Para entender melhor o que seriam esses dois grupos e o problema por trás da comparação de seus resultados, imagine que um gestor público esteja interessado em avaliar os efeitos de uma política de financiamento estudantil sobre a decisão de cursar ensino superior, como por exemplo, do Fundo de Financiamento Estudantil (Fies), no Brasil. O Fies, a partir de 2012 até 2015, disponibiliza linhas de financiamento para estudantes, com taxas de juros abaixo do mercado, voltadas para famílias com rendimento bruto abaixo de 20 salários mínimos. O grupo de tratamento, nesse caso, seriam membros de famílias com rendimento de até 20 salários mínimos brutos e o grupo de controle, aqueles membros de famílias com rendimentos superiores a esse limite.

A partir do exemplo do Fies, pode-se pensar no que há por trás de uma análise de regressão. Se a questão é saber se o Fies foi capaz de alterar a decisão de cursar ensino superior no Brasil entre os indivíduos de

menor renda, por que não utilizar simplesmente uma comparação entre a média de anos de estudos dos dois grupos? A análise de regressão com a inclusão de covariadas nos permite considerar diferenças observáveis entre os dois grupos, como, por exemplo, a escolaridade dos pais, o fato de se trabalhar ou não, a idade, entre outras características que afetam a decisão de estudar, além da própria restrição de crédito, que estaria sendo relaxada pela política pública. Além disso, vamos discutir, mais adiante, todas as consequências da existência e como tratar as chamadas diferenças não observáveis, como esforço, habilidade, entre outras, quando estivermos falando de um modelo de regressão linear.

Tem-se, portanto, alguns pontos a serem discutidos ao longo deste manual. O primeiro deles se refere a explorar a relação entre duas ou mais variáveis, apenas como uma correlação, sem qualquer interpretação de causa e efeito; é uma forma de se resumir um conjunto de dados. Em geral, os economistas, os cientistas sociais ou gestores públicos querem mais do que isso. E por que eles querem ir além da simples correlação entre variáveis? Na maioria das vezes porque querem prever o resultado de uma intervenção ou política pública, como a de um programa de alfabetização no tempo certo sobre o desempenho acadêmico das crianças, os efeitos da construção de uma rodovia sobre a comércio exterior ou o aumento do período da licença maternidade sobre a oferta de trabalho das mulheres. Além disso, pode haver um interesse em fazer previsões sobre o comportamento de variáveis macroeconômicas, como as taxas de inflação ou crescimento do Produto Interno Bruto (PIB).

Capítulo 2 – O modelo de regressão linear simples

Como destacado no capítulo anterior, o instrumental da econometria é utilizado para analisar, qualitativamente e quantitativamente, relações entre variáveis. Chama-se de variável dependente ou variável endógena, *y*, aquela cujo comportamento será explicado pela variável *x*, chamada de variável explicativa, regressor ou variável independente. A ideia aqui é bastante simples; é, praticamente, estimar a equação de uma reta, como a do Gráfico 1. Tal equação é descrita como $y=a+bx$. O ponto central é, portanto, encontrar valores para a e b. Em outras palavras, queremos estimar a inclinação da reta utilizando uma amostra aleatória de dados de *x* e y. A inclinação nos fornece o efeito em *y* da mudança de uma unidade em *x.*

Stock e Watson (2010) trazem o exemplo da estimação do efeito do tamanho da turma, ou número de alunos por sala de aula, sobre o desempenho acadêmico dos estudantes. Nos Estados Unidos, em muitos distritos de escolas, o desempenho acadêmico é medido por testes padronizados. Poder-se-ia perguntar, para um gestor de uma superintendência regional de ensino, qual seria o efeito de se diminuir o tamanho da turma sobre o resultado médio nesses testes padronizados. A mesma pergunta poderia ser replicada para o caso brasileiro, utilizando os resultados da Prova Brasil, que é aplicada a cada dois anos, pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), no quinto e nono ano do ensino fundamental, e cuja medida de desempenho acadêmico é construída por resultados em testes padronizados. Qual relação deve-se estimar para responder tal pergunta? Uma proposta é utilizar uma equação da reta $y=a+bx,$ em que *y* será o desempenho no teste padronizado dos alunos e *x* será o número de alunos por turma. Na equação acima, a será o intercepto vertical e b é a inclinação da reta.

Mas por que é relevante tentar encontrar uma resposta para tal pergunta? Existe aí uma política pública diretamente dependente de

um resposta quantitativa para o efeito do tamanho da turma sobre o desempenho. Imagine um secretário estadual de educação decidindo se contrata mais professores ou não. Do lado dos pais/responsáveis dos alunos, há uma demanda por turmas menores sob o argumento de que, quanto menor o número de alunos, mais individualizada tenderá a ser a atenção do professor para com o aprendizado do aluno. Contudo, na ótica do gasto público, uma redução do tamanho das turmas está atrelada a um aumento de gastos. Então o que está por trás dessa avaliação é, em última instância, uma análise de benefício marginal vis-à-vis um custo marginal decorrente da necessidade de mais professores para viabilizar a redução do tamanho das turmas. Tal relação pode ser mensurada, de forma indireta, considerando-se a variação no escore padronizado no teste de desempenho do aluno, dada uma variação no número de estudantes por classe.

Dentro dessa ideia de observar relações entre variáveis, imagine um estudo que pretenda analisar os impactos das mudanças climáticas sobre a produtividade agrícola. Uma forma de apresentar evidências da relação entre variáveis de clima e resultados do setor primário seria olhar para o comportamento conjunto de uma variável de clima, por exemplo, média de temperatura no verão, e o produto do setor agropecuário. O Gráfico 1 é uma forma de descrever essa relação.

O Gráfico 1 é um gráfico de dispersão das variáveis temperatura no verão e produto agropecuário. Cada um dos pequenos pontos representa uma observação no banco de dados, ou melhor, refere-se a valores de temperatura média no verão no eixo horizontal e produto agrícola no eixo vertical para cada um dos municípios brasileiros. A linha reta é resultado exatamente de uma regressão linear em que o comportamento do produto agropecuário é explicado pela temperatura média no verão.

A ideia por trás do modelo de regressão linear é estimar uma reta que melhor descreva a relação entre variáveis. No exemplo do Gráfico 1, pode-se pensar na reta como uma forma de se resumir a informação contida na nuvem de pontos, essa é uma reta de regressão linear.

A reta de regressão depende de cinco estatísticas básicas: a) média de X ($\overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$), b) desvio-padrão de X $\left(S_X = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2}\right)$; c) média de Y; d) desvio-padrão de Y; e) correlação de X e Y $\left(r = \frac{1}{N}\sum_{i=1}^{n} \frac{(X_i-\overline{X})}{S_X}.\frac{(Y_i-\overline{Y})}{S_Y}\right)$.

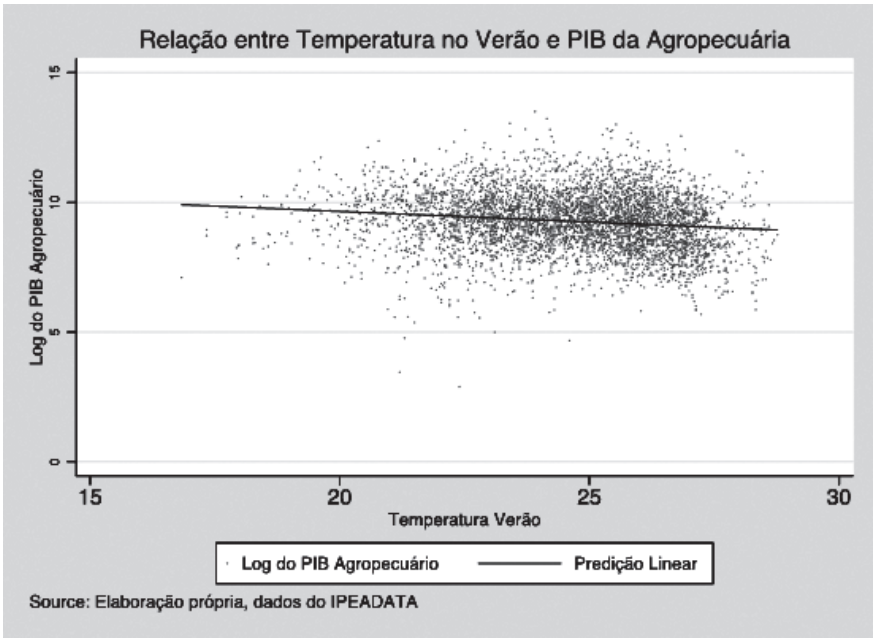
Logo, com base nessas estatísticas, podemos calcular a reta de regressão, sabendo que a regressão de Y em X passa pelos pontos médios ($\overline{X}, \overline{Y}$). A inclinação da reta é:

$$\beta_1 = \frac{r.S_y}{S_x} \quad (1)$$

O intercepto da reta de regressão populacional será:

$$\overline{Y} - \beta_1 \overline{X} \quad (2)$$

Gráfico 1 – Exemplo correlação entre clima e produto agropecuário



A reta de regressão é o arcabouço que irá permitir prever o comportamento de *Y* a partir de *X.* No exemplo representado no Gráfico 1, a reta de regressão é a linha que permite prever o comportamento do PIB agropecuário a partir da temperatura média no verão. Ressalte-se que a correlação é um importante elemento na análise de regressão, além de trazer muita informação sobre a forma do gráfico que relaciona as duas variáveis, também permite identificar se uma regressão linear resume de forma adequada a relação de interesse.

Cabe, nesse contexto, questionar qual a motivação para se pensar em uma relação como a do Gráfico 1. Bem, um ministro da agricultura, com base nas discussões recentes sobre mudança climática, poderia perguntar qual o efeito do aquecimento global sobre o PIB da agropecuária. A resposta para tal pergunta exige, assim como no caso da discussão sobre redução do tamanho da turma, uma declaração quantitativa, ou algo como: se a temperatura no verão aumentar em 1 grau Celsius, haverá uma perda de *x* reais no produto interno da área agrícola. Nesse caso, pode-se pensar em uma equação como:

$$pib_{agro} = \beta_0 + \beta_1 temp_{verão} \quad (3)$$

onde pib_{agro} (produto interno da agropecuária) é a variável dependente, β_0 é o intercepto, ou valor médio do PIB na agropecuária, β_1 é a inclinação da reta, ou o efeito de uma variação na temperatura no verão, $temp_{verão}$, sobre o pib_{agro}. Mas a equação (1) não nos permite recuperar o valor de pib_{agro}, ainda que se conheçam os valores de β_0 e β_1. E por que não? Simplesmente porque existem outros fatores além da temperatura no verão que influenciam o comportamento do PIB da agropecuária e que não estão incluídos na equação (1). Tal equação pode, nesse sentido, ser reescrita incluindo um termo de erro (ε):

$$pib_{agro} = \beta_0 + \beta_1 temp_{verão} + \varepsilon \quad (4)$$

A equação (4), que relaciona PIB da agropecuária com a temperatura no verão, pode ser reescrita de uma forma mais geral. Suponha que você possa observar os valores médios do PIB na agropecuária nos municípios do Brasil, bem com as temperaturas médias no verão, então pode-se reescrever a equação (4) como:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \text{ (5)}$$

onde Y_i é a média do PIB na agropecuária observado em cada município *i* do Brasil e representa a variável dependente; β_0 é a constante ou intercepto; X_i é o único regressor ou variável independente na equação e representa as temperaturas médias no verão observadas em cada município *i* da amostra; ε_i é o termo de erro e capta todos os fatores não observados no modelo.

Destaca-se que $Y_i = \beta_0 + \beta_1 X_i$ nos dá a equação da reta estimada, a linha de regressão populacional ou a função de regressão populacional, como apresentada no Gráfico 1, sendo β_1 o efeito médio das temperaturas no verão sobre o PIB na agropecuária. β_0 e β_1 são também os coeficientes da linha de regressão populacional, também denominados de parâmetros do modelo populacional.

Já o termo u_i, como destacam Heij *et al*. (2004) e Stock e Watson (2010), é o erro que se comete ao estimar Y_i por meio da variável X_i. Esse termo capta todos os fatores, além de X, responsáveis pela diferença na média do PIB da agropecuária no município *i* e o valor predito pela linha de regressão populacional.

Um ponto importante que é necessário destacar é o fato de que a equação (5) se refere a um modelo populacional, tal relação não é observada diretamente, então, é preciso estimá-la utilizando um modelo amostral, ou seja, um modelo estimado a partir de uma amostra da população, que se concretiza na forma de uma base de dados. O modelo amostral correspondente ao da equação (5) pode ser escrito como:

$$Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{u_i} \text{ (6)}$$

Cabe destacar que, na equação (5), temos parâmetros β_0 e β_1 e um erro u_i. Os parâmetros são constantes, com valores fixos. O problema é que, em geral, os parâmetros são desconhecidos, uma vez que é muito difícil termos dados de um população inteira. A solução, portanto, é estimar os parâmetros da equação (5), por meio de uma amostra. Nesse sentido, a equação (6), o modelo amostral, fornecerá estimativas para os parâmetros populacionais por meios dos estimadores $\widehat{\beta_0}, \widehat{\beta_1}$. Os estimadores dependem de uma amostra, estão relacionados a uma distribuição de probabilidade, são, dessa forma, variáveis aleatórias, têm valor esperado e variância.

O objetivo do modelo de regressão linear é, a partir dos valores observados na base de dados, obter valores para $\widehat{\beta_0}, \widehat{\beta_1}$ e suas variâncias.

A equação (6) é, portanto, o ponto de partida para se pensar o modelo de regressão linear simples. Cabe, a partir daqui, discutir como estimá-la. Na seção seguinte, será apresentado o estimador linear de mínimos quadrados ordinários, cuja proposta é estimar a reta apresentada no Gráfico 1, de modo que as distâncias de cada ponto observado até a reta seja minimizada.

2.1 O estimador de mínimos quadrados ordinários

De acordo com Stock e Watson (2010), o estimador de mínimos quadrados ordinários escolhe os coeficientes de modo que a linha de regressão estimada fique o mais próxima possível dos dados observados. Como medir tal proximidade? A proximidade é medida pela soma dos quadrados dos resíduos obtidos ao se medir *Y* dado *X*.

Assim, dado o modelo amostral, $Y_i = \widehat{\beta_0} + \widehat{\beta_1} X_i + \widehat{u_i}$, $\widehat{\beta_0}$ e $\widehat{\beta_1}$ são encontrados de modo que o resíduo da regressão de *Y* em *X* para a *i* observação é $Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$. A soma dos resíduos das predições de Y_i a partir do modelo amostral para todas as *n* observações será:

$$\sum_{i=1}^{n} \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i\right)^2 \quad (7)$$

Logo, o estimador mínimos quadrados ordinários (MQO) para a inclinação β_1 e para o intercepto β_O será:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n}(X_i-\bar{X})^2} = \frac{S_{XY}}{S_X^2} \quad (8)$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \quad (9)$$

Os valores preditos por MQO para a variável dependente *Y* podem ser encontrados por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \text{ para } i=1..n \quad (10)$$

Já os resíduos são definidos do modelo amostral como:

$$\hat{u}_{i=} Y_{i-} \hat{Y}_i, \text{ para i=1..n} \quad (11)$$

Retornando ao exemplo do Gráfico 1, quando o estimador de MQO é utilizado para estimar a relação da temperatura no verão com o produto da agropecuária (medido em logaritmo), utilizando 4.967 observações referentes a municípios brasileiros com informações disponíveis, a inclinação é -0,081 e o intercepto é igual a 11,286. A linha de regressão estimada por MQO é, portanto:

$$lnPIB_{agro_i} = 11{,}286 - 0{,}081 tempver_i + \hat{u}_i \quad (12)$$

O Quadro 1 traz o detalhamento dessa estimação realizada com a utilização do pacote econométrico Stata© 14, por meio da linha de comando *reg lagro tempver.* Além dos coeficientes estimados, o Quadro 1 traz ainda outras informações, como o erro padrão dos coeficientes, intervalos de confiança e medidas de ajuste do modelo, que serão apresentadas mais adiante.

Quadro 1 – Estimação por MQO da relação entre PIB agropecuário e a temperatura média no verão (Exemplo 1)

Source	SS	df	MS	Number of obs	=	4,967
				F(1, 4965)	=	104.39
Model	125.282521	1	125.282521	Prob > F	=	0.0000
Residual	5958.42312	4,965	1.20008522	R-squared	=	0.0206
				Adj R-squared	=	0.0204
Total	6083.70564	4,966	1.22507161	Root MSE	=	1.0955

lagro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tempver	-.0815801	.0079845	-10.22	0.000	-.0972332	-.0659271
_cons	11.28602	.197066	57.27	0.000	10.89969	11.67236

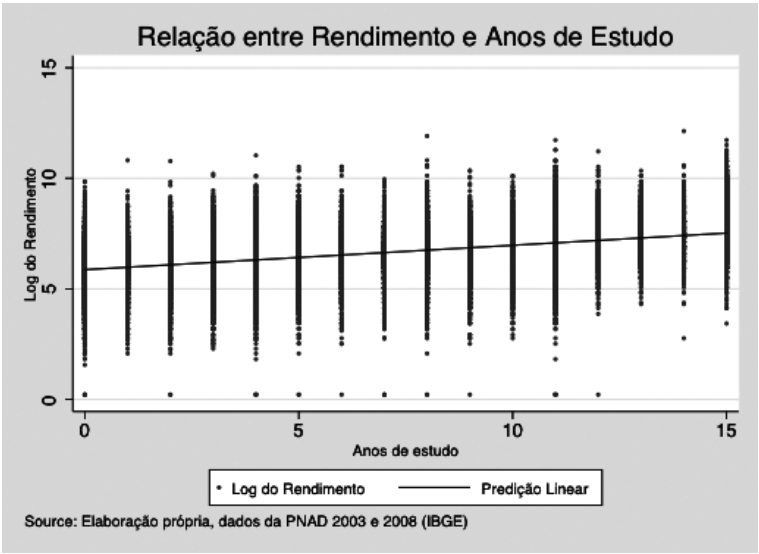
Fonte: elaboração própria a partir de dados da Ipeadata.

É importante destacar que, como, no exemplo, a variável dependente $lnPIB_{agro}$ está sendo medida em logaritmo, o coeficiente de -0,081 significa que o aumento de 1 grau Celsius na temperatura no verão geraria uma queda de cerca de 8% do PIB na agropecuária. Quando se mede a variável dependente em log e a variável independente em nível, tem-se o que é chamado de modelo log-nível. A interpretação usual, nesse caso, para o $\hat{\beta}_1$, é a de semielasticidade, em que:

$$\frac{\%\Delta E(Y|X)}{\Delta X} = 100.\hat{\beta}_1 \quad (13)$$

O estimador de MQO pode ser utilizado para estimar diferentes relações entre uma variável dependente e outra independente, que trazem em si questões importantes em termos de políticas públicas. Pense, por exemplo, na questão do investimento em capital humano, ou, especificamente, em educação formal. Um dos argumentos para se investir em educação está atrelado ao retorno salarial. Nesse contexto, a relação entre o rendimento do trabalho principal e os anos de estudo ou escolaridade pode ser estimada por MQO com base em informações da Pesquisa Nacional por Amostragem de Domicílios (PNAD), coletada pelo IBGE. O Gráfico 2 descreve essa relação.

Gráfico 2 – Exemplo correlação entre anos de estudo e rendimento do trabalho



É importante notar que a nuvem de dispersão representada no Gráfico 1 é bastante distinta da representação do Gráfico 2. Isso porque a variável independente, nesse último caso, é uma variável discreta, com valores inteiros de 0 a 15, daí, em vez de se enxergar uma nuvem de pontos, há várias fileiras de pontos. A linha reta, assim como no Gráfico 1, representa a reta de regressão linear. O modelo estimado por MQO para os microdados da PNAD para os anos de 2003 e 2008, considerando um total de 178.131 indivíduos do sexo masculino, pode ser expresso como:

$$\ln\text{rendpri} = 5{,}87 + 0{,}11anest + \hat{u}_i \quad (14)$$

Assim como no exemplo anterior, o modelo em (14) também apresenta a variável dependente em logaritmo. A estimativa do coeficiente da variável anos de estudo (anest) mostra que o aumento de um ano de estudo gera um aumento de rendimento em torno de 11%. O Quadro 2, mostra os detalhes do modelo estimado.

Quadro 2 – Estimação por MQO da relação entre rendimento do trabalho principal e anos completos de estudo (Exemplo 2)

Source	SS	df	MS	Number of obs	=	178,131
				F(1, 178129)	=	65033.71
Model	41760.6855	1	41760.6855	Prob > F	=	0.0000
Residual	114383.596	178,129	.642139101	R-squared	=	0.2674
				Adj R-squared	=	0.2674
Total	156144.281	178,130	.876574869	Root MSE	=	.80134

lnrendpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anest	.1106838	.000434	255.02	0.000	.1098331	.1115345
_cons	5.867154	.0037335	1571.50	0.000	5.859837	5.874472

Fonte: elaboração própria a partir de dados da PNAD 2003 e 2008.

Cabe, nesse ponto, questionar: por que utilizar o estimador de MQO para estimar a relação entre temperatura no verão e o comportamento do PIB na agropecuária, bem como a relação entre rendimento e escolaridade, além de tantas outras? A resposta é que, sob determinadas hipóteses, que serão apresentadas a seguir, o estimador de MQO é consistente e não apresenta viés. Além disso, sob hipóteses adicionais, também será um estimador eficiente, ou seja, de menor variância.

2.1.1 Hipóteses do modelo linear simples

O estimador de MQO apresenta um conjunto de quatro hipóteses fundamentais sobre as quais se sustenta o seu uso como o melhor estimador linear não viesado:

HIPÓTESE 1: O MODELO É LINEAR NOS PARÂMETROS.

A primeira hipótese se refere ao fato de o modelo ser linear nos parâmetros, ou seja, os betas do modelo populacional (5) entram de forma linear na equação.

HIPÓTESE 2: A AMOSTRAGEM É ALEATÓRIA.

A segunda hipótese diz que existe uma amostra aleatória de tamanho N, $f(X_i,Y_i)$, $i=1...N$, proveniente de um modelo populacional. Cabe

lembrar que, em muitos casos, problemas de seleção amostral estarão presentes, daí será necessário tratar de forma especial os casos em que a hipótese de amostragem aleatória não estiver presente. Pode-se pensar no exemplo de retornos salariais do investimento em capital humano. Em geral, em base de dados com informações individuais, como a PNAD e o Censo Demográfico, observam-se os rendimentos do trabalho apenas para aqueles trabalhadores que estão ocupados, ou seja, empregados, logo, a amostra de rendimentos pode apresentar um viés de seleção, sob a hipótese de que os trabalhadores ocupados são aqueles mais produtivos.

HIPÓTESE 3: VARIAÇÃO AMOSTRAL DA VARIÁVEL INDEPENDENTE (X).

A terceira hipótese estabelece que os resultados amostrais de *X* não têm todos o mesmo valor. Essa hipótese também é conhecida como a hipótese de variabilidade do regressor. Na maior parte das aplicações, esta hipótese sempre estará presente. Faz pouco sentido tentar explicar a variação de *y* por variações em *X,* se *X* não varia.

HIPÓTESE 4: MÉDIA CONDICIONAL DO ERRO IGUAL A ZERO.

O erro tem valor esperado igual a 0 dado *X*. A quarta hipótese se refere ao fato de que a distribuição condicional dos erros, dada a variável independente, apresenta média zero. Em termos matemáticos, tem-se que:

$$E(u_i|X_i)=0 \quad (15)$$

Como explicam Stock e Watson (2010), esta hipótese é uma afirmação matemática formal sobre os "outros fatores" contidos nos erros (u_i) e assevera que esses outros fatores são não correlacionados com X_i, de modo que, dado um valor de X_i, a média da distribuição desses outros valores é zero. Voltando ao exemplo de retornos à educação, a hipótese 4 estabelece que numa equação de rendimentos, não há nada no erro que seja correlacionado com a decisão de investimento em educação e que afete o rendimento, como, por exemplo, habilidades inatas.

Assim, sob as hipóteses **H1 – H4**, os estimadores de MQO são estimadores não viesados dos parâmetros da população, ou seja;

Propriedade 1 dos estimadores de MQO: $\mathbb{E}\left[\hat{\beta}_0 \middle| X\right] = \beta_0$, $\mathbb{E}\left[\hat{\beta}_1 \middle| X\right] = \beta_1$ (16)

- O estimador de MQO é um estimador NÃO VIESADO.
- A distribuição de $\hat{\beta}_1$ está centrada no lugar correto.

É importante compreender que condicionar nos valores de X é o mesmo que manter X_i como se fosse fixo em amostras repetidas. Como se fixássemos X_i em n valores, e construíssemos amostras de Y baseados em amostras de ε.

Mas, esta propriedade não garante que a estimativa que obtemos usando certa base de dados é "correta".

- Hipótese-chave para esse resultado: $\mathbb{E}[\varepsilon|X] = 0$

Pode-se perguntar, portanto, se o valor estimado do retorno à educação de 11% apresenta viés ou não? Ou, em outras palavras, o que poderia diferenciar o valor esperado de $\hat{\beta}$, do verdadeiro parâmetro populacional? Uma das possíveis fontes de viés está relacionada à omissão de variável relevante, que fará com que a hipótese 4 seja violada.

A violação da hipótese 4, média condicional do erro igual a zero, ocorrerá, no modelo de regressão linear simples, se houver alguma variável omitida correlacionada com a variável independente e que também afete a variável dependente.

No exemplo da estimação do efeito de anos completos de estudo sobre o rendimento, é preciso pensar em alguma variável omitida que seja correlacionada com anos completos de estudo e que também afete o rendimento do trabalho principal, por exemplo, aptidão. A ideia aqui seria a de que aptidão está correlacionada com o investimento em capital humano e também com o rendimento do trabalho principal, uma vez que aptidão pode afetar também a produtividade do trabalho. A não inclusão dessa variável fará com que a variável independente *x* esteja correlacionada com os resíduos da regressão.

HIPÓTESE 5: o erro tem a mesma variância para qualquer valor da variável explicativa:

$$\mathbb{V}ar[\varepsilon|X] = \sigma^2 \ (17) \qquad (17)$$

- σ^2 é a variância incondicional do erro.

- σ é o desvio-padrão do erro.

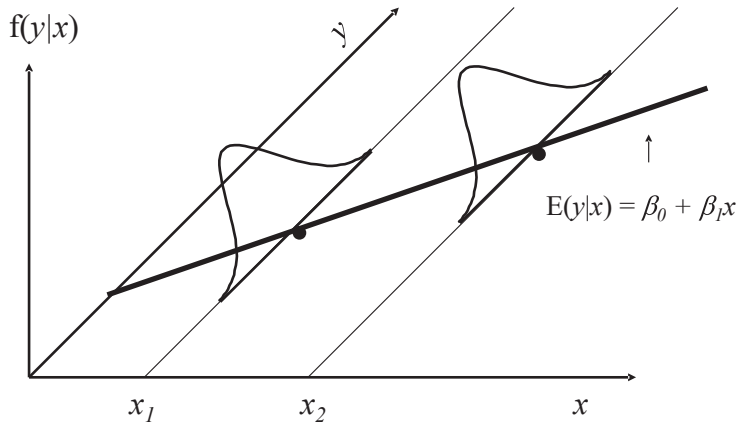
- Esta hipótese é conhecida como homocedasticidade, ou variância constante dos erros.

Sob as hipóteses **H4** e **H5,** podemos calcular a esperança e variância de Y

$$\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i \ (18)$$

$$\mathbb{V}ar[Y_i|X_i] = \sigma^2 \ (19)$$

Gráfico 3 – Ilustração do caso de homocedasticidade



Fonte: Anderson (2002)

Voltando à estimação da equação de salário (Quadro 2), o nosso modelo populacional sob **H4** é:

$$\mathbb{E}[\text{lnsalário}|\text{anest}] = \beta_0 + \beta_1 \text{anest}$$

Como a hipótese $\mathbb{V}ar[\varepsilon|\text{anest}]$ afeta o nosso modelo populacional? O salário médio pode crescer com educação, no entanto a taxa de crescimento do salário (ou a variabilidade do salário em relação à média) é constante para todos os valores de anos de estudo.

2.1.2 Medidas de ajuste do modelo amostral

O passo seguinte à estimação do modelo de regressão linear é investigar o quão bem a linha de regressão descreve os dados. Trata-se de verificar o ajuste do modelo ou, como colocam Stock e Watson (2010), se o regressor explica muito ou pouco da variação na variável dependente e quão dispersas estão as observações ao redor da linha de regressão.

Duas medidas fundamentais para entender o ajuste de uma regressão são o R^2 e os erros padrões. O R^2 varia entre 0 e 1 e mede qual a fração da variância de *Y* é explicada por *X*. Já o erro padrão de uma regressão mede o quão distante a variável dependente *Y* está do seu valor predito.

2.1.2.1 O R^2

Matematicamente, o R^2 pode ser escrito como a proporção dos quadrados dos resíduos explicados dada a soma total dos quadrados. Como explicam Stock e Watson (2010), A **soma dos quadrados explicados (SQE)** é definida como a soma dos quadrados dos desvios dos valores preditos de Y_i em relação à sua média. Já a **soma dos quadrados totais (SQT)** é a soma dos desvios de Y_i da sua média.

$$SQE = \sum_{i=1}^{n}\left(\widehat{Y}_i - \bar{Y}\right)^2 \quad (20)$$

$$SQT = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad (21)$$

O R^2 é a razão entre a soma dos quadrados explicados sobre os quadrados totais:

$$R^2 = \frac{SQE}{SQT} \quad (22)$$

De forma alternativa, como esclarecem Stock e Watson (2010), o R^2 pode ser definido em termos da fração da variância de Y_i que não é explicada por *X*, ou seja:

$$R^2 = 1 - \frac{SQR}{SQT} \text{ (23)}$$

onde *SQR* é a soma dos quadrados dos resíduos.

O R^2 irá variar entre 0 e 1, quanto mais próximo de 1, maior será o poder de explicação ou o ajuste do modelo de regressão. Voltando ao exemplo do Quadro 2, verifica-se que o modelo de regressão linear simples proposto explica cerca de 26% do comportamento do rendimento do trabalho principal.

2.1.2.2 O erro padrão de uma regressão

O erro padrão de uma regressão é um estimador do desvio padrão dos erros da regressão. Stock e Watson (2010) esclarecem que os erros-padrão são uma medida de dispersão das observações ao redor da linha de regressão, adotando a mesma medida da variável dependente. Nesse sentido, se a variável dependente for rendimentos do trabalho em reais, o erro padrão dá a magnitude do desvio típico da reta de regressão em reais.

Como os erros da regressão não são conhecidos, o estimador do desvio padrão da regressão (*SER – Standard Error of Regression*) é calculado a partir dos resíduos da regressão por mínimos quadrados ordinários, isto é:

$$SER = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 \text{ (24)}$$

em que $\hat{u}_i^2$ são os resíduos estimados da regressão por MQO ao quadrado.

Destaca-se que a divisão por *n-2* na equação (24) se faz necessária em função do fato de que dois parâmetros da regressão são estimados (β_0 e β_1). Como destacam Stock e Watson (2010), essa é a chamada correção de graus de liberdade.

2.1.3 A significância de uma estimativa

Como destacam Heij *et al*. (2004), o objetivo do modelo de regressão linear simples é explicar a variação em uma variável dependente *y* em termos de variações em uma variável explicativa *x*. Tal proposta só faz sentido se *y* estiver de fato relacionado a *x*, ou seja, voltando ao exemplo da estimação da relação do clima sobre produto agropecuário, se β_1, na equação (5), for diferente de zero. Desse modo, queremos aplicar um teste para a hipótese nula de que $\beta_1 = 0$, contra a hipótese alternativa que $\beta_1 \neq 0$. A hipótese nula será rejeitada se o coeficiente diferir significativamente de zero.

A hipótese H_0: $\beta_1 = 0$, pode ser testada usando a estatística:

$$\frac{b-\beta}{\sqrt{Var(b)}} = \frac{b}{\sqrt{Var(b)}} \quad (25)$$

que tem distribuição *t* com *n – 2* graus de liberdade.

Mas, como podemos calcular a variância dos estimadores? Sob as hipóteses **H1-H5**

$$\mathbb{V}ar\left[\hat{\beta}_1 \middle| X\right] = \frac{\sigma^2}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \quad (26)$$

$$\mathbb{V}ar\left[\hat{\beta}_0 \middle| X\right] = \frac{\sigma^2 \frac{\sum_{i=1}^{N} X_i^2}{N}}{\sum_{i=1}^{N}\left(X_i - \overline{X}\right)^2} \quad (27)$$

Olhando para $\mathbb{V}ar\left[\hat{\beta}_1\right]$:

- Quanto maior a variabilidade do erro, maior a variabilidade de $\hat{\beta}_1$
- Quanto maior a variabilidade de X, menor a variabilidade de $\hat{\beta}_1$

A equação (26) deixa claro que, para estimar a variância dos estimadores de MQO, temos que estimar a variância do erro. Então, como estimar a variância do erro?

Primeiro, pela fórmula da variância

$\mathbb{E}[\varepsilon^2] = \sigma^2$

Um possível estimador seria:

$\frac{1}{N}\sum_{i=1}^{N} \varepsilon_i^2$ (28)

Mas não podemos obter esse estimador, pois não observamos o erro. Podemos substiuir o erro pelo resíduo.

$\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i^2$ (29)

Mas esse estimador é VIESADO. Por quê? Ele não leva em consideração as duas restrições que devem ser satisfeitas pelos resíduos de MQO: $\sum_{i=1}^{N} \hat{\varepsilon}_i = 0$ e $\sum_{i=1}^{N} x_i\hat{\varepsilon}_i = 0$. Se você conhece $n-2$ resíduos, os outros são determinados por essas restrições. Assim, os resíduos de MQO tem $n-2$ graus de liberdade e não n.

Um estimador NÃO viesado para σ^2 é:

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-2}$ (30)

Sob **H1-H5**

$\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

Usando esse estimador, podemos obter o desvio-padrão para $\hat{\beta}_1$ e $\hat{\beta}_0$:

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{N} (X_i - \overline{X})^2}} \quad (31)$$

$$sd[\hat{\beta}_0] = \sigma \cdot \sqrt{\frac{\frac{\sum_{i=1}^{N} X_i^2}{N}}{\sum_{i=1}^{N} (X_i - \overline{X})^2}} \quad (32)$$

Como $\hat{\beta}_1$ e $\hat{\beta}_0$ são estimadores, eles não apenas variam de amostra para amostra, mas, para certa amostra, eles são depedentes um do outro. Essa tendência é medida pela covariância entre eles.

$$Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{X}Var(\hat{\beta}_1)$$

$$= -\overline{X}\left(\frac{\sigma^2}{\sum_{i=1}^{N} (X_i - \overline{X})^2}\right) \quad (33)$$

Propriedade 2 dos estimadores de MQO: Sob **H1-H5**, os estimadores são eficientes. Eles têm a variância mínima na classe de todos os estimadores lineares não viesados. Para fazer inferência, precisamos de encontrar a distribuição dos estimadores. Essa distribuição será uma função da distribuição do erro.

HIPÓTESE 6 (distribuição normal): cada ε_i tem uma distribuição normal condicional a X_i

$$\mathbb{E}[\varepsilon_i|X_i] = 0$$

$$\mathbb{V}ar[\varepsilon_i|X_i] = \mathbb{E}\left[\varepsilon_i^2\middle|X_i\right] = \sigma^2$$

$$\mathbb{C}ov(u_i, u_j) = 0$$

De outra forma

$$\varepsilon_i|X_i \sim \mathcal{N}(0, \sigma^2) \text{ (34)}$$

Qualquer função linear de variáveis aleatórias com distribuição normal também é normalmente distribuída. $\hat{\beta}_0$ e $\hat{\beta}_1$ são funções lineares de $\varepsilon_i \Rightarrow \hat{\beta}_0$ e $\hat{\beta}_1$ tem distribuição normal.

Sob **H1-H6**, $\hat{\beta}_1$ tem uma distribuição normal condicional a X_i com

$$\mathbb{E}\left[\hat{\beta}_1\middle|X\right] = \beta_1$$

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

De outra maneira

$$\hat{\beta}_1|X \sim \mathcal{N}\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right) \text{ (35)}$$

$$Z = \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \sim \mathcal{N}(0,1)$$

Sob **H1-H6**, $\hat{\beta}_0$ tem uma distribuição normal condicional a X_i com

$$\mathbb{E}\left[\hat{\beta}_0\middle|X\right] = \beta_0$$

$$\sigma^2_{\hat{\beta}_0} = \frac{\sigma^2 \frac{\sum_{i=1}^{N} X_i^2}{N}}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$$

De outra maneira

$$\hat{\beta}_0 | X \sim \mathcal{N}\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right) \text{ (36)}$$

$$Z = \frac{\hat{\beta}_0 - \beta_0}{\sigma_{\hat{\beta}_0}} \sim \mathcal{N}(0,1)$$

Sob **H1-H6**,

$$(N-1)\left(\frac{\hat{\sigma}^2}{\sigma^2}\right) \sim \mathcal{X}^2_{N-2}$$

A distribuição de ($\hat{\beta}_0$, $\hat{\beta}_1$) é independente de σ^2.

Sob **H1-H6**,

$$Y_i | X_i \sim \mathcal{N}(\beta_0 + \beta_1 X_i, \sigma^2) \text{ (37)}$$

Conhecendo-se as distribuições dos estimadores, pode-se, agora, construir os intervalos de confiança para as estimativas decorrentes do modelo de mínimos quadrados ordinários. Utilizando-se os resultados anteriores, tem-se que:

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim \mathcal{N}(0,1) \text{ (38)}$$

No entanto, não observamos $\sigma_{\hat{\beta}_1}$. Na prática, usamos um estimador não viesado $\hat{\sigma}_{\hat{\beta}_1}$ Substituindo $\hat{\sigma}_{\hat{\beta}_1}$ por $\sigma_{\hat{\beta}_1}$, temos:

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{(\hat{\beta}_1 - \beta_1) \cdot \sum_{i=1}^{N} (X_i - \overline{X})^2}{\hat{\sigma}} \text{ (39)}$$

$$t \sim t - \text{student}_{N-2}$$

O intervalo de confiança parte do fato de que:

$$\Pr\left[-t_{\frac{\alpha}{2}} \leq t \leq t_{\frac{\alpha}{2}}\right] = 1 - \alpha \text{ (40)}$$

No caso dos estimadores de MQO:

$$\Pr\left[-t_{\frac{\alpha}{2}} \leq \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \leq t_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

$$\Rightarrow \Pr\left[\hat{\beta}_1 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_1}\right] = 1 - \alpha \text{ (41)}$$

Da mesma maneira,

$$\Pr\left[\hat{\beta}_0 - t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + t_{\frac{\alpha}{2}} \cdot \hat{\sigma}_{\hat{\beta}_0}\right] = 1 - \alpha \text{ (42)}$$

De (41) e (42) pode-se notar que a amplitude do intrevalo de confiança depende do desvio-padrão. O desvio-padrão é uma medida da precisão do esimador, ou seja, da exatidão com que o estimador mede o valor da população.

De volta ao **Exemplo 2**: rendimento no trabalho principal e anos de estudo completos.

Voltando ao modelo do Quadro 2, em que o comportamento do rendimento no trabalho principal em logaritmo é explicado pelos anos de estudo completos, encontramos:

$$\widehat{\log(\text{rendpri})}_i = \underset{(0.0037335)}{5.867154} + \underset{(0.000434)}{0.11068}\text{anest}$$

$$R^2 = 0.2674; N = 178.131$$

Vamos calcular o intervalo de confiança para cada um dos parâmetros β_1 e β_0

$$N = 178.131, \alpha = 0.05, t_{0.025} = 1.96$$

IC_{β_0}=[5,867154-1,96*0,0037335; 5,867154+1,96*0,0037335]=[5,859837;5,874472]

IC_{β_1}=[0,1106838-1,96*0,000434; 0,1106838+1,96*0,000434]=[0,1098331;0,1115345]

Os valores dos intervalos de confiança calculados acima podem ser checamos no cálculo realizado pelo programa STATA© e reportados no Quadro 2.

Podemos, ainda, calcular o intervalo de confiança para σ^2, considerando a distribuição chi-quadrado

$$\mathcal{X}^2 = (N-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \mathcal{X}^2_{N-2}$$

Para conseguir um intervalo de confiança para distribuição chi-quadrado, temos que levar em conta que a distribuição é assimetrica:

$$\Pr\left[\mathcal{X}^2_{1-\frac{\alpha}{2}} \leq \mathcal{X}^2 \leq \mathcal{X}^2_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Substituindo e reorganizando a expressão acima, temos

$$\Pr\left[(N-2)\frac{\widehat{\sigma}^2}{\mathcal{X}^2_{\frac{\alpha}{2}}} \leq \sigma^2 \leq (N-2)\frac{\widehat{\sigma}^2}{\mathcal{X}^2_{1-\frac{\alpha}{2}}}\right] = 1 - \alpha \text{ (43)}$$

Ressalte-se que o intervalo de confiança nos dá a região de aceitação do teste de hipótese. Suponha que queremos testar se β_1 = 0.08. Se 0.08 estiver dentro do intervalo de confiança, não rejeitamos H_0. Se cair fora, rejeitamos H_0.

É exatamente essa ideia que utilizamos para testar a significância dos coeficientes da regressão. Formalmente temos:

- Hipotése nula: β_1 = 0

- Estatística:

$$t = \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \text{ (44)}$$

- Regra de decisão: rejeitamos quando $|t| > c$

- Distribuição da estatística: $t \sim t - student_{N-2}$

- Rejeitamos para $|t| > t_{\frac{\alpha}{2}}$

- Região de aceitação:

$$\Pr\left[-t_{\frac{\alpha}{2}} \leq \frac{\widehat{\beta}_1 - \beta_1}{\widehat{\sigma}_{\widehat{\beta}_1}} \leq t_{\frac{\alpha}{2}}\right] = 1 - \alpha$$

Retornando aos exemplos dos Quadros 1 e 2, constatamos que os coeficientes das variáveis *tempver* (Quadro 1) e *anest* (Quadro 2) são significativos a 1%. Note que o p-valor é reportado na coluna (p>|t|), em ambos os quadros. O p-valor é o nível de significância exato ou observado. É o menor nível de signficância ao qual a hipótese nula pode ser rejeitada. Se o p-valor é menor que α, temos evidência que podemos rejeitar a hipótese nula a 5%

Além do teste *t,* anteriormente apresentado, a análise de regressão pode ser complementada pelo estatística *F*, que decorre da decomposição da variância. Já sabemos que:

$$\underbrace{\sum_{i=1}^{N}\left(Y_i-\overline{Y}\right)^2}_{STQ}=\underbrace{\sum_{i=1}^{N}\left(\hat{Y}_i-\overline{Y}\right)^2}_{SQE}+\underbrace{\sum_{i=1}^{N}\hat{\varepsilon}_i^2}_{SQR}$$

Cada um dos termos apresenta os seguintes graus de liberdade:

STQ	$N-1$
SQE	1
SQR	$N-2$

A estatística *F* é definida como:

$$F=\frac{\frac{SQE}{1}}{\frac{SQR}{N-2}} \quad (45)$$

Podemos usar F para testar a hipótese nula: $H_0=\beta_1=0$. Sob a hipótese de que os erros são normalmente distribuídos e sob H_0:

$$F\sim\mathcal{F}_{1,N-2}$$

A ideia subjacente ao teste é a de que, se $\beta_1 = 0$, a variável X não tem qualquer influência sob Y e toda variação de Y é explicada pelos resíduos. Desse modo, a hipótese nula é rejeitada para F grande, $F>\mathcal{F}_{\alpha}$

É importante salientar que o teste de t e o teste de F são duas formas alternativas, mas complementares, de testar a hipótese de que $\beta_1 = 0$. Vamos ver que o teste de F é muito importante para o caso de regressão múltipla.

Logo, para avaliar a qualidade de ajuste de um modelo de regressão linear, além de analisar se o R^2 é alto ou baixo, precisamos verificar se os sinais dos coeficientes estão de acordo com a teoria e se o modelo satisfaz as propriedades do modelo de regressão linear, ou seja, as hipóteses H1-H6, anteriormente apresentadas.

Capítulo 3 – O modelo de regressão linear múltipla

Vimos, no capítulo anterior, que, no modelo de regressão simples, como reportado a seguir:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

a hipótese-chave é a de que a média dos erros condicionada na variável explicativa é igual a zero, ou seja:

$$\mathbb{E}[\varepsilon_i | X_i] = 0$$

Uma forma simples de compreender essa hipótese é admitir que não existe nenhum fator não observado em ε que seja **correlacionado** com *X*, ou seja, com a variável explicativa. Será que essa hipótese é de fato realista? Na verdade não. Muitas vezes, existem vários outros fatores que podem estar, assim como *X*, afetando a nossa variável de resposta e serem também correlacionados com *X.*

Nesse sentido, surge a proposta do modelo de regressão linear múltipla. O conceito por trás desse modelo é o de *ceteris paribus*. Tal expressão tem suas origens no latim e é muito utilizada nos modelos econômicos. A ideia é de que "tudo o mais constante", ou mantendo-se outros fatores fixos, podemos estimar o efeito de *X* (variável explicativa) sobre *Y* (variável explicada ou dependente).

Logo, no modelo de regressão múltipla, por exemplo, com duas variáveis explicativas, isto é:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i \quad (46)$$

temos β_1 e β_2 como coeficientes **parciais** de regressão.

Vamos, novamente, retornar aos exemplos dos Quadros 1 e 2, mas, agora, iremos adicionar outras variáveis explicativas.

Exemplo 1: relação entre o PIB da agropecuária e a temperatura média no verão. Vamos acrescentar, como variável explicativa, além da temperatura média no verão, a precipitação no verão.

O nosso interesse é no efeito na temperatura média no verão sobre o PIB da agropecuária, mantendo-se constante o nível de precipitação no verão, ou condicional ao nível de preciptação no verão. O Quadro 3 reporta as estimações para o novo modelo de regressão linear múltipla.

Quadro 3 – Exemplo 1 revisto

Source	SS	df	MS	Number of obs	=	4,967
				F(2, 4964)	=	201.33
Model	456.462828	2	228.231414	Prob > F	=	0.0000
Residual	5627.24281	4,964	1.13361056	R-squared	=	0.0750
				Adj R-squared	=	0.0747
Total	6083.70564	4,966	1.22507161	Root MSE	=	1.0647

lagro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tempver	-.0326518	.0082713	-3.95	0.000	-.0488672	-.0164364
precver	.0036747	.000215	17.09	0.000	.0032533	.0040962
_cons	9.449757	.2196033	43.03	0.000	9.019237	9.880277

Fonte: elaboração própria.

Nesse novo modelo, para termos estimadores não viesados, vamos assumir que o termo de erro é não correlacionado com a temperatura média no verão e com a precipitação média no verão.

Exemplo 2: efeito dos anos de estudo sobre o rendimento do trabalho principal. Nesse caso, vamos agora incluir também a experiência como uma variável explicativa, além dos anos de estudo.

Quadro 4 – Exemplo 2 revisto

Source	SS	df	MS	Number of obs	=	178,131
				F(2, 178128)	=	43958.14
Model	51599.0188	2	25799.5094	Prob > F	=	0.0000
Residual	104545.263	178,128	.586910888	R-squared	=	0.3305
				Adj R-squared	=	0.3304
Total	156144.281	178,130	.876574869	Root MSE	=	.7661

lnrendpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anest	.1227917	.0004253	288.68	0.000	.121958	.1236254
exp	.0002554	1.97e-06	129.47	0.000	.0002515	.0002593
_cons	5.395011	.0051028	1057.27	0.000	5.38501	5.405013

Fonte: elaboração própria.

Assim, como nas estimações com uma única variável explicativa, é utilizado, nos modelos de regressão múltipla, o estimador de mínimos quadrados ordinários (MQO). Logo, considerando o modelo de regressão linear múltipla:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

temos como hipótese central:

$$\mathbb{E}[\varepsilon_i | X_{1i}, X_{2i}] = 0 \quad (47)$$

A hipótese acima estabelece que o valor esperado do erro é o mesmo para qualquer combinação possível entre X_{1i} e X_{2i}. Desse modo, não existe correlação entre as variáveis no termo de erro e X_1 e X_2.

Exemplo 1: retornando à estimação do efeito da temperatura no verão sobre o PIB da agropecuária:

$$lpib_{agro} = \beta_0 + \beta_1 temp_{verão\ i} + \beta_1 prec_{verão\ i} + \varepsilon_i \quad (48)$$

Hipótese-chave:

$$\mathbb{E}[\varepsilon_i | \text{temp}_{verãoi}, prec_{verãoi}] = 0 \quad (49)$$

A hipótese anterior estabelece que outros fatores que afetam o PIB, e são capturados pelos erros, não são correlacionados na média com a temperatura no verão e a precipitação no verão. No caso de variáveis geográficas, parece mais fácil assumir essa hipótese de exogeneidade das variáveis explicativas.

Exemplo 2: voltando à estimação do efeito dos anos de estudos no rendimento do trabalho principal:

$$Lrendpri_i = \beta_0 + \beta_1 \text{anest}_i + \beta_2 exp_i + \varepsilon_i \quad (50)$$

Hipótese-chave:

$$\mathbb{E}[\varepsilon_i | \text{anest}_i, exp_i] = 0 \quad (51)$$

Nesse caso, a hipótese acima indica que outros fatores que afetam o salário e estão, portanto, incluídos nos erros, não são correlacionados na média com anos de estudo e experiência. Nesse caso, essa pode ser uma hipótese forte, se pensarmos que fatores não observáveis que afetam o rendimento podem também afetar os anos de estudo.

Comparativamente ao exemplo 1, a hipótese (51) parece ser mais forte. Na verdade, esse é um problema enfrentado por economistas do trabalho ao tentar estimar o retorno à educação. Para entender um pouco mais esse problema, pense que a habilidade do indivíduo é uma característica inata que afeta o seu rendimento no trabalho principal, uma vez que define a sua produtividade. Essa habilidade é algo não observável pelo econometrista, portanto, não pode ser incluída no modelo com uma variável explicativa adicional, logo estará incluída no erro do modelo. Se a habilidade também afetar a decisão por estudar, teremos a violação da hipótese (51).

Podemos, de formar mais genérica, apresentar o modelo de regressão linear múltipla para *k* variáveis explicativas como:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad (52)$$

Nesse caso, temos $k + 1$ parâmetros, com:

β_0 : intercepto

$\beta_1, \ldots, \beta_k$: são "inclinações", embora na prática elas não sejam a inclinação da função.

ε_i: termo de erro

A hipótese-chave é definida como:

$$\mathbb{E}[\varepsilon_i | X_{1i}, X_{2i}, \ldots, X_{ki}] = 0 \quad (53)$$

Como interpretamos a hipótese (53)? Tal hipótese estabelece que nenhum fator no termo de erro pode ser correlacionado com qualquer variável explicativa.

Já afirmamos, anteriormente, que o estimador utilizado para o modelo de regressão linear múltipla será obtido também pelo método de mínimos quadrados ordinários (MQO).

Desse forma, para o caso de duas variáveis explicativas, a equação estimada por MQO é dada por:

$$Y_i = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}}_{\hat{Y}_i} + \hat{\varepsilon}_i \quad (54)$$

em que $\hat{\varepsilon}_i$ é o resíduo da regressão.

O método de MQO escolhe os valores para os parâmetros desconhecidos que minimizam a soma dos quadrados dos resíduos da regressão. Com *N* observações de Y, X_1 e X_2, $\hat{\beta}_0$, $\hat{\beta}_1$ e $\hat{\beta}_2$ são escolhidos simultaneamente, de modo a fazer com que o valor da expressão abaixo seja o menor possível:

$$\min \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \sum_{i=1}^{N} \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i}\right)^2 \quad (55)$$

A resolução matemática de (55) mostra que os estimadores de MQO satisfazem as equações normais:

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}_1 + \hat{\beta}_2 \overline{X}_2 \quad (56)$$

$$\sum_{i=1}^{N} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{N} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{N} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{N} X_{1i} X_{2i} \quad (57)$$

$$\sum_{i=1}^{N} Y_i X_{2i} = \hat{\beta}_0 \sum_{i=1}^{N} X_{2i} + \hat{\beta}_1 \sum_{i=1}^{N} X_{1i} X_{2i} + \hat{\beta}_2 \sum_{i=1}^{N} X_{2i}^2 \quad (58)$$

Com intercepto:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2 \quad (59)$$

E inclinações:

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^{N} Y_i X_{1i}\right)\left(\sum_{i=1}^{N} X_{2i}^2\right) - \left(\sum_{i=1}^{N} Y_i X_{2i}\right)\left(\sum_{i=1}^{N} X_{1i} X_{2i}\right)}{\left(\sum_{i=1}^{N} X_{2i}^2\right)\left(\sum_{i=1}^{N} X_{1i}^2\right) - \left(\sum_{i=1}^{N} X_{1i} X_{2i}\right)^2} \quad (60)$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^{N} Y_i X_{2i}\right)\left(\sum_{i=1}^{N} X_{1i}^2\right) - \left(\sum_{i=1}^{N} Y_i X_{1i}\right)\left(\sum_{i=1}^{N} X_{1i} X_{2i}\right)}{\left(\sum_{i=1}^{N} X_{2i}^2\right)\left(\sum_{i=1}^{N} X_{1i}^2\right) - \left(\sum_{i=1}^{N} X_{1i} X_{2i}\right)^2} \quad (61)$$

Na interpretação do modelo de regressão linear múltipla, como em (62):

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \quad (62)$$

tem-se que:

$\hat{\beta}_0$: o valor predito de Y quando $X_1 = 0$ e $X_2 = 0$

$\hat{\beta}_1$ e $\hat{\beta}_2$: efeitos parciais

$$\Delta \hat{Y}_i = \hat{\beta}_1 \Delta X_{1i} + \hat{\beta}_2 \Delta X_{2i} \quad (63)$$

Se X_2 é mantido fixo em (63), $\Delta X_{2i} = 0$

$$\Delta \hat{Y}_i = \hat{\beta}_1 \Delta X_{1i} \quad (64)$$

Por outro lado, se X_1 é mantido fixo, $\Delta X_{1i} = 0$

$$\Delta\hat{Y}_i = \hat{\beta}_2 \Delta X_{2i} \text{ (65)}$$

3.1 Propriedades numéricas do estimador de MQO

A partir do modelo de regressão linear múltipla (66) e dos resíduos estimados (67), podemos elencar as propriedades númericas do estimador de MQO.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki} \text{ (66)}$$

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i \text{ (67)}$$

1. A soma dos resíduos é igual a 0: $\sum_{i=1}^{N} \hat{\varepsilon}_i = 0$

2. $\overline{y} = \overline{\hat{y}}$

3. A covariância amostral entre cada variável independente e os resíduos de MQO é zero.

$$\sum_{i=1}^{N} \hat{\varepsilon}_i X_{ij} = 0, j = 1, \ldots, k$$

A covariância amostral entre cada variável independente e o valor predito de MQO é zero.

$$\sum_{i=1}^{N} \hat{y}_i X_{ij} = 0, j = 1, \ldots, k$$

O ponto $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \overline{y})$ sempre está na reta de regressão de MQO:

$$\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}_1 + \hat{\beta}_2 \overline{x}_2 + \ldots + \hat{\beta}_k \overline{x}_k$$

Outro ponto relevante a considerar refere-se à comparação entre o modelo de regressão simples de Y em X_1 e o modelo de regressão múltipla de Y em X_1 e X_2. Para entender quando os dois modelos produzirão os mesmos resultados, é preciso estabelecer os valores de *Y* estimados nas duas situações.

No modelo de regressão simples temos:

$$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i}$$

Já no modelo de regressão múltipla:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

Temos então que:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\delta}$$

onde $\hat{\delta}$ mede o efeito de X_{1i} em X_{2i}. É o coeficiente da regressão linear de X_{i2} em X_{i1}

Logo, se $\hat{\beta}_2 = 0$, $\tilde{\beta}_1 = \hat{\beta}_1$ De outro modo, se X_1 e X_2 são não correlacionados na amostra, $\hat{\delta} = 0 \Rightarrow \tilde{\beta}_1 = \hat{\beta}_1$

3.2 Medidas de ajustamento no modelo de regressão linear múltipla

Vamos retomar, agora, a definição do coeficiente de ajustamento do modelo de regressão linear, considerando agora o modelo de regressão linear múltipla. O conceito aqui é exatamente o mesmo já apresentando no contexto de regressão linear simples. Apenas para recordar, temos:

Soma total dos quadrados:

$$STQ = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$$

Soma dos quadrados explicados pela regressão:

$$SQE = \sum_{i=1}^{N} \left(\hat{y}_i - \overline{\hat{y}}\right)^2 = \sum_{i=1}^{N} (\hat{y}_i - \overline{y})^2$$

Soma do quadrado dos resíduos

$$SQR = \sum_{i=1}^{N} \hat{\varepsilon}_i^2$$

A variação total em y é a soma da variação que foi explicada pela regressão com a variação que não foi explicada:

$$STQ = SQE + SQR$$

Daí, temos o coeficiente de ajustamento

$$R^2 = \frac{SQE}{STQ}$$

Podemos, então, definir o R^2 como o quadrado do coeficiente de correlação entre o valor atual de Y_i e o valor predito $\hat{Y}_i$

$$R^2 = \frac{\left(\sum_{i=1}^{N}(y_i-\overline{y})\left(\hat{y}_i-\overline{\hat{y}}\right)\right)^2}{\left(\sum_{i=1}^{N}(y_i-\overline{y})^2\right)\cdot\left(\sum_{i=1}^{N}\left(\hat{y}_i-\overline{\hat{y}}\right)\right)^2} \quad (68)$$

Verificamos que o R^2 nunca decresce, e, em geral, cresce quando adicionamos uma outra variável explicativa na regressão. Isso porque a soma do quadrado dos resíduos nunca aumenta quando adicionamos uma nova variável explicativa na regressão. Torna-se, portanto, um pouco mais difícil decidir sobre a inclusão ou não de uma variável adicional no modelo. Na verdade, desejaríamos saber se essa variável tem um efeito parcial em *Y.*

Muitas vezes, é interessante utilizar o conceito do R-quadrado ajustado, que irá considerar o número de variáveis explicativas no modelo, ou seja:

$$\overline{R}^2 = 1 - \frac{\frac{SQR}{N-k-1}}{\frac{STQ}{N-1}} \quad (69)$$

$$= 1 - \frac{\hat{\sigma}^2}{\frac{STQ}{N-1}}$$

Note que, na equação (69), ponderamos a variação explicada pelo graus de liberdade, ou seja, há um custo, de certo modelo, associado à inclusão de uma variável explicativa. Logo, para $k > 2$, $\overline{R}^2 < R^2$

$$\overline{R}^2 = 1 - (1 - R^2)\cdot\frac{N-1}{N-k-1}$$

Além disso, o $\overline{R}^2$ pode ser negativo.

3.3 Estimador de MQO e não-viés

Vamos aqui, assim como no caso do modelo de regressão linear simples, apresentar as hipóteses do estimador de MQO, em modelos de regressão linear múltipla.

A primeira hipótese a ser considerada diz respeito à forma funcional do modelo de regressão linear múltipla, ou seja:

HIPÓTESE 1 (H1): o modelo é linear nos parâmetros:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad (70)$$

De forma complementar, a hipótese 2 estabelece a aleatoriedade da amostra:

HIPÓTESE 2 (H2): temos uma amostra aleatória de N observações

$$\{(X_{1i}, \ldots, X_{ki}, Y_i), i\ 1, \ldots, N\}$$

que segue o modelo populacional em (70).

De outro lado, a hipótese 3 refere-se ao comportamento das variáveis explicativas, isto é:

HIPÓTESE 3 (H3): na amostra e, consequentemente, na população, nenhuma variável explicativa é constante. Além disso, não há uma relação linear exata entre quaisquer duas variáveis explicativas, o que exclui a possibilidade de colinearidade perfeita entre variáveis explicativas.

Para entendermos esse hipótese, vamos pensar no modelo do exemplo 1, efeito de variáveis climáticas sobre o PIB da agropecuária, de uma forma revista, em que:

$$lpib_{agro} = \beta_0 + \beta_1 temp_{verão\ i} + \beta_1 prec_{verão\ i} + \beta_2 prec_{primavera\ i} + \beta_3 prec_{outono\ i} + \beta_{14} prec_{inverno\ i+} + \beta_1 prec_{ano\ i} + \varepsilon_i$$

em

$$prec_{ano\ i} = prec_{verão\ i} + prec_{primavera\ i} + prec_{outono\ i} + prec_{inverno\ i}$$

Nesse caso, teríamos que a variável $prec_{ano\ i}$, precipitação total no ano, seria a soma das demais variáveis que medem a precipitação

em cada uma das estações, ou seja, haveria uma colinearidade perfeita entre essas variáveis, e, portanto, a violação de **H3**.

Cabe destacar que existem casos em que, embora não haja colinearidade perfeita entre as variáveis, existe uma correlação muito forte entre as mesmas, gerando um problema de multicolinearidade, em que se torna difícil olhar para os parâmetros estimados de forma independente.

Uma solução simples para o caso de colinearidade perfeita ou multicolinearidade é retirar uma das variáveis.

A hipótese **H3** pode ser violada, também, no caso em que o tamanho da amostra é muito pequeno em relação ao número de parâmetros a serem estimados. Para estimar $k + 1$ parâmetros, precisamos de pelo menos $k + 1$ observações.

A próxima hipótese estabelece o valor para a média do erro condicionada às variáveis explicativas, isto é:

HIPÓTESE 4 (H4): o valor esperado do erro, dado qualquer valor das variáveis independentes, é zero.

$$\mathbb{E}[\varepsilon|X_1, \ldots, X_k] = 0$$

Existem duas formas de violar essa hipótese:

(a) Forma funcional errada

(b) Omitir um fator importante que é correlacionado com $X_1, \ldots, X_k$

A hipótese **H4** também é conhecida com hipótese de **exogeneidade** das variáveis explicativas ou independentes. Assim, se uma das variáveis explicativas X_j é correlacionada com ε, dizemos que a variável explicativa é **endógena**.

Sob as hipóteses **H1 – H4**,

$\mathbb{E}\left[\hat{\beta}_j \middle| X_1, \ldots, X_k\right] = \beta_j, j = 0, \ldots, k$ para qualquer valor dos parâmetros populacionais.

Podemos, desse modo, afirmar que o estimador de MQO é um estimador não viesado se o processo fosse replicado para *n* amostras aleatórias possíveis. Em outras palavras, ao estimarmos um modelo pelo método de MQO, esperamos obter, a partir de uma amostra aleatória,

uma estimativa perto do valor populacional (desconhecido), mas não é possível ter certeza sobre isso. O que podemos afirmar é que sob as hipóteses **H1 – H4**, não há porque acreditar que a estimativa encontrada seja muito pequena ou muito grande em relação ao verdadeiro parâmetro populacional.

Logo, uma questão relevante seria: o que acontece se incluirmos variáveis irrelevantes no modelo? É importante esclarecer que a inclusão de variáveis explicativas que não apresentam efeito parcial na variável depedente *Y* não causará viés no estimador de MQO se as hipóteses **H1 – H4** forem válidas.

Suponha um modelo populacional:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

que satisfaz as hipóteses **H1 – H4**

Agora, considere que X_3 não tem efeito parcial em Y, quando controlamos por X_1 e $X_2 \Rightarrow \beta_3 = 0$, isto é,

$$\mathbb{E}[Y_i | X_{1i}, X_{2i}, X_{3i}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

Como não sabemos que $\beta_3 = 0$, incluimos na nossa regressão:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{\beta}_3 X_{3i}$$

A inclusão de X_3 não gera viés nos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$ sempre que **H1 – H4** forem verdadeiras.

Por outro lado, se, ao invés de considerarmos a inclusão de uma variável irrelevante, tivermos a omissão de uma variável relevante, será que isso gerará viés? Sim, nesse caso, teremos viés nos estimadores de MQO.

Considere, novamente, o seguinte modelo populacional:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

que satisfaz hipóteses **H1 – H4**

Suponha que estamos interessados em β_1, mas omitimos a variável. De modo que estimamos o seguinte modelo:

$$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i}$$

Vamos, portanto, retornar ao exemplo do Quadro 2. Queremos estimar o efeito dos anos de estudo sobre o rendimento do trabalho principal, a partir do seguinte modelo populacional:

$$Lrendpri_i = \beta_0 + \beta_1 \text{anest}_i + \beta_3 habilidade_i + \varepsilon_i$$

Como não observamos habilidade, estimamos o seguinte modelo:

$$Lrendpri_i = \beta_0 + \beta_1 \text{anest}_i + \upsilon_i$$

onde $\upsilon_i = \beta_3$ habilidade $_i + \varepsilon_i$, gerando, dessa forma, um viés de variável omitida. Mas qual será a direção desse viés?

Sabemos que

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \cdot \widehat{\delta}$$

onde $\hat{\beta}_1$, $\hat{\beta}_2$ são os estimadores da regressão de Y_i em X_{1i} ,X_{2i} . $\widehat{\delta}$ é a inclinação da regressão linear simples de X_{2i} em X_{1i}

Nesse caso,

$$\mathbb{E}\left[\tilde{\beta}_1 \middle| X_1, X_2\right] = \beta_1 + \beta_2 \cdot \widehat{\delta}$$

e o viés

$$\beta_2 \cdot \widehat{\delta}$$

Portanto, o estimador será não viesado se:

(a) X_1 e X_2 são não correlacionados $\Rightarrow \hat{\delta} = 0$.

(b) $\beta_2 = 0$, X_2 não aparece no modelo populacional.

No caso de viés do estimador, teremos as seguintes possibilidades (Quadro 5):

Quadro 5 – Direção do viés de variável omitida

	Corr$(X_1, X_2) > 0$	Corr$(X_1, X_2) < 0$
$\beta_2 > 0$	viés positivo	viés negativo
$\beta_2 < 0$	viés negativo	viés positivo

Fonte: elaboração própria.

3.3.1 O caso geral: duas ou mais variáveis explicativas

Entender a direção do viés de variável omitida no caso com múltiplas variáveis explicativas é bem mais complicado. Isso porque a correlação de uma única variável explicativa com o erro gera viés em todos os estimadores de MQO

Vamos, portanto, analisar o caso de três variáveis explicativas. Seja o seguinte modelo populacional:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

satisfazendo as hipóteses **H1 – H3**

Porém, para estimar o modelo, omitimos X_3

$$Y_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i} + \tilde{\beta}_2 X_{2i}$$

Sabemos que X_2 e X_3 são não correlacionados, mas X_1 e X_3 são correlacionados. Logo, $\tilde{\beta}_1$ e $\tilde{\beta}_2$ são viesados. Determinar a direção do viés é extremamente complicado, a não ser que assumamos que X_1 e X_2 são não correlacionados.

Assim, se X_1 e X_2 são não correlacionados, $\tilde{\beta}_2$ é não viesado, ou seja:

$$E\left[\tilde{\beta}_1\right] = \beta_1 + \beta_3 \cdot \frac{\sum_{i=1}^{N} \left(X_{1i} - \overline{X_1}\right) X_{3i}}{\sum_{i=1}^{N} \left(X_{1i} - \overline{X_1}\right)^2}$$

Para ilustrar, vamos retornar ao exemplo 2 revisto, conforme apresentado no Quadro 4. Assim:

$$Lrendpri_i = \beta_0 + \beta_1 \text{anest}_i + \beta_2 exp_i + \beta_3 habilidade_i \varepsilon_i$$

Qual seria a direção do viés em $\tilde{\beta}_1$ devido à omissão da variável habilidade no modelo estimado? Se considerarmos que habilidade e anos de estudo são positivamente correlacionados, a omissão da variável não observada habilidade do modelo fará com que haja uma sobreestimação do retorno à educação. Dito de outro modo, é como se o coeficiente dos anos de estudo captasse parte do efeito da habilidade sobre o salário e não apenas do aumento dos anos de estudo.

3.4 A variância do estimador de MQO

Para encontrar as variâncias dos estimadores de MQO, adicionamos a hipótese de homocedasticidade, já conhecida no modelo de regressão linear simples.

HIPÓTESE 5 (H5): o erro tem a mesma variância dado qualquer valor das variáveis explicativas.

$$\mathbb{V}ar[\varepsilon_i|X_{1i}.X_{2i},\ldots,X_{ki}] = \sigma^2 \text{ (71)}$$

De acordo com **H5**, a variância do erro é a mesma para qulaquer combinação de variáveis explicativas.

Com as hipóteses **H1 – H5**, podemos derivar a esperança condicional e a variância condicional de Y:

$$\mathbb{E}[Y|X_{1i}.X_{2i},\ldots,X_{ki}] = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} \text{ (72)}$$

$$\mathbb{V}ar[Y|X_{1i}.X_{2i},\ldots,X_{ki}] = \sigma^2 \text{ (73)}$$

Podemos, também, obter a variância condiconal do estimador de MQO

$$Var\left[\hat{\beta}_j\right] = \frac{\sigma^2}{\left(\sum_{i=1}^{N}\left(X_{ji}-\overline{X}_j\right)^2\right)\left(1-R_j^2\right)}, j = 0,..,k \text{ (74)}$$

onde R_j^2 é o R-quadrado da regressão de X_j nas outras variáveis explicativas.

Por que buscamos um estimador de variância mínima? Porque quanto maior a variância, menor a precisão da estimativa, maiores os intervalos de confiança e menos exatos serão os testes de hipóteses.

A variância do estimador de MQO, conforme estabelecida em (74), possui os seguintes componentes:

1. **A variância do erro** (σ^2): quanto maior a variância do erro, menor a precisão do estimador. Uma das formas de reduzir a variância do erro é acrescentar outras variáveis explicativas ao modelo.

2. **O total da variância em X_j (SST)**: quanto maior a variância em X_j, menor a variância do estimador. O aumento da variabilidade de *X* pode ser alcançado aumentado-se o tamanho da amostra.

3. **A relação linear entre as variáveis explicativas** (R_j^2): um R_j^2 muito alto significa que as demais variáveis independentes explicam muito da variabilidade em X_j. Quando R_j^2 aumenta, a variabilidade de $\hat{\beta}_j$ também aumenta. A melhor situação ocorre quando $R_j^2 = 0$, e a variância explode quando $R_j^2 = 1$. Ressalte-se que os casos em que $R_j^2 = 1$ estão associados a problemas de multicolinearidade, ou seja, de correlação muito alta entre as variáveis explicativas, e, portanto, violação de **H3**. Note, porém, que R_j^2 perto de **1** não viola a hipótese **H3**.

3.4.1 Variância do estimador em modelos especificados de forma incorreta

Vamos discutir o que acontece com a variância se incluirmos uma variável irrelevante ou excluirmos uma variável relevante. Para tanto, vamos considerar o seguinte modelo que satisfaz as hipóteses **H1 – H5**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Vamos considerar dois estimadores:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$$

$$\tilde{Y}_i = \tilde{\beta}_0 + \tilde{\beta}_1 X_{1i}$$

Nesse caso, $\tilde{\beta}_1$ é viesado. Como comparamos as variâncias?

$$\mathbb{V}ar\left[\hat{\beta}_1\right] = \frac{\sigma^2}{\left(\sum_{i=1}^{N}\left(X_{ji} - \overline{X}_j\right)^2\right)(1 - R_1^2)}$$

$$\mathbb{V}ar\left[\tilde{\beta}_1\right] = \frac{\sigma^2}{\left(\sum_{i=1}^{N}\left(X_{ji} - \overline{X}_j\right)^2\right)}$$

Logo,

$$\mathbb{V}ar\left[\tilde{\beta}_1\right] < \mathbb{V}ar\left[\hat{\beta}_1\right]$$

a não ser que X_1 seja não correlacionado com X_2. O Quadro 6 a seguir resume os efeitos da omissão de variável sobre a variância dos estimadores.

Quadro 6 – Variância e viés e variável omitida

	$\text{Corr}(X_1, X_2) \neq 0$	
	$\beta_2 \neq 0$	$\beta_2 = 0$
$\hat{\beta}_1$	não viesado, > variância	não viesado, > variância
$\tilde{\beta}_1$	viesado, < variância	não viesado, < variância

Fonte: elaboração própria.

Note que se $\text{Corr}(X_1, X_2) = 0$, $\hat{\beta}_1 = \tilde{\beta}_1$. Além disso, a inclusão de uma variável irrelevante no modelo, embora não cause viés no estimador, aumenta a variância. Já a omissão de uma variável relevante gera viés, mas reduz a variância.

3.5 Variância dos erros no modelo de regressão múltipla

Assim como no caso de regressão simples, usamos a soma do quadrado dos resíduos ponderada pelos graus de liberdade.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{N} \hat{\varepsilon}_i^2}{N-k-1} \quad (75)$$

Este estimador é não viesado:

$$\mathbb{E}[\hat{\sigma}^2 | X_1, .., X_k] = \sigma^2$$

Estimamos o erro padrão de $\hat{\beta}_j$ como:

$$\hat{\sigma}_{\beta_j} = \frac{\hat{\sigma}}{\sqrt{\left(\sum_{i=1}^{N} (X_{ji} - \overline{X}_j)^2\right)(1-R_1^2)}} \quad (76)$$

Se as hipóteses **H1 – H5** são válidas, o estimadores de MQO são os estimadores lineares não viesados, mais eficientes, ou seja, possuem variância mínima.

3.6 Inferência nos modelos de regressão linear múltipla

Para encontrarmos a distribuição dos estimadores de MQO, temos que impor a hipótese de normalidade.

HIPÓTESE 6 (H6): condicional a todas as variáveis explicativas, o erro tem uma distribuição normal com média 0 e variância σ^2

$$\varepsilon|X_1, X_2, \ldots, X_k \sim \mathcal{N}(0, \sigma^2) \quad (77)$$

Sob as hipóteses **H1 – H6**:

$$Y|X_1, X_2, \ldots, X_k \sim \mathcal{N}(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k, \sigma^2) \quad (78)$$

Sob as hipóteses **H1 – H6**:

$$\hat{\beta}_j \left| X_1, X_2, \ldots, X_k \sim \mathcal{N}\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right) \right. \quad (79)$$

onde

$$\sigma^2_{\hat{\beta}_j} = \frac{\sigma^2}{\left(\sum_{i=1}^{N} \left(X_{ji} - \overline{X}_j\right)^2\right)\left(1 - R_j^2\right)} \quad (80)$$

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\beta_j}} \sim \mathcal{N}(0,1)$$

Podemos afirmar que qualquer combinação linear de $\hat{\beta}_1, .., \hat{\beta}_k$ tem uma distribuição normal.

Assim, podemos aplicar o teste bilateral para verificar se os efeitos estimados por MQO são significativos ou não:

Hipótese nula:

$$H_0: \beta_j = 0$$

Estatística t:

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}} = \frac{\hat{\beta}_j}{\sigma_{\beta_j}}$$

Rejeita a hipótese nula se $|t| > t_{\frac{\alpha}{2}}$

É importante lembar que os testes só são válidos sob as hipóteses de **exogeneidade** e **homocedasticidade**.

Por fim, muitas vezes queremos testar se os coeficientes são conjuntamente iguais a zero. Nesse caso, utilizamos o teste de F.

Consideremos o seguinte modelo não restrito:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Vamos supor que queremos testar a hipótese de que as últimas q variáveis do modelo têm coeficientes iguais a zero.

$H_0: \beta_{k-q+1} = 0, \ldots, \beta_k = 0$

A hipótese alternativa é que pelo menos um dos parâmetros é diferente de zero.

Consideremos, então, o modelo restrito:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{k-q} X_{k-qi} + \varepsilon_i$

O que precisamos é de uma estatística que compare a SQR nos dois modelos. Vamos, olhar, portanto, para a estatística F:

$$F = \frac{\frac{(SQR_r - SQR_{nr})}{q}}{\frac{SQR_{nr}}{N-k-1}}$$

Note que a estatística F é sempre não-negativa (≥ 0). Lembre que a soma dos quadrados dos resíduos nunca aumenta quando aumentamos o número de variáveis explicativas.

$F \sim \mathcal{F}_{q,N-k-1}$

Rejeitamos H_0 se $\mathrm{F} > c$, onde $c = F_{1-\alpha}$. Se H_0 é rejeitada, dizemos que $X_{k-q+1}, \ldots, X_k$ são estatisticamente significativas conjuntamente a 5% (ou 10%, 1%) de significância.

Retornando ao exemplo do Quadro 1, que testa o efeito da temperatura no verão sobre o PIB na agropecuária, vamos agora estimar um modelo que inclua, além da variável temperatura no verão, as temperaturas no inverno, outorno e primavera, bem como as precipitações nas quatro estações do ano.

Modelo 1:

$lpib_{agro} = \beta_0 + \beta_1 temp_{verão\ i} + \varepsilon_i$

Quadro 7 – Estimação do modelo restrito para efeito da temperatura sobre o PIB agropecuário

Source	SS	df	MS	Number of obs	=	4,967
				F(1, 4965)	=	104.39
Model	125.282521	1	125.282521	Prob > F	=	0.0000
Residual	5958.42312	4,965	1.20008522	R-squared	=	0.0206
				Adj R-squared	=	0.0204
Total	6083.70564	4,966	1.22507161	Root MSE	=	1.0955

lagro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tempver	-.0815801	.0079845	-10.22	0.000	-.0972332	-.0659271
_cons	11.28602	.197066	57.27	0.000	10.89969	11.67236

Fonte: elaboração própria.

Modelo 2:

$$ib_{agro} = \beta_0 + \beta_1 temp_{verão\,i} + \beta_2 temp_{inverno\,i} + \beta_3 temp_{primavera\,i} + \beta_4 temp_{outono\,i} + \beta_5 prec_{verão\,i} + \beta_6 prec_{primavera\,i} + \beta_7 prec_{outono} + \beta_8 prec_{inverno\,i+} + \beta_9 prec_{ano\,i}\varepsilon_i$$

Quadro 8 – Estimação do modelo não restrito para efeito da temperatura e precipitação sobre o PIB agropecuário

Source	SS	df	MS	Number of obs	=	4,967
				F(8, 4958)	=	98.51
Model	834.349133	8	104.293642	Prob > F	=	0.0000
Residual	5249.35651	4,958	1.05876493	R-squared	=	0.1371
				Adj R-squared	=	0.1358
Total	6083.70564	4,966	1.22507161	Root MSE	=	1.029

lagro	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
tempver	-.1432653	.0410509	-3.49	0.000	-.2237433	-.0627873
tempout	-.2059371	.0481377	-4.28	0.000	-.3003083	-.111566
temppri	.2708589	.0402192	6.73	0.000	.1920115	.3497063
tempinv	.0956549	.0535966	1.78	0.074	-.0094181	.200728
precinv	.0019529	.0007051	2.77	0.006	.0005705	.0033352
precout	.000568	.0006834	0.83	0.406	-.0007717	.0019077
precpri	.0081451	.0007219	11.28	0.000	.0067298	.0095605
precver	-.000299	.0006217	-0.48	0.631	-.0015177	.0009198
_cons	7.465086	.3247802	22.99	0.000	6.828373	8.101799

Fonte: elaboração própria.

A partir das estimações dos modelos 1 e 2, vamos testar se as variáveis *tempout, tempinv, temppri, precinv, precout, precpri, precver* são conjuntamente diferentes de zero.

$$F = \frac{\frac{(\ 5958{,}42312 - 5249{,}35651)}{7}}{\frac{5249{,}35651}{4958}} = 95{,}67$$

O teste F, portanto, rejeita a hipótese nula de que *tempout, tempinv, temppri, precinv, precout, precpri, precver* são conjuntamente iguais a zero. Se olharmos na tabela, o valor de F com significância de 5% é igual a 2,60, como 95,67 > 2,60, podemos rejeitar a hipótese nula.

Cabe destacar que podemos reescrever a estatística F a partir do valor de R-quadrado, isso porque sabemos que existe uma relação entre a SQR e o R^2

$$SQR = SQT(1 - R^2)$$

Então, escrevendo a estatística F usando o R^2, temos:

$$F = \frac{\frac{R^2_{nr} - R^2_r}{q}}{\frac{1 - R^2_{nr}}{N-k-1}} \quad (81)$$

No nosso exemplo, teríamos

$$F = \frac{\frac{(0{,}1371 - 0{,}0206)}{7}}{\frac{1-0{,}1371}{4958}} = 95{,}62$$

Se olharmos o resultado do teste F que o *software* STATA© realiza, como reportado nos Quadros 8 (F=104,39) e 9 (F=98,51), temos que ele testa a significância total da regressão.

H_0: $X_1, X_2, \ldots, X_k$ não ajudam a explicar Y

$H_0: \beta_1 = 0, \ldots, \beta_k = 0$

O modelo restrito é

$Y_i = \beta_0 + \varepsilon_i$

Neste caso, a estatística de F é

$$F = \frac{\frac{R^2}{q}}{\frac{1 - R^2}{N-k-1}} = \frac{\frac{SQE}{q}}{\frac{SQR}{N-k-1}}$$

Testa-se a significância de todas as variáveis explicativas conjuntamente.

3.7 Teoria assintótica e estimadores de MQO

É importante destacarmos que todos os resultados até agora eram válidos para um *N* com valor finito. Portanto, as propriedades vistas até o momento são conhecidas como propriedades em pequenas amostras. Mas existem propriedades que são válidas apenas em grandes amostras, ou seja, aquelas que são definidas quando $N \to \infty$.

Qual a importância das propriedades assintóticas? Note que, por exemplo, o não-viés é uma propriedade que nem sempre é satisfeita por um estimador. Por outro lado, os estimadores devem ser, no mínimo, consistentes.

Um estimador será **consistente** se, quando N tender ao infinito, a distribuição de $\hat{\beta}_j$ convergir para β_j. Se as hipóteses **H1 – H4** são válidas, o estimador de MQO $\hat{\beta}_j$ é consistente para β_j, para todo $j = 1, .., k$. Por outro lado, o estimador de MQO deixa de ser consistente se houver correlação entre ε e qualquer variável explicativa $(X_1, \ldots, X_k)$.

Além da violação da propriedade de não-viés do estimador de MQO, pode ocorrer também a violação da propriedade de normalidade dos estimadores. No caso em que os erros são variáveis aleatórias oriundas de uma outra distribuição que não a normal, $\hat{\beta}_j$ não terá uma distribuição normal, e, por consequência, os teste t e F não serão válidos. E não são raras as vezes em que a hipótese de normalidade é violada.

Daí, a importância da definição de **normalidade assintótica**. Segundo a propriedade de normalidade assintótica, quando $N \to \infty$, a distribuição do estimador de MQO pode ser aproximada por uma distribuição normal. Se as hipóteses **H1 – H5** são válidas:

$$\frac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{\beta}_j}} \sim \mathcal{N}(0,1) \quad (81)$$

Ressalte-se que a qualidade da aproximação, em (81), depende do tamanho da amostra, N, e do número de graus de liberdade.

3.8 Formas funcionais dos modelos de regressão múltipla

Na análise dos modelos de regressão, tão importante quanto realizar inferências, verificando a significância dos estimadores, é a interpretação das estimativas. Vamos começar pela forma logarítimica, já apresentada nos exemplos dos Quadros 1 e 2.

$$\widehat{\log(Y)}_i = \hat{\beta}_0 + \hat{\beta}_1 \log(X_1)_i + \hat{\beta}_2 X_{2i}$$

Como devemos interpretar $\hat{\beta}_2$? Vimos anteriormente que uma aproximação pode ser obtida por $100.\hat{\beta}_2$, como sendo a variação percentual em *Y*, dado a variação de uma unidade em X_{2i} , mas algumas vezes tal aproximação se torna muito pouco precisa na aplicação. Nesses casos, é necessário calcular a porcentagem exata da mudança, ou seja:

Fixando X_1

$$\Delta\widehat{\log(Y)}_i = \hat{\beta}_2 \Delta X_{2i}$$

$$\Rightarrow \%\widehat{\Delta Y}_i = 100 \cdot \left[\exp\left(\hat{\beta}_2 \Delta X_{2i}\right) - 1\right]$$

A Tabela 1 fornece as fórmulas dos coeficientes angulares e de elasticidades para diferentes especificações do modelo de regressão linear.

Tabela 1 – Formas funcionais e interpretação no modelo de regressão linear

Modelo	Equação	Coeficiente angular $\left(=\frac{dY}{dX}\right)$	Interpretação de β_2	Elasticidade $\left(=\frac{dY}{dX}\frac{X}{Y}\right)$
Nível-nível	$Y = \beta_1 + \beta_2 X$	β_2	$\Delta Y = \beta_1 \Delta X$	$\beta_2 \left(\frac{Y}{X}\right)^*$
Log-log	$\ln Y = \beta_1 + \beta_2 lnX$	$\beta_2 \left(\frac{Y}{X}\right)$	$\%\Delta Y = \beta_1 \%\Delta X$	β_2
Log-nível	$\ln Y = \beta_1 + \beta_2 X$	$\beta_2 (Y)$	$\%\Delta Y = (100\beta_1)\Delta X$	$\beta_2 (X)^*$
Nível-log	$Y = \beta_1 + \beta_2 \ln X$	$\beta_2 \left(\frac{1}{X}\right)$	$\Delta Y = (\beta_1/100)\%\Delta X$	$\beta_2 \left(\frac{1}{X}\right)^*$

Nota: o * indica que a elasticidade varia do valor assumido por X ou Y ou ambos. Quando tais valores não são especificados, as elasticidades são medidas pelos valores médios de X e Y.

Fonte: Gujarati e Porter (2006) e Wooldridge (2011).

Além das formas funcionais que envolvem logarítimicos das variáveis, outra especificação bastante utilizada é a da forma quadrática. A especificação quadrática é utilizada para captar retornos marginais crescentes ou descrescentes.

Considere o modelo mais simples com uma variável explicativa

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{1i}^2$$

Nesse caso, a aproximação é dada por:

$$\Delta\hat{Y}_i \approx \hat{\beta}_1 + 2\hat{\beta}_2 X_{1i}$$

Para exemplificar, vamos retomar a estimação do efeito dos anos de estudo e experiência sobre o rendimento do trabalho principal.

O valor máximo de y é atingido quando

$$x^* = \left| \frac{\hat{\beta}_1}{2\hat{\beta}_2} \right|$$

Quadro 9 – Estimação com forma funcional quadrática do efeito dos anos de estudo e experiência sobre o rendimento do trabalho principal

Source	SS	df	MS	Number of obs	=	185,809
				F(3, 185805)	=	13860.95
Model	1.6122e+11	3	5.3740e+10	Prob > F	=	0.0000
Residual	7.2038e+11	185,805	3877070.54	R-squared	=	0.1829
				Adj R-squared	=	0.1829
Total	8.8160e+11	185,808	4744675.88	Root MSE	=	1969

rendpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anest	197.765	1.068876	185.02	0.000	195.67	199.86
exp	116.2874	2.406311	48.33	0.000	111.5711	121.0038
exp2	-.9317243	.0305698	-30.48	0.000	-.9916405	-.8718082
_cons	-3060.567	44.86068	-68.22	0.000	-3148.493	-2972.641

Fonte: elaboração própria.

Nesse exemplo, o efeito da variação de um ano de experiência sobre o rendimento do trabalho principal pode ser aproximado como 116,2874+2.(-0,9317243)=114,42395. Isso significa que 1 ano a mais de experiência está associado a um aumento de R$ 114, 42 reais no

rendimento do trabalho principal. Além disso, podemos perceber que o efeito da experiência sobre o rendimento do salário principal é positivo, mas a taxas decrescentes. Desse modo, o valor da experiência que maximiza o trabalho é dado por 116,2874/2.(-0,9317243), ou seja, o salário é maximizado quando o indivíduo alcança 62,40 anos de experiência.

Podemos, agora, analisar o comportamento do valor esperado de *y*, a partir da relação entre os coeficientes das variáveis em nível e quadrática.

Se os coeficentes referentes à variável em nível e na sua forma quadrática têm o mesmo sinal (ambos positivos ou ambos negativos), não existirá um valor no qual o efeito muda de sinal para X>0.

De outro lado, se $\beta_1 > 0$ e $\beta_2 > 0$, o menor valor esperado de Y acontece quando $X = 0$ e aumentos de X têm um efeito positivo em Y. Se $\beta_1 < 0$ e $\beta_2 < 0$, o maior valor esperado de Y acontece quando $X = 0$ e aumentos de X têm um efeito negativo em Y

3.9 Variável dummy em modelos de regressão linear

Uma variável dummy é uma variável binária que assume valores 0 ou 1. Por exemplo, sexo (0-mulher; 1-homem); raça (0-brancos, 1-não brancos); condição de ocupação (0-desocupados; 1-ocupados). A inclusão de uma variável dummy em um modelo em nível muda o intercepto do modelo.

Retornando ao exemplo do Quadro 9, vamos agora reestimar o modelo incluindo uma dummy para migrante:

Quadro 10 – Modelo determinantes do rendimento do trabalho principal, com *dummy* para migrante

Source	SS	df	MS	Number of obs	=	178,131
				F(4, 178126)	=	26612.90
Model	58408.701	4	14602.1752	Prob > F	=	0.0000
Residual	97735.5804	178,126	.548687897	R-squared	=	0.3741
				Adj R-squared	=	0.3741
Total	156144.281	178,130	.876574869	Root MSE	=	.74073

lnrendpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anest	.1224384	.0004113	297.66	0.000	.1216322	.1232446
exp	.0906406	.0009459	95.82	0.000	.0887867	.0924946
exp2	-.0008965	.000012	-74.82	0.000	-.00092	-.000873
migrante	.174536	.0035797	48.76	0.000	.1675199	.1815521
_cons	3.697893	.0176648	209.34	0.000	3.66327	3.732515

Fonte: elaboração própria.

A inclusão da *dummy* migrante está capturando uma diferença de intercepto ou do rendimento médio do trabalho principal entre migrantes e não migrantes. O coeficiente estimado, 0,174436, mostra que os migrantes recebem, em média, 19% a mais em relação aos não migrantes.

Algumas vezes, as variáveis *dummy* são utilizadas como forma de estimar efeitos não lineares. Podemos, por exemplo, em vez de estimar o retorno de cada ano de estudo, optar por estimar o efeito do diploma. Nesse caso, substituímos a variável anos de estudo por grupos de anos de estudo. O Quadro 11 apresenta os resultados para essa estratégia.

Quadro 11 – Modelo determinantes do rendimento do trabalho principal, com grupos de escolaridade

Source	SS	df	MS	Number of obs	=	175,478
				F(6, 175471)	=	17698.71
Model	57703.2216	6	9617.2036	Prob > F	=	0.0000
Residual	95348.2125	175,471	.543384448	R-squared	=	0.3770
				Adj R-squared	=	0.3770
Total	153051.434	175,477	.872202249	Root MSE	=	.73715

lnrendpri	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
anest	.1308801	.0013867	94.38	0.000	.1281621	.1335981
exp	.0884711	.0009473	93.39	0.000	.0866144	.0903279
exp2	-.0008806	.000012	-73.45	0.000	-.0009041	-.0008571
_Ig_anest_2	-.1545445	.0082038	-18.84	0.000	-.1706237	-.1384652
_Ig_anest_3	-.2464869	.0121736	-20.25	0.000	-.2703469	-.222627
_Ig_anest_4	.1148409	.0180438	6.36	0.000	.0794756	.1502063
_cons	3.877476	.0181707	213.39	0.000	3.841862	3.91309

Fonte: elaboração própria.

No exemplo do Quadro 11, foram considerados quatro grupos de escolaridade. A categoria omitida é a de 0 a 5 anos de estudos. O grupo 2 inclui indivíduos com 6 a 9 anos de estudo; o grupo 3 engloba aqueles com 9 a 12 anos de estudo; e, no grupo 4, estão indivíduos com 13 anos ou mais de estudo. É interessante notar que indivíduos com 6 a 9 anos de estudo ganham em média 14,32% a menos do que aqueles com até 5 anos de estudo.

Outra forma de inclusão de variáveis *dummy* nos modelos de regressão linear é por meio da interação com outras variáveis explicativas. Se, no caso da inclusão direta da dummy na regressão, o objetivo é estimar as diferenças de intercepto, no caso da inclusão da dummy interagindo com uma variável explicativa, o que se almeja é identificar diferenças na inclinação da reta de regressão para diferentes grupos ou categorias.

Capítulo 4 – Violação das hipóteses básicas do modelo de regressão linear

4.1 Violação da hipótese de homocedasticidade

Vamos considerar o seguinte modelo linear

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Qual a consequência da violação da hipótese de homocedasticidade? A existência de heterocedasticidade não causa viés nos estimadores, embora ocasione viés nos estimadores da vairância do MQO, tornando não válidos os testes *F* e *t*.

Precisamos, portanto, obter um estimador para a variância do estimador de MQO que seja válido na presença de heterocedasticidade de forma desconhecida.

Consideremos o modelo com uma variável explicativa:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

Vamos substituir **H5** por:

$$\mathbb{V}ar[\varepsilon_i | X_{1i}] = \sigma_i^2$$

Neste caso,

$$\mathbb{V}ar\left[\hat{\beta}_1 \middle| X_{1i}\right] = \frac{\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2 \cdot \sigma_i^2}{\left(\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2\right)^2}$$

Um estimador da variância robusto à heterocedasticidade é:

$$\widehat{\mathbb{V}ar}\left[\hat{\beta}_1 \middle| X_{1i}\right] = \frac{\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2 \cdot \hat{\varepsilon}_i^2}{\left(\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2\right)^2}$$

O desvio-padrão robusto a heterocedasticidade é:

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\sqrt{\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2 \cdot \hat{\varepsilon}_i^2}}{\sum_{i=1}^{N} \left(X_{1i} - \overline{X}\right)^2}$$

Note que o estimador da variância robusto a heterocedasticidade é consistente, mas é viesado.

Uma vez que sabemos dos efeitos gerados pela violação da hipótese de heterocedasticidade, torna-se importante discutir a forma de se identificar essa violação. Para tanto, podemos adotar um teste para heterocedasticidade. Consideremos o nosso modelo linear com *k* variáveis explicativas:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Vamos supor que as hipóteses **H1 – H4** são válidas.

Queremos testar se a hipótese **H5** é verdadeira:

$$H_0: \mathbb{V}ar[\varepsilon_i | X_{1i}, \ldots, X_{ki}] = \sigma^2$$

Se não rejeitamos essa hipótese a 5%, iremos concluir que heterocedasticidade não é um problema. Asumindo que $\mathbb{E}[\varepsilon_i | X_{1i}, \ldots, X_{ki}] = 0$, podemos reescrever a nossa hipótese como:

$$H_0: \mathbb{E}\left[\varepsilon_i^2 \middle| X_{1i}, \ldots, X_{ki}\right] = \sigma^2$$

Precisamos, desse modo, testar se ε_i^2 é relacionado a uma ou mais variáveis explicativas. Para tanto, assumimos que o erro é uma função linear das variáveis explicativas, isto é:

$$\varepsilon_i^2 = \delta_0 + \delta_1 X_{1i} + \ldots + \delta_k X_{ki} + \nu_i \quad (82)$$

Testamos a hipótese de homocedasticidade pela seguinte hipótese nula:

$$H_0: \delta_1 = \delta_2 = \ldots = \delta_k = 0$$

e aplicamos o teste de F.

Cabe lembrar que não podemos observar o erro, como em (82), mas sim os resíduos:

$$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 X_{1i} + \ldots + \delta_k X_{ki} + \nu_i$$

daí, basta utilizarmos o teste de F do STATA©

$$F = \frac{\frac{R^2_{\hat{\varepsilon}^2}}{k}}{\frac{1 - R^2_{\hat{\varepsilon}^2}}{N-k-1}}$$

Podemos sintetizar as etapas de realização desse teste de homocedasticidade da seguinte forma:

(a) Estimamos o modelo original por MQO, geramos a variável de resíduos ao quadrado e salvamos.

(b) Estimamos um modelo que relaciona a variável de resíduos ao quadrado com as variáveis explicativas.

(c) A partir do $R^2_{\hat{\varepsilon}^2}$, obtido na etapa (b), calculamos o teste de F.

Outra forma de testarmos a homocedasticidade se dá por meio do chamado **teste de White**. Neste teste, estimamos um modelo com os resíduos ao quadrado com variável depedente e os quadrados e produtos cruzados de *X* como variáveis explicativas, ou seja, em um modelo com três variáveis explicativas, teríamos:

$$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i} + \delta_4 X_{1i}^2 + \delta_5 X_{2i}^2 + \delta_6 X_{3i}^2 + \delta_7 X_{1i} X_{2i} + \delta_8 X_{1i} X_{3i} + \delta_9 X_{2i} X_{3i} + \nu_i$$

Uma vez estimado o modelo, fazemos um teste de F para a hipótese nula, $H_0: \delta_1 = \ldots = \delta_9 = 0$. A grande limitação do teste de White é que o número de restrições testadas cresce com o número de variáveis explicativas, aumentando a perda de graus de liberdade.

Alternativamente, podemos utilizar o valor predito pelo modelo em vez das variáveis explicativas. Ou melhor, considerando o modelo:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_k X_{ki}$$

podemos fazer um teste para heterocedasticidade usando o valor predito de *Y* para estimar os resíduos ao quadrado, isto é:

$$\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2 + \nu_i$$

Daí, testamos, utilizando o teste *F*, a seguinte hipótese nula:

$$H_0: \delta_1 = \delta_2 = 0$$

4.1.1 O método dos mínimos quadrados ponderados

Vimos, anteriormente, como testar a hipótese de homocedasticidade. Agora, vamos identificar uma forma para corrigir a heterocedasticidade.

Vamos substituir **H5** por:

$$\mathbb{V}ar[\varepsilon | X_1, X_2, \ldots, X_k] = \sigma^2 h(X_1, X_2, \ldots, X_k)$$

onde $h(X_1, X_2, \ldots, X_k)$ é uma função que determina a heterocedasticidade.

Agora, assumimos que $h(X_1, X_2, \ldots, X_k)$ é conhecida. Para determinada amostra de $(X_1, X_2, \ldots, X_k)$, podemos escrever

$$\sigma_i^2 = \mathbb{V}ar[\varepsilon_i | X_1, X_2, \ldots, X_k]$$
$$= \sigma^2 \cdot \underbrace{h(X_{1i}, X_{2i}, \ldots, X_{ki})}_{h_i}$$

É importante observar que $h(X_{1i}, X_{2i}, \ldots, X_{ki})$ muda para cada observação, pois as variáveis independentes mudam para cada observação.

Exemplo: Suponha que queremos estimar a relação entre poupança e renda

$$\text{poupanca}_i = \beta_0 + \beta_1 \text{renda}_i + \varepsilon_i$$

e assumimos que

$$\mathbb{V}ar[\varepsilon_i | \text{renda}_i] = \sigma^2 \cdot \text{renda}_i$$

Voltando ao modelo originial....

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Vamos fazer uma transformação no erro:

$$\varepsilon_i^* = \frac{\varepsilon_i}{\sqrt{h_i}}$$

Sob hipótese **H4**:

$$\mathbb{E}[\varepsilon_i^* | X_1, X_2, \ldots, X_k] = 0$$

e sob a forma de heterocedastcidade acima

$$\mathbb{E}\left[\varepsilon_i^{*2} \middle| X_1, X_2, \ldots, X_k\right] = \frac{\mathbb{E}[\varepsilon_i^2 | X_1, X_2, \ldots, X_k]}{h_i} = \frac{\sigma^2 \cdot h_i}{h_i} = \sigma^2$$

Para corrigirmos o problema de heterocedasticidade, basta dividirmos o nosso modelo original por $\sqrt{h_i}$:

$$\underbrace{\frac{Y_i}{\sqrt{h_i}}}_{Y_i^*} = \beta_0 + \beta_1 \underbrace{\frac{X_{1i}}{\sqrt{h_i}}}_{X_{1i}^*} + \ldots + \beta_k \underbrace{\frac{X_{ki}}{\sqrt{h_i}}}_{X_{ki}^*} + \underbrace{\frac{\varepsilon_i}{\sqrt{h_i}}}_{\varepsilon_i^*}$$

Daí, estimamos o modelo ponderado por MQO. Desse modo, na ausência de heterocedasticidade, podemos realizar os testes de *F* e *t*. Os coeficientes são interpretados na sua forma original.

É preciso estarmos atentos, contudo, que uma má especificação da variância faz com que os testes *F* e *t* percam a sua validade. Logo, para

que o método de MQO ponderado tenha sucesso em lidar com a violação da hipótese de homocedasticidade, é preciso estimar de forma correta h_i.

4.2 Violação da hipótese de endogeneidade

A violação da hipótese **H4**, em geral, está relacionada à utilização de forma funcional errada ou problema de variável omitida.

Desse modo, é fundamental escolhermos o modelo correto, senão os nossos estimadores serão viesados e inconsistentes. Quando erramos na forma funcional, estamos errando na especificação do modelo, ou melhor, temos todas as variáveis explicativas necessárias ao modelo, mas a forma como estas se relacionam com a variável dependente está incorreta.

Uma forma de testar a má espeficiação do modelo é pelo teste de significância conjunta, ou melhor, o teste F. Mas existem também testes específicos para a forma funcional, como, por exemplo, o teste de erro de especificação.

Note que se o modelo:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

satisfaz **H4**, então nenhuma função não linear das variáveis explicativas deveria ser signifcativa quando adicionada ao modelo. Podemos fazer o teste de *F*, mas nunca iremos detectar todas as não linearidades possíveis.

Contudo, o maior problema relacionado à endogeneidade é, exatamente, a omissão de uma variável explicativa relevante por não sermos capazes de observá-la. O exemplo clássico, nesse sentido, segue o apresentado no Quadro 2, na estimação da equação de rendimentos, em que não somos capazes de observar a habilidade dos indivíduos.

Uma das formas de solucionar a questão da omissão de variável explicativa relevante é a adoção de uma variável *proxy* para a mesma. Essa variável *proxy* seria alguma variável, que podemos observar, que seja fortemente correlacionada com a variável não observada.

Porém, para que a estratégia de adoção de uma *proxy* seja capaz de gerar estimadores de MQO consistentes, é preciso que ela satisfaça algumas hipóteses. Considerando o modelo com três variáveis explicativas:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

onde X_{3i}^* é uma variável *proxy* para X_{3i}, precisamos que X_{3i}^* seja realcionado com X_{3i}.

$$X_{3i} = \delta_0 + \delta_3 X_{3i}^* + \nu_{3i}$$

o δ_3 nos dá uma idéia da associação entre X_{3i} e sua variável *proxy*. O ideal é que δ_3 seja alto.

Assim, para que a utilização de X_{3i}^* possibilite alcançarmos um estimador de MQO consistente:

1. Precisamos que ε seja não correlacionado com X_1, X_2 e X_3

2. Precisamos que ε seja não correlacionado com X_3^*. Isto é, quando controlamos por X_1, X_2 e X_3, X_3^* se torna uma variável irrelevante.

3. O erro ν_3 tem que ser não correlacionado com X_1, X_2 e X_3^*. Esta hipótese assegura que X_3^* é uma boa proxy para X_3.

Para estimar o modelo utilizando a *proxy*, simplesmente substituimos a variável endógena no modelo, isto é:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3(\delta_0 + \delta_3 X_{3i}^* + \nu_{3i}) + \varepsilon_i$$

Estimamos o seguinte modelo

$$Y_i = (\beta_0 + \beta_3\delta_0) + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3\delta_3 X_{3i}^* + \underbrace{\beta_3\nu_{3i} + \varepsilon_i}_{e}$$

Temos estimadores viesados de β_0 e β_3, mas, se as hipóteses 1-3, anteriormente elencadas, forem satisfeitas, temos bons estimadores para β_1 e β_2

Mas, suponha que a terceira hipótese seja violada,

$$X_{3i} = \delta_0 + \delta_1 X_{1i} + \delta_2 X_{2i} + \delta_3 X_{3i}^* + \nu_{3i}$$

Nesse caso, o modelo com a proxy é

$$Y_i = (\beta_0 + \beta_3\delta_0) + (\beta_1 + \beta_3\delta_1)X_{1i} + (\beta_2 + \beta_3\delta_2)X_{2i} + \beta_3\delta_3 X_{3i}^* + \underbrace{\beta_3\nu_{3i} + \varepsilon_i}_{e}$$

o que causa um viés em β_1 e β_2

Capítulo 5 – Estimação por diferenças em diferenças

Vimos no capítulo anterior que um dos problemas do estimador de MQO é a violação da hipótese de exogeneidade das variáveis explicativas. Infelizmente, essa é uma questão bastante frequente quando se utilizam ferramentas econométricas para responder temas associados à formulação e avaliação de políticas públicas.

Como este manual é apenas uma introdução ao uso da análise de regressão, vamos, aqui, neste último capítulo, apresentar o estimador de diferenças em diferenças, que parte do modelo de regressão linear e do estimador de MQO.

Por que falar, aqui, do estimador de diferenças em diferenças? A principal motivação baseia-se no fato de ser este um estimador amplamente utilizado para avaliação de políticas públicas e de fácil compreensão, uma vez que se conheçam, ainda que de forma bem básica, os modelos de regressão linear.

Já existe hoje uma vasta literatura com aplicações diversas do método de diferenças em diferenças, que engloba o texto clássico de Ashenfelter e Card (1985) e estudos como os de Gruber (1994); Blundell, Duncan e Meghir (1998); e Di Tella e Schargrodsky (2004).

A ideia subjacente à proposta do estimador de diferenças em diferenças é a do chamado "experimento natural", ou seja, utiliza-se uma mudança política para definir os chamados grupos de controle e tratamento. O grupo de tratamento envolve os indivíduos que foram expostos ou beneficiados pelo programa ou política, enquanto que o grupo de controle é definido como aquele que apresenta indivíduos com características semelhantes ao do grupo de tratamento, a não ser pelo fato de não terem sido afetados pela política.

Como a própria nomenclatura sugere, o método de diferenças em diferenças (DD) é basedo no cálculo de uma dupla subtração ou diferença. A primeira diferença refere-se à diferença das médias da variável de resultado

entre os períodos anterior e posterior ao programa, tanto para o grupo de tratamento como para o grupo de controle. Já a segumda diferença refere-se à diferença da diferença anterior calculada entre os dois grupos. A grande limitação para utilização dessa técnica é que ela requer informação para ambos os grupos para pelo menos um período antes da implatanção da política ou programa e um depois dessa implantação.

De outro lado, a grande vantagem reside na sua capacidade de lidar com o viés de seleção associado às características não observáveis dos indivíduos, em especial aquelas que se mantêm constantes no tempo.

No primeiro período de tempo, ou seja, no chamado *baseline*, nem o grupo de controle nem o grupo de tratamento foram expostos à política ou programa. Já no segundo período, apenas o grupo de tratados foi afetado pela política ou programa.

Com *cross-sections* repetidas, podemos escrever o modelo para um membro genérico de um dos grupos como:

$$y = \beta_o + \beta_1 dB + \delta_0 d2 + \delta_1 d2.dB + u \text{ (83)}$$

onde y é o resultado de interesse, dB é uma dummy para o grupo de tratamento, $d2$ é uma variável *dummy* para o segundo período de tempo. O coeficiente de interesse δ_1 multiplica a interação da *dummy* de tratamento e da *dummy* de tempo.

O estimador de diferenças em diferenças é:

$$\hat{\delta} = (\overline{y}_{B,2} - \overline{y}_{B,1}) - (\overline{y}_{A,2} - \overline{y}_{A,1}) \text{ (84)}$$

Em alguns casos, é possível se ter uma análise de mudança de política mais convincente pelo refinamento da definição dos grupos de tratamento e controle.

Exemplo: Mudança em uma política de saúde voltada para idosos (+65 anos) – y variável de resultado em saúde.

- Possibilidade: utilizar dados apenas do estado afetado pela mudança de política, antes e depois, com o grupo de controle sendo as pessoas abaixo de 65 anos, e o de tratamento, as pessoas com 65 anos ou mais. Problema potencial? Estamos comparando dois grupos muito diferentes.

- Outra possibilidade: utilizar outro estado como grupo de controle – idosos. Problema potencial? É preciso garantir que, no outro Estado, não tenha ocorrido nenhuma política que pudesse ter beneficiado os idosos.

- Análise mais robusta: utilizar tanto um grupo de controle de outra faixa etária dentro do mesmo Estado, como grupo de controle de outro Estado na mesma faixa etária. Considere dois períodos de tempo, B representa o estado onde a política foi implementada, E é o grupo de idosos. Uma versão expandida da equação (83) seria:

$$y = \beta_0 + \beta_1 dB + \beta_2 dE + \beta_3 dB.dE + \delta_0 d2 + \delta_1 d2.dB + \delta_2 d2.dE + \delta_3 d2.dB.dE + u \quad (85)$$

O coeficiente de interesse agora é δ_3 o coeficiente do termo de interação tripla, $d2.dB.dE$ o estimador OLS, $\hat{\delta}_3$, pode ser expresso como:

$$\hat{\delta}_3 = (\bar{y}_{B,E,2} - \bar{y}_{B,E,1}) - (\bar{y}_{A,E,2} - \bar{y}_{A,E,1}) - (\bar{y}_{B,N,2} - \bar{y}_{B,N,1}) \quad (86)$$

onde o subscrito A representa o estado que não implementou a política e o subscrito N representa os não idosos. O estimador em (86) é chamado de estimador de diferenças em diferenças em diferenças (DDD).

O estimador de tripla diferença começa com a mudança temporal nas médias dos idosos no estado de tratamento e então retira a mudança na média dos idosos no estado de controle e a mudança nas médias para os não idosos no estado de tratamento. A ideia é que com isso se esteja controlando para duas formas de tendências potenciais que poderiam viesar os resultados:

- Mudanças no estado de saúde dos idosos entre os estados não relacionadas com a política.

- Mudanças no estado de saúde de todas as pessoas que vivem no estado de tratamento (devido a outras políticas estaduais que afetam a saúde de todos ou mudanças específicas na economia que afetam a saúde da população como um todo).

5.1 A incerteza e o arcabouço geral de diferenças em diferenças

De acordo com Angrist e Pischke (2009), a abordagem padrão de DD assume que toda a incerteza na inferência entra pelo erro amostral em estimar as médias de cada combinação grupo/período.

As metodologias de DD e DDD podem ser aplicadas em mais de dois períodos de tempo. No primeiro caso, um conjunto completo de *dummies* de período de tempo é adicionado a (83) e uma *dummy* de política substitui $d2.dB$; a *dummy* de política é definida como sendo igual a 1 para os grupos e períodos de tempo sujeitos à política.

Nessa estratégia, a restrição é de que a política tem o mesmo efeito em todos os anos. No modelo de tripla diferença, DDD, um conjunto completo de *dummies* é incluído para cada dois tipos de grupos e todos os períodos de tempo, bem como todos os pares de interações.

Por fim, quando se tem muitos períodos e grupos, a estrutura geral proposta por Bertrand, Duflo e Mullainathan (2004) é bastante útil. A equação no nível individual é:

$$y_{igt} = \lambda_t + \alpha_g + x_{gt}\beta + z_{igt}\gamma_{gt} + v_{gt} + u_{igt}, i = 1, \ldots, M_{gt} \quad (87)$$

i - indivíduo

g - grupo

t - tempo

Um forma de se escrever a equação (87) é:

$$y_{igt} = \delta_{gt} + \mathbf{z}_{igt}\gamma_{gt} + u_{git}, i = 1, \ldots, M_{gt} \quad (88)$$

O modelo em (88) é um modelo em nível individual no qual tanto o intercepto como as inclinações podem variar entre todos os (g,t) pares. Logo, vemos δ_{gt} como:

$$\delta_{gt} = \lambda_t + \alpha_g + \mathbf{x}_{gt}\beta + v_{gt} \quad (89)$$

Uma forma de estimar e fazer inferência em (89) é ignorar v_{gt}, e tratar as observações no nível individual como independentes. Quando v_{gt} está presente, a inferência resultante pode ser muito viesada.

5.2 Uma aplicação do modelo de diferenças em diferenças

Faria e Chein (2016) utilizam o modelo de diferenças em diferenças para avaliar o efeito da política de aceleração de aprendizagem do Estado de Minas Gerais, o Programa de Intervenção Pedagógica/Alfabetização no Tempo Certo (PIP/ATC), no desempenho dos alunos do 5º ano do ensino fundamental das escolas estaduais de Minas Gerais, bem como os possíveis efeitos do programa Acelerar para Vencer.

O PIP/ATC foi uma ação da SEE de Minas Gerais, com início em 2007, e que, inicialmente, abrange as escolas estaduais do estado inteiro. O programa surgiu com a principal meta de que toda criança estaria lendo e escrevendo até os 8 anos de idade. Posteriormente avançou tanto em termos de objetivos quanto em termos de público-alvo. O PIC/ATC visava apoiar as escolas e orientar os professores, os gestores da escola e os pais dos alunos, por meio de diversas ações de natureza pedagógica e com a finalidade principal de melhorar o desempenho dos alunos, principalmente em termos de leitura, escrita e operações matemáticas básicas.

O artigo do investiga se os alunos afetados pelo programa melhoraram seu rendimento em português e matemática, devido ao efeito isolado do PIP/ATC. Assume-se que um efeito positivo do programa vá gerar, no futuro, uma contribuição para redução da desigualdade, dada a teoria exposta nesta seção.

A análise empírica considera o acompanhamento do rendimento dos alunos da 5ª série do ensino fundamental nos anos de 2007, 2009 e 2011, a partir de dados da Prova Brasil, coletada de dois em dois anos pelo Inep. No ano de 2007, o programa foi implementado para os três primeiros anos do ensino fundamental e, com isso, o 5º ano ainda não havia sido atingido, o que aconteceria no ano seguinte. Nos anos de 2009 e 2011, os alunos do 5º ano da rede estadual já haviam sofrido os efeitos do programa. Dessa forma, os autores conseguem observar as diferenças dos níveis de proficiência em português e matemática desses alunos, de 2007 (não afetados) para 2009 e 2011 (efetivamente afetados pela política).

Para estimar o impacto do programa pelo método de diferenças em diferenças, Faria e Chein (2016) definem *T* = {1,0} como a participação ou não no programa, e t = {1,0} como os períodos anterior e posterior à intervenção, de forma que o estimador possa ser expresso como:

$$\beta dd = \{E[Y|T = 1, t = 1] - E[Y|T1, t0]\} - \{E[Y|T = 0, t = 1] - [E[Y|T = 0, t = 0]\} \quad (90)$$

Como esclarecem os autores, o estimador é a diferença da variação temporal do grupo de tratados em relação à variação temporal do grupo de controle.

Tal estimador pode, ainda, ser expresso pela dupla diferença:

$$\beta dd = \{E[Y|T = 1, t = 1] - E[Y|T = 0, t = 1]\} - \{E[Y|T = 1, t = 0] - E[Y|T = 0, t = 0]\} \quad (91)$$

A expressão (91) é um rearranjo da expressão (90), sendo interpretada como a diferença das diferenças de médias entre os dois grupos nos períodos anterior e posterior à implementação do programa.

É importante, destacar, aqui, as hipóteses assumidas pelos autores, para a identificação do efeito do programa avaliado, PIP/ATC, sobre o rendimento acadêmico dos alunos:

1. O grupo de tratamento teria a mesma trajetória da variável de resultado do grupo de controle, caso a intervenção não tivesse ocorrido.
2. A composição de ambos os grupos não se altera de forma significativa entre o período anterior e o período posterior ao tratamento, já que uma vez que isso ocorra, partes dos resultados obtidos poderiam estar se devendo a essa mudança de composição, e não ao efeito do tratamento.
3. Os grupos de tratamento e controle não são afetados de forma distinta por mudanças que ocorrem após o tratamento. Se isso ocorre, a mudança na variável de resultado pode deixar de estar representando o grupo de controle como o contra factual do grupo tratado, ou seja, o que teria ocorrido caso o tratamento não tivesse sido aplicado. São utilizadas variáveis de controle

observáveis em relação ao professor, ao diretor, ao aluno e à escola para isolar o efeito dessas variáveis na variação da variável dependente.

As Tabelas 2 e 3 a seguir apresentam os resultados das estimativas por diferenças em diferenças do efeito do programa PIP/ATC sobre a proficiência dos alunos de Minas Gerais encontrados por Faria e Chein (2016). O ano de 2007 é considerado o *baseline* pelos autores, é, portanto, a *dummy* de tempo omitida no modelo dos autores. As interações entre a *dummy* de 2009 e a *dummy* de beneficiários do programa e entre a *dummy* de 2011 e a *dummy* de beneficiários do programa representam os estimadores de diferenças em diferenças, ou seja, captam a diferença entre a variação na proficiência dos beneficiários e a variação na proficiência do grupo de controle ou comparação.

Ressalte-se que os autores encontram resultados positivos e significativos do PIP/ATC tanto para a proficiência em matemática como também para a proficiência em português.

Tabela 2 – Estimativas do efeito do PIP/ATC sobre a proficiência em matemática – método diferenças em diferenças

Variáveis	(1)	(2)	(3)	(4)	(5)	(6)
Ano 2009	13,37***	12,41***	12,24***	12,30***	12,21***	12,80***
	(0,0866)	(0,110)	(0,159)	(0,167)	(0,262)	(0,336)
Ano 2011	17,08***	15,16***	15,91***	15,63***	17,75***	17,29***
	(0,0880)	(0,109)	(0,158)	(0,176)	(0,319)	(0,390)
Beneficiários PIP/ATC	13,70***	14,29***	15,18***	14,91***	15,62***	15,62***
	(0,156)	(0,200)	(0,294)	(0,303)	(0,450)	(0,450)
Ano 2009 x Beneficiários do PIP/ATC	8,018***	6,950***	8,209***	8,312***	8,746***	8,604***
	(0,251)	(0,320)	(0,438)	(0,448)	(0,605)	(0,607)
Ano 2011 x Beneficiários do PIP/ATC	5,753***	4,555***	3,935***	4,544***	2,853***	2,577***
	(0,218)	(0,268)	(0,391)	(0,431)	(0,563)	(0,566)
Controles incluídos						
Características individuais[1]	Não	Sim	Sim	Sim	Sim	Sim

Background familiar[2]	Não	Não	Sim	Sim	Sim	Sim
Características do professor[3]	Não	Não	Não	Sim	Sim	Sim
Características do diretor[4]	Não	Não	Não	Não	Sim	Sim
Caracterísitcas da escola[5]	Não	Não	Não	Não	Não	Sim
Constante	190,6***	169,5***	161,5***	157,5***	150,9***	150,4***
	(0,0588)	(1,445)	(1,909)	(2,761)	(5,159)	(5,175)
Observações	1.884.229	1.046.910	508.072	430.576	215.498	213.945
R^2	0,048	0,186	0,206	0,209	0,225	0,225

Fonte: Elaboração própria com base na Prova Brasil e Censo Escolar (INEP) para os anos de 2007, 2009 e 2011

[1] nota: Vetor de características individuais inclui as variáveis: Raça, idade, sexo, nível socioeconômico, trabalho doméstico, trabalho fora, quando entrou na escola, tipo de escola que estudou e já reprovou

[2] nota: Vetor de *background* familiar inclui as variáveis: Escolaridade da mãe e escolaridade do pai

[3] nota: Vetor de características do professor inclui as variáveis: Escolaridade do professor, experiência lecionando e experiência na escola atual

[4] nota: Vetor de características do diretor inclui as variáveis: Experiência em educação, experiência como diretor, forma de contratação, escolaridade, programa de abandono, programa de reprovação, financiamento federal, financiamento

[5] nota: vetor de características da escola inclui o componente principal de infraestrutura escolar

*** Significativo a 1% ** Significativo a 5% * Significativo a 10%

Fonte: Faria e Chein (2016).

Tabela 3 – Estimativas do efeito do PIP/ATC sobre a proficiência em português – método diferenças em diferenças

Variáveis	(1)	(2)	(3)	(4)	(5)	(6)
Ano 2009	9,691***	8,353***	8,365***	8,375***	8,275***	9,073***
	(0,0833)	(0,106)	(0,152)	(0,159)	(0,250)	(0,320)
Ano 2011	14,72***	12,56***	12,96***	12,70***	14,55***	13,84***
	(0,0845)	(0,104)	(0,151)	(0,168)	(0,305)	(0,373)
Beneficiários PIP/ ATC	11,26***	11,46***	12,39***	12,42***	12,18***	12,22***
	(0,150)	(0,192)	(0,281)	(0,289)	(0,430)	(0,430)
Ano 2009 x Beneficiários do PIP/ATC	5,369***	4,593***	5,625***	5,792***	6,503***	6,300***
	(0,241)	(0,307)	(0,419)	(0,427)	(0,577)	(0,579)

Ano 2011 x Beneficiários do PIP/ATC	5,976***	5,157***	4,624***	5,226***	4,465***	4,234***
	(0,209)	(0,257)	(0,374)	(0,411)	(0,538)	(0,541)
Controles incluídos						
Características individiduais[1]	Não	Sim	Sim	Sim	Sim	Sim
Background familiar[2]	Não	Não	Sim	Sim	Sim	Sim
Características do professor[3]	Não	Não	Não	Sim	Sim	Sim
Características do diretor[4]	Não	Não	Não	Não	Sim	Sim
Caracterísitcas da escola[5]	Não	Não	Não	Não	Não	Sim
Constante	174,1***	165,3***	157,5***	155,6***	148,0***	147,4***
	(0,0565)	(1,385)	(1,824)	(2,634)	(4,926)	(4,942)
Observações	1.884.367	1.046.959	508,088	430.586	215.500	213.947
R^2	0,037	0,185	0,217	0,218	0,231	0,231

Fonte: Elaboração própria com base na Prova Brasil e Censo Escolar (INEP) para os anos de 2007, 2009 e 2011

[1] nota: Vetor de características individuais inclui as variáveis: Raça, idade, sexo, nível socioeconômico, trabalho doméstico, trabalho fora, quando entrou na escola, tipo de escola que estudou e já reprovou

[2] nota: Vetor de *background* familiar inclui as variáveis: Escolaridade da mãe e escolaridade do pai

[3] nota: Vetor de características do professor inclui as variáveis: Escolaridade do professor, experiência lecionando e experiência na escola atual

[4] nota: Vetor de características do diretor inclui as variáveis: Experiência em educação, experiência como diretor, forma de contratação, escolaridade, programa de abandono, programa de reprovação, financiamento federal, financiamento

[5] nota: vetor de características da escola inclui o componente principal de infraestrutura escolar

*** Significativo a 1% ** Significativo a 5% * Significativo a 10%

Fonte: Faria e Chein (2016).

Comentários finais

O intuito principal deste pequeno manual foi introduzir o leitor às ferramentas básicas de econometria, em especial, aos modelos de regressão linear. É apenas um primeiro passo para adentrar esse imenso universo das ferramentas estatísticas e econométricas que podem ser de grande utilidade para a avaliação de políticas públicas.

Ao invés de apresentar de todas possibilidades dos modelos de regressão linear, o que se pretendeu aqui foi simplesmente abrir a tampa de uma imensa caixa de utensílios, e apontar as limitações e hipóteses subjacentes aos modelos econométricos mais simples.

De outro lado, cabe uma lembrança final de que o instrumental estatístico, seja qual for, não subsiste sem a teoria e o conhecimento da questão que se busca analisar. Logo, para se avaliar uma política pública, é preciso, antes de mais nada, conhecer sobre tal política, o que motivou a sua elaboração, o que se pretende com a mesma, quem é o seu público-alvo, quais podem ser seus desdobramentos e todas as demais questões a ela atreladas.

Por fim, muitos são os pacotes econométricos capazes de dar respostas rápidas, estimações complexas, mas nada disso tem significado sem o conhecimento técnico do cientista social, do economista, do epidemiologista ou de qualquer outro profissional que domine o tema associado aos modelos estimados.

Referências bibliográficas

Anderson, Patricia. Fall 02 term of Economics 20 - Econometrics at Dartmouth College. available at http://www.dartmouth.edu/~econ20pa/.

Angrist, J. D.; Pischke, J.-S. *Mostly harmless econometrics:* an empiricist's companion. Massachusettts Institute of Technology and The London school of Economics, 2009. https://doi.org/10.1017/CBO9781107415324.004

Ashenfelter, Orley; Card, David. Using longitudinal structure of earnings to estimate the effect of trainning programs. *The Review of Economics and Statistics*, v. 67, p. 648-60, 1985.

Bertrand, M.; Duflo, E.; Mullainathan, S. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, v. 119, n. 1, p. 249-275, fev. 2004. https://doi.org/10.1162/003355304772839588

Blundell, Richard; Duncan, Alan; Meghir, Costas. Estimating labor supply responses using tax reforms. *Econometrica*, v. 66, n. 4, p. 827-861, 1998.

Heij, Christiaan *et al*. *Econometric methods with applications in business and economics*. New York: Oxford University Press Inc., 2004. ISBN 0–19–926801–0

Di Tella, Rafael; Schargrodsky, Ernesto. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *The American Economic Review*, v. 94, n. 1, p. 115-133, 2004.

Faria, Victor N.; Chein, Flávia. Alfabetização e desempenho escolar: uma análise de intervenções recentes em Minas Gerais. Planejamento e Políticas Públicas, Rio de Janeiro, Ipea, n. 46, p-293-332, 2016.

Gruber, J. The incidence of mandated maternity benefits. *The American Economic Review*, v. 84, n. 3, p. 622-641, 1994.

Gujarati, D. N.; Porter, D. C. *Basic Econometrics* . 5h Ed. New York: McGraw-Hill Irwin, 2009

Stock, J. H.; Watson, M. W. *Introduction to Econometrics.* 3. ed. Addison-Wesley Series in Economics, v. 1. Addison-Wesley, 2010.

Heckman, J. J. Econometric causality. *International Statistical Review*, v. 76, n. 1, p. 1-27, 2008. https://doi.org/10.1111/j.1751-5823.2007.00024.x

Wooldridge, J. Introductory econometrics: a modern approach. 4. Ed. South-Western Cengage Learning, 2009.