Revista de Contabilidade e Organizações
www.rco.usp.br

DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2018.152100

Journal of Accounting and Organizations
www.rco.usp.br

O ensino do modelo clássico de regressão linear por meio de simulação de Monte Carlo

Teaching the classical linear regression model using Monte Carlo simulation

Marcelo Sanches Pagliarussi[a]

[a] *Universidade de São Paulo*

Palavras-chave
Simulação de Monte Carlo.
Modelo clássico de regressão linear.
Distribuição amostral.
Estimadores de mínimos quadrados.

Resumo

Este trabalho apresenta um conjunto de estudos de Monte Carlo, usando *softwares* de planilha eletrônica, que pode ser usado para facilitar a aprendizagem do conceito de distribuição amostral em um contexto de aprendizagem do modelo clássico de regressão linear. A partir da construção de duas planilhas básicas, uma para regressão simples e outra para regressão múltipla, outras planilhas podem ser facilmente obtidas com pequenas alterações no processo gerador de dados. As alterações que podem ser introduzidas incluem variações no tamanho das amostras e em diversas características do termo de erro, como sua variância, valor médio e função de probabilidade. Também podem ser introduzidas correlações entre os regressores no modelo de regressão múltipla. Um professor de econometria introdutória pode usar o conjunto de planilhas de modo a obter figuras e tabelas que facilitam a visualização do desempenho dos estimadores de mínimos quadrados ordinários para diferentes situações. Deste modo, os estudantes podem compreender na prática como as violações nas premissas do modelo clássico de regressão linear afetam om desempenho dos estimadores de mínimos quadrados ordinários e dos testes de hipóteses usualmente empregados no contexto da análise de regressão. As violações trabalhadas no presente trabalho incluem heterocedasticidade, omissão de variáveis relevantes, erros não normais e multicolinearidade.

Keywords
Monte Carlo simulation.
Classical linear regression model.
Sampling distribution.
Least square estimators.

Abstract

This work presents a series of Monte Carlo studies using spreadsheet software aimed at facilitating the understanding of the concept of sampling distributions when students are learning the classical linear regression model. Starting from two basic spreadsheets, one for simple regression and the other for multiple regression, other spreadsheets can be easily built by introducing minor alterations in the data-generating process. The modifications that can be introduced include variations in sample size, and several characteristics of the error term, such as its variance, expected value and probability function. Different degrees of correlation between regressors can also be included. A teacher of basic econometrics can use the set of spreadsheets to obtain graphics and tables that enable the visualization of the performance of ordinary least squares estimators in different situations. Thus, students can understand in practice how violations in the underlying hypotheses of the classical linear regression model affect the performance of least square estimators, as well as the tests of hypotheses that usually accompanying the process of regression analysis. The violations analyzed in the present work include heteroscedasticity, omission of relevant variables, non-normal errors and multicollinearity.

Informações do Artigo
Recebido: 26 de novembro de 2018
Aceito: 18 de dezembro de 2018
Publicado: 28 de dezembro de 2018

Implicações práticas

O artigo apresenta um conjunto de práticas de ensino que pode facilitar o entendimento da inferência estatística na análise de regressão linear simples e múltipla. Assim, tem o potencial de contribuir para a melhoria do ensino de tais conceitos nos cursos de graduação em ciências contábeis, administração e economia.

Autor Correspondente: Tel. (16) 3315-9028
E-mail: marcelosp@usp.br (M. S. Pagliarussi)
Universidade de São Paulo. Faculdade de Economia, Administração e Contabilidade de Ribeirão Preto. Avenida Bandeirantes, 3900 - Monte Alegre - Ribeirão Preto/SP - 14040-905, Brasil.

1 INTRODUÇÃO

Muito provavelmente, um professor de econometria básica que segue os bons livros existentes sobre o assunto irá desenvolver aulas que requerem um nível razoavelmente elevado de raciocínio teórico e matemático. Tal abordagem é comum inclusive em cursos direcionados ao público da área de negócios (Becker e Greene, 2001). Este mesmo professor, ao abordar o conceito de distribuição amostral de um estimador, provavelmente irá perceber nos olhos dos seus alunos a dificuldade de compreender o seu significado. Diversos professores-autores relataram tal percepção. Becker e Greene (2001) observaram que seus alunos compreendiam facilmente o papel que o acaso representa na obtenção de uma dada amostra. Porém, os autores notaram que os alunos têm imensa dificuldade em entender que as estatísticas calculadas a partir de tal amostra são igualmente fruto do acaso, cujos valores podem ser representados em um histograma de modo a produzir a distribuição amostral da estatística.

Kennedy (1998) afirma que, contrariamente ao que imaginam os professores, ao final do curso de econometria básica a ampla maioria dos estudantes não compreende a lógica fundamental da estatística, representada pelo conceito de distribuição amostral. O autor ressalta que os alunos aprendem a realizar procedimentos, como calcular a variância, executar uma regressão, testar uma hipótese, e eles sabem que serão aprovados no curso se memorizarem como tais técnicas funcionam. Entretanto, usualmente os cursos levam os alunos à percepção de que a estatística é um ramo da matemática, e estes não desenvolvem a habilidade de usar a estatística como uma lente para enxergar o mundo, pois o conceito de distribuição amostral constitui essa lente (Kennedy, 1998).

Barreto e Howland (2005), Chance, Garfield e del Mas (1999) e Dyck e Gee (1998) manifestaram essencialmente o mesmo desconforto com suas observações de que muitos estudantes aprovados com boas notas em econometria básica não desenvolvem a compreensão necessária do conceito de distribuição amostral, nem a capacidade de aplicar o conceito em uma linha de raciocínio coerente. Chance, del Mas e Garfield (2004) explicam que a dificuldade em apreender o conceito de distribuição amostral pode residir no fato de que o conceito requer que os estudantes integrem e apliquem vários outros conceitos obtidos em diferentes momentos do curso de estatística, assim como raciocinem a respeito do comportamento hipotético de muitas amostras. Na visão dos autores, mesmo que o Teorema do Limite Central forneça um modelo teórico para o comportamento das distribuições amostrais, os alunos têm dificuldade em aplicar tal modelo em contextos práticos.

Kennedy (1998) sugere que uma mudança fundamental deveria ocorrer nos cursos de econometria básica: a inserção de estudos de Monte Carlo como ferramenta pedagógica para investigação das propriedades da distribuição amostral de um estimador. O autor inclusive chega a afirmar que havia suprimido a maior parte das demonstrações matemáticas em seus cursos. Para Kennedy (1998), a investigação de distribuições amostrais por meio de estudos de Monte Carlo deveria constituir a maior parte da carga horária dos cursos de econometria, pois tal atividade permite aos estudantes alcançar a compreensão de todas as dimensões do curso. O autor conclui que a apresentação de técnicas de estimação avançadas não agrega nada se os estudantes não compreendem os princípios fundamentais que as sustentam.

Apesar dos fortes argumentos de Kennedy (1998), o uso de estudos de Monte Carlo no ensino de econometria é pouco difundido. Becker e Greene (2001) destacam que a maioria dos professores deixa o desenvolvimento do conceito de distribuição amostral a cargo da imaginação dos estudantes, mesmo tendo à sua disposição computadores e softwares que permitem o desenvolvimento real do histograma dos valores possíveis de uma estatística amostral. Bekkerman (2015) também chama a atenção para o pouco uso de simulações no ensino de econometria, possivelmente pela falta de conhecimento dos professores a respeito dos potenciais benefícios pedagógicos da ferramenta.

Barreto e Howland (2005) ressaltam sua frustração com o ensino de econometria baseado em equações e na prova de proposições. Tal abordagem, segundo os autores, resulta na ênfase na memorização ao invés do uso efetivo do conceito em situações reais. O uso de simulações permite a visualização dinâmica e a repetição de situações interessantes. Por exemplo, por meio do Excel os estudantes podem instantaneamente obter resultados novos e reconstruir tabelas e gráficos após terem alterado o valor de um parâmetro ou obtido uma nova amostra (Barreto e Howland, 2005). Os autores defendem que, por meio do uso de planilhas, os estudantes conseguem associar os valores com os símbolos abstratos presentes nas equações, e conseguem ver um teorema em operação quando um resultado esperado é observado repetidamente em muitas amostras. Barreto e Howland (2005) concluem que é irônico que as simulações desempenhem papel proeminente nos estudos avançados de econometria, enquanto que o ensino da disciplina padece nos métodos ultrapassados de memorização e prova.

Com base nas recomendações de Kennedy (1998), Judge (1999) desenvolveu um exercício de Monte Carlo simples em Excel, direcionado a permitir que os estudantes simulem a obtenção de 100 amostras aleatórias contendo observações de duas variáveis, X e Y, e calculem as estimativas dos parâmetros de um modelo de regressão simples $Y=\beta_1+\beta_2 X+u$ para cada amostra. Entre os objetivos do exercício estavam a análise da distribuição amostral dos 100 valores obtidos do estimador de mínimos quadrados ordinários $\hat{\beta}_2^{MQO}$, como sua média, variância e a discussão de questões como viés e ausência de viés. Além disso, os estudantes precisavam analisar se o histograma construído com os 100 valores obtidos de $\hat{\beta}_2^{MQO}$ se assemelhava a uma distribuição normal. Craft (2003) oferece contribuição semelhante ao detalhar as etapas necessárias para modelar o processo gerador de dados, obter amostras aleatórias repetidas e calcular as estimativas dos parâmetros de uma regressão simples usando planilhas eletrônicas.

Mais recentemente, Briand e Hill (2013) expõem detalhadamente a realização de estudos de Monte Carlo usando planilhas em uma aplicação envolvendo regressão linear simples. Os autores desenvolvem dois exercícios. No primeiro, explicam como os alunos podem obter a distribuição amostral do estimador de inclinação por meio do procedimento de mínimos quadrados ordinários. No segundo, são obtidas as estimativas de intervalos de confiança para o coeficiente de inclinação $\hat{\beta}_2^{MQO}$.

O presente artigo estende as propostas de Judge (1999), Craft (2003) e Briand e Hill (2013) ao apresentar vários exercícios desenvolvidos por meio da aplicação da simulação de Monte Carlo aplicada em situações de regressão linear simples como múltipla. Por meio da ferramenta pedagógica apresentada aqui, os estudantes poderão desenvolver simulações com o objetivo de: (1) obter a distribuição amostral dos estimadores $\hat{\beta}^{MQO}$ calculados para 1.000 amostras aleatórias repetidas; (2) analisar as propriedades da distribuição amostral dos estimadores, como média, variância e forma; (3) analisar como o tamanho da amostra impacta no desempenho do teste F para significância global da regressão, e do teste t para significância dos coeficientes individuais; (4) analisar como as violações das premissas do modelo clássico de regressão linear afetam as propriedades da distribuição amostral dos estimadores e o desempenho dos testes F e t. As violações analisadas incluem termo de erro com variância heterocedástica, termo de erro com distribuição não normal, termo de erro com média diferente de zero, omissão de variáveis relevantes e existência de multicolinearidade entre regressores.

A seção 2 a seguir apresenta a técnica de simulação de Monte Carlo e sua aplicação em um contexto de análise de regressão linear. Na seção 3, são explicados os procedimentos para obtenção das amostras repetidas, estimação pontual e intervalar dos parâmetros do modelo de regressão e obtenção das estatísticas da regressão, como R^2, F, t e suas respectivas significâncias. A seção 4 apresenta uma breve discussão da abordagem proposta e suas possíveis extensões. Por fim, a última seção conclui brevemente a proposta.

2 ESTUDOS DE MONTE CARLO E SUA APLICAÇÃO NO CONTEXTO DA ANÁLISE DE REGRESSÃO LINEAR

Simulação de Monte Carlo refere-se ao emprego de modelos artificiais para representar processos reais de geração de dados, de modo a obter uma maior compreensão de tais processos (Barreto e Howland, 2005). As simulações usam geradores de números aleatórios para recriar os processos estocásticos, e o fazem repetidas vezes para observar os resultados obtidos (Barreto e Howland, 2005; Hill, Griffiths e Lim, 2011). Por meio da simulação, podem ser criadas muitas amostras de tamanho N e assim examinar as propriedades de diferentes métodos de estimação, inclusive o seu comportamento em situações distantes do ideal, como é o caso de muitas aplicações na área de negócios (Hill, Griffiths e Lim, 2011). A Figura 1 a seguir apresenta o fluxo de trabalho em um estudo de Monte Carlo.

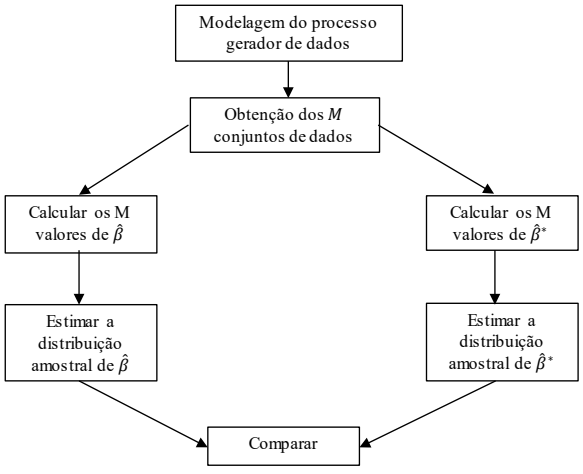

Figura 1. Estrutura de um estudo de Monte Carlo. Adaptada de Kennedy (2003)

Do ponto de vista pedagógico, uma razão importante para o uso de experimentos de Monte Carlo é propiciar o entendimento dos conceitos de amostragem repetida e propriedades da distribuição de amostragem de um estimador, que são conceitos cruciais para o entendimento de econometria (Kennedy, 2003). O autor descreve as etapas do desenvolvimento de um estudo de Monte Carlo (2003):

1. Modelar o processo gerador de dados: um estudo de Monte Carlo se inicia com a construção do modelo que permita ao computador imitar o processo gerador de dados, incluindo seu componente estocástico. Por exemplo, pode ser especificado que N valores de X_1 um termo de erro u irão produzir N valores de Y de acordo com a equação $Y=\beta_1+\beta_2 X_1+u$. Na equação, β_1e β_2 são números específicos e conhecidos, as N observações de X_1 correspondem a realizações exógenas dos valores da variável, e os N valores de u são obtidos aleatoriamente a partir de uma distribuição normal com média zero e variância conhecida σ^2. Quaisquer características especiais do processo gerador de dados podem ser incluídas no modelo. Por exemplo, o termo de erro pode ser gerado a partir de uma distribuição normal com média diferente de zero e variância conhecida σ^2. Também é possível fazer com que a variância do termo de erro dependa seja função de X_1. Por fim, os erros podem ser gerados a partir de uma distribuição de probabilidade diferente da normal. Um aspecto importante a destacar é que todos os valores dos parâmetros são conhecidos, porque a pessoa que conduz o estudo é que escolhe tais valores.

2. Criação dos conjuntos de dados: após o modelo do processo gerador de dados ter sido construído e inserido no computador, os dados artificiais podem ser criados. Deste modo, uma amostra completa com N valores de Y,X_1 e u é obtida. Note que tal conjunto artificial de dados pode ser enxergado como um exemplo dos dados reais que um pesquisador iria obter quando tivesse que lidar com o problema de estimação que o modelo representa. É importante destacar também que o conjunto de dados depende crucialmente dos valores obtidos para o termo de erro. Um conjunto diferente de N valores de u iria alterar significativamente os valores de Y obtidos para o mesmo problema. Se tal processo de amostragem for repetido 1.000 vezes, por exemplo, teremos 1.000 conjuntos de amostras com tamanho N, chamadas amostras repetidas.

3. Cálculo das estimativas: cada uma das amostras obtidas será usada como input para o cálculo do valor do estimador $\hat{\beta}_2$. Então, se estivermos trabalhando com 1.000 amostras, podemos obter 1.000 estimativas $\hat{\beta}_2$ para o parâmetro β_2. As estimativas podem ser vistas como 1.000 sorteios aleatórios de valores retirados da distribuição de $\hat{\beta}_2$.

4. Estimação das propriedades da distribuição amostral: as 1.000 extrações aleatórias da distribuição amostral de $\hat{\beta}_2$ podem ser usadas como dados para estimar as propriedades dessa distribuição. As propriedades de maior interesse são o valor esperado e a variância, os quais podem ser usados para estimar o viés e o erro quadrático médio do estimador. Na etapa 3 as estimativas obtidas por meio de um estimador alternativo $\hat{\beta}_2^*$ também podem ser obtidas, de modo que as propriedades da distribuição amostral de $\hat{\beta}_2^*$ podem ser comparadas com as propriedades da distribuição de $\hat{\beta}_2$.

Assim, de acordo com a abordagem de Briand e Hill (2013), e as orientações de Kennedy (1998, 2003), as aplicações a seguir foram desenvolvidas com o objetivo de servirem como ferramentas pedagógicas e contribuir para o entendimento das propriedades dos estimadores e testes associados à regressão linear simples e múltipla, a partir da simulação de um processo de amostragem repetida.

3 APLICAÇÕES DA SIMULAÇÃO DE MONTE CARLO NA ANÁLISE DE REGRESSÃO USANDO O EXCEL

Nesta seção são desenvolvidas duas aplicações da simulação de Monte Carlo. Na Aplicação 1, as etapas da construção do processo de simulação que irá resultar na obtenção de 1.000 amostras de pares de valores (x,y) em uma planilha eletrônica são detalhadamente descritas. Na sequência da obtenção das amostras, é apresentado o passo a passo da obtenção das estimativas de ponto e de intervalo, das estatísticas de regressão. Também são obtidas as estatísticas F e t e os resultados dos respectivos testes de hipóteses. A Aplicação 2 estende os procedimentos da Aplicação 1 para o contexto de regressão linear múltipla.

3.1 Aplicação 1: estimação dos parâmetros na RLS usando diferentes tamanhos de amostra

O desenvolvimento de estudos de Monte Carlo começa com a definição do processo gerador de dados. Assim, a partir de um modelo $Y=\beta_1+\beta_2 X_1+u$ é necessário definir os valores dos parâmetros β_1 e β_2 e os valores de X na amostra. Em seguida, deve-se inserir a fórmula para obtenção dos valores de Y. Isto é feito usando a função ALEATÓRIO do Excel para gerar os valores de u, de modo a tornar a relação entre Y e X não determinística. A Tabela 1 a seguir apresenta os parâmetros definidos para as simulações da Aplicação 1.

Tabela 1. Parâmetros do processo gerador de dados

$Y=\beta_1+\beta_2 X_1+u$

	A	B	C	D
1	N=	20	sigma=	25
2	X11=	100	beta1=	50
3	X12=	200	beta2=	0.25
4				
5	X	y		
6	=$B$2			
7	=$B$2			
...				
15	=$B$2			
16	=$B$3			
...				
25	=$B$3			

Fonte: Elaborado pelo autor.

Inicialmente a o processo envolverá amostras aleatórias com 20 observações de *X, u* e *Y*. Em cada amostra, 10 observações terão valor x_1=100 e 10 terão valor x_2=200. Tais valores irão permanecer fixos no processo de amostragem repetida.

Os valores escolhidos para β_1 e β_2 são 50 e 0,25 respectivamente, e os valores de *u* em cada amostra são distribuídos de forma independente e normal, com média zero e variância homocedástica, ou seja, σ^2=625 para qualquer valor de *x*. O termo de erro de cada observação é gerado a partir da combinação entre as funções do Excel INV.NORM.N e ALEATÓRIO. A primeira função retorna o inverso da distribuição cumulativa normal para valores específicos de média e desvio padrão. Sua sintaxe é INV.NORM.N(probabilidade;média;desv_padrão), na qual a probabilidade é $P(X\leq x)$, definida entre 0 e 1.

Deste modo, para obter valores aleatórios de um termo de erro com distribuição normal, média zero e variância constante igual a 625, basta inserir na fórmula de *y* o termo INV.NORM.N(ALEATÓRIO();0;25), uma vez que a função ALEATÓRIO() retorna um valor aleatório uniformemente distribuído entre 0 e 1. Como *x* é fixo nas amostras repetidas, resulta que *y* irá se distribuir normalmente com média $E(y \mid x)=\beta_1+\beta_2 x$. A Tabela 2 mostra a fórmula utilizada para gerar os valores de *y* da primeira amostra (y_1), com os pares de valores (x,y_1). As fórmulas inseridas nas células B6:B25 devem ser coladas no mesmo intervalo das colunas C a ALM, produzindo assim 1.000 amostras, $\{x,y_1\},\{x,y_2\},\{x,y_3\},\ldots,\{x,y_{1000}\}$.

Tabela 2. Configurações do processo gerador de dados para um estudo de Monte Carlo em regressão linear simples

	A	B	C	D
1	N=	20	sigma=	25
2	X11=	100	beta1=	50
3	X12=	200	beta2=	0.25
4				
5	x	y1	y2	y3
6	100	=$D$2+$D$3*A6+INV.NORM.N(ALEATÓRIO();0;$D$1)		
7	100			
...				
15	100	...		
16	200			
...				
25	200	=$D$2+$D$3*A25+INV.NORM.N(ALEATÓRIO();0;$D$1)		

Fonte: Elaborado pelo autor.

A Figura 2 a seguir apresenta os resultados obtidos com a simulação de 1.000 amostras contendo pares de realizações de *x* e *y*, sendo que os valores de *x* são fixos em todas as amostras. Tais resultados foram obtidos colando as fórmulas do intervalo B6:B25 da planilha para o mesmo intervalo nas colunas C à ALM.

	A	B	C	D	E	F	ALI	ALJ	ALK	ALL	ALM
1	N=	20		sigma=	25						
2	X11	100		beta1=	50						
3	X12	200		beta2=	0.25						
4											
5	x	y1	y2	y3	y4	y5	y996	y997	y998	y999	y1000
6	100	59.966	113.946	77.866	111.651	66.344	97.910	61.237	54.111	103.871	49.079
7	100	81.968	68.045	103.446	93.802	123.801	64.987	41.679	59.224	57.270	79.419
8	100	81.522	88.330	41.141	100.375	73.378	88.184	84.175	66.399	65.649	19.360
9	100	46.393	142.252	45.181	80.286	23.152	147.200	81.420	54.633	67.493	100.949
10	100	109.506	34.583	79.962	89.125	52.020	48.038	31.248	67.798	79.364	93.482
11	100	91.666	43.460	13.507	79.780	96.049	78.090	60.710	41.599	71.588	71.455
12	100	80.417	82.526	97.166	70.391	100.895	77.595	71.292	48.949	104.944	53.276
13	100	38.894	63.031	50.113	110.649	43.796	82.165	110.189	35.275	70.434	75.795
14	100	109.348	94.231	60.103	107.875	89.560	97.202	74.225	91.651	39.841	53.738
15	100	108.380	80.951	39.351	44.491	28.719	46.277	97.011	94.213	79.589	53.524
16	100	57.779	44.315	83.065	134.153	35.208	92.257	54.324	63.854	88.146	63.144
17	100	115.713	78.858	66.407	75.092	84.636	55.220	56.242	64.040	13.283	115.303
18	100	87.442	108.764	123.893	96.144	94.476	63.413	107.811	79.245	72.520	34.208
19	100	90.876	45.105	72.339	42.517	77.781	82.465	78.474	61.313	67.927	91.909
20	100	94.681	91.053	66.814	100.566	64.046	93.524	68.579	109.906	81.474	34.635
21	100	67.281	53.382	95.114	52.695	76.194	71.274	88.269	67.053	100.837	81.303
22	100	107.170	73.827	93.411	90.165	84.982	114.762	90.624	32.228	98.682	107.628
23	100	61.496	95.651	54.797	66.370	89.985	77.496	61.267	82.684	98.377	77.233
24	100	36.027	110.856	75.360	15.931	95.119	12.390	131.381	40.603	25.526	68.084
25	100	87.497	64.771	34.667	94.866	77.070	70.848	105.123	87.744	115.454	85.791

Figura 2. Recorte de tela com os parâmetros do modelo e os valores simulados para 1.000 amostras de valores (*x,y*)

Com as 1.000 amostras disponíveis, o próximo passo é proceder com a estimativa dos parâmetros de regressão para cada *y* em função de *x*. A função PROJ.LIN do Excel calcula as estatísticas da regressão linear obtidas por meio do processo de estimação por mínimos quadrados ordinários e retorna uma matriz com tais estatísticas. Como a função retorna uma matriz de valores, ela precisa ser inserida como uma fórmula de matriz, mas é possível obter o valor de uma célula específica da matriz combinando a função PROJ.LIN com a função ÍNDICE, como será descrito a seguir.

A função PROJ.LIN parte da equação da reta $y = mx + b$ ou $y = m_1 x_1 + m_2 x_2 + ... + b$, em que os valores *y* são função dos valores *x* e de uma constante *b*, e uma matriz contendo as estimativas para os coeficientes m_i e *b* e as estatísticas de regressão adicionais, como o erro padrão dos coeficientes, o R-quadrado, o erro padrão de *y*, a estatística *F*, os graus de liberdade, a soma dos quadrados da regressão e dos resíduos. A Tabela 3 mostra a ordem em que as estatísticas são retornadas.

Tabela 3. Matriz obtida com a aplicação da função PROJ.LIN

	1	2	3	4	5	6
1	m_n	$m_{(n-1)}$	...	m_2	m_1	b
2	ep_n	$ep_{(n-1)}$	...	ep_2	ep_1	ep_b
3	R^2	ep_y				
4	F	gl				
5	SQ_{reg}	SQ_{res}				

Fonte: Elaborado pelo autor.

A sintaxe da função PROJ.LIN é: PROJ.LIN(val_conhecidos_y, [val_conhecidos_x], [constante], [estatísticas]).

Na qual val_conhecidos_y representa a coluna com os valores de *y*, [val_conhecidos_x] representa a(s) coluna(s) com os valores de x_i, [constante] deve assumir um valor "0" ou "Falso" caso a regressão seja estimada sem o termo *b*, e "1" ou "Verdadeiro" caso queiramos que a regressão seja estimada com o termo *b*. Igualmente, se quisermos as estatísticas adicionais da regressão deveremos colocar "1" ou "Verdadeiro" em [estatísticas], ou "0" ou "Falso" caso não queiramos as estatísticas adicionais.

É possível desmembrar as células da matriz resultante da função PROJ.LIN por meio da função ÍNDICE, a qual retorna o valor que existe dentro de uma tabela. Deste modo, podemos obter o resultado de cada célula da matriz descrita na Tabela 3 individualmente, o qual pode ser convenientemente posicionado na planilha sem estar vinculado a uma matriz. A sintaxe da função ÍNDICE é: ÍNDICE(matriz; núm_linha; [núm_coluna]).

No nosso caso, PROJ.LIN representará a matriz a que se refere a sintaxe da função ÍNDICE. Os termos núm_linha e [núm_coluna] são usados para especificar a célula da matriz que contém o valor a ser exibido. Por exemplo, se quisermos obter a estatística *F* da regressão entre os valores de y_1 e *x* apresentados na Figura 2, devemos selecionar uma célula vazia da planilha e inserir a fórmula =ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);4;1). Os dois últimos termos da fórmula se referem à posição da estatística F na matriz dos resultados de PROJ.LIN, conforme indicado na Tabela 3. Do mesmo modo, se quisermos obter o R-quadrado da mesma regressão, devemos selecionar uma célula vazia da planilha e inserir a fórmula =ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);3;1), e assim por diante.

A partir da aplicação das funções PROJ.LIN e ÍNDICE, iremos obter, para cada regressão de y*i* em função de *x*, as seguintes informações:

•A estatística $R^{2;}$

•A estatística *F;*

•A informação se o valor da estatística *F* é significante a 5% ou não;

•A estimativa do parâmetro β_1, obtida por meio do estimador de mínimos quadrados ordinários $\hat{\beta}_{1;}$

•A estimativa do intervalo de confiança a 95% para o parâmetro $\beta_{1;}$

•A informação se o intervalo de confiança contêm $\beta_{1;}$

•O valor-p de $\hat{\beta}_{1;}$

•A informação se a estimativa para β_1 é significante a 5%;

•A estimativa do parâmetro β_2, obtida por meio do estimador de mínimos quadrados ordinários $\hat{\beta}_{2;}$

•A estimativa do intervalo de confiança a 95% para o parâmetro β_2 *;*

•A informação se o intervalo de confiança contêm $\beta_{2;}$

•O valor-p de $\hat{\beta}_{2;}$

•A informação se a estimativa para β_2 é significante a 5%.

De posse de tais informações, iremos contar:

•O número de regressões para as quais foi obtido um valor de *F* significante a 5%;

•O número de regressões para as quais os intervalos de confiança contêm $\beta_{1;}$

•O número de regressões para os quais o valor estimado para β_1 é significante a 5%;

•O número de regressões para as quais os intervalos de confiança contêm $\beta_{2;}$

•O número de regressões para os quais o valor estimado de β_2 é significante a 5%.

Para obter tais informações é necessário fazer uso de outras funções do Excel, tais como:

•DIST.F: retorna o resultado da distribuição de probabilidade *F*, para um dado valor de *F* e seus respectivos graus de liberdade;

•SE: permite que sejam feitas comparações lógicas entre um valor e aquilo que se espera;

•OU: determina se alguma condição em um teste é verdadeira;

•CONT.SE: conta o número de células que atendem a um critério;

•DIST.T.BC: retorna o valor da distribuição t de Student bicaudal, para um dado valor de t e seu respectivo número de graus de liberdade;

•INV.T.BC: retorna o inverso bicaudal da distribuição t de Student, para uma dada probabilidade e um número de graus de liberdade.

A Tabela 4 mostra as fórmulas usadas para obter as informações das regressões realizadas com as 1.000 amostras obtidas na simulação.

Tabela 4. Fórmulas para obter as estimativas dos parâmetros, estatísticas adicionais e demais informações na regressão simples

Célula	Fórmula	Resultado	Copiada para
B30	=INV.T.BC(0.05;C29)	*t* crítico (95%)	
B31	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);3;1)	R2	C31:ALM31
B32	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);4;1)	Estatística *F*	C32:ALM32
B33	=DIST.F(B32;$B$29;$C$29;FALSO)	*F* de significância	C33:ALM33
B34	=SE(OU(B33<0.05);"Sim";"Não")	Sim ou Não	C34:ALM34
B35	=CONT.SE(B34:ALM34;"Sim")	Número de regressões com *F* significante a 5%	
B37	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);1;2)	Estimativa para beta 1	C37:ALM37
B38	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);2;2)	Erro padrão da estimativa para beta 1	C38:ALM38
B39	=B38-$B$30*B39	Limite inferior (95%) para beta 1	C39:ALM39
B40	=B38+$B$30*B39	Limite superior (95%) para beta 1	C40:ALM40
B41	=SE(OU($E$2<B40;$E$2>B41);"Não";"Sim")	Reporta se o IC contém beta 1	C41:ALM41
B42	=CONT.SE(B42:ALM42;"Sim")	Número de IC que contém beta 1	
B43	=DIST.T.BC(ABS(B38/B39);$C$29)	Valor-p da estimativa de beta 1	C43:ALM43
B44	=SE(OU(B44<0.05);"Sim";"Não")	Se a estimativa de beta 1 é significante ou não	C44:ALM44
B45	=CONT.SE(B45:ALM45;"Sim")	Número de regressões que produziram estimativas de beta 1 significantes a 5%	C45:ALM45
B47	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);1;1)	Estimativa para beta 2	C47:ALM47
B48	=ÍNDICE(PROJ.LIN(B6:B25;$A$6:$A$25;1;1);2;1)	Erro padrão da estimativa de beta 2	C48:ALM48
B49	=B49-$B$30*B50	Limite inferior (95%) para beta 2	C49:ALM49
B50	=B49+$B$30*B50	Limite superior (95%) para beta 2	C50:ALM50
B51	=SE(OU($E$3<B51;$E$3>B52);"Não";"Sim")	Reporta se o IC contém beta 2	C51:ALM51
B52	=CONT.SE(B53:ALM53;"Sim")	Número de IC que contém beta 2	
B53	=DIST.T.BC(ABS(B49/B50);$C$29)	Valor-p da estimativa de beta 2	C31:ALM31
B54	=SE(OU(B55<0.05);"Sim";"Não")	Se a estimativa de beta 2 é significante ou não	C31:ALM31
B55	=CONT.SE(B56:ALM56;"Sim")	Número de regressões que produziram estimativas de beta 2 significantes a 5%	

Fonte: Elaborado pelo autor.

Deste modo, obtemos uma planilha que contém a simulação de 1.000 amostras com valores de *y* e *x*, e para cada amostra obtivemos também os resultados e as estatísticas da análise de regressão por mínimos quadrados ordinários. A Figura 3 apresenta um recorte de tela da planilha obtida.

	A	B	C	D	E	F	G	H	I	J
1	N=	20		sigma=	25					
2	X11	100		beta1=	50					
3	X12	200		beta2=	0.25					
4										
5	x	y1	y2	y3	y4	y5	y6	y7	y8	y9
6	100	68.043	69.242	89.793	121.891	57.750	103.012	122.262	62.518	93.226
7	100	37.501	55.555	80.926	72.541	46.010	87.694	34.842	50.157	60.344
8	100	72.459	57.733	34.245	86.767	50.382	89.409	67.566	112.765	76.861
9	100	77.030	100.557	57.212	83.017	119.115	37.985	25.380	65.398	109.535
10	100	74.688	64.527	123.119	33.649	107.177	20.340	69.491	119.052	130.074
11	100	102.879	72.905	56.927	57.874	121.850	56.808	91.530	53.455	90.099
12	100	61.175	67.108	66.428	59.681	118.253	83.460	82.248	71.565	86.931
13	100	32.899	70.853	122.036	92.680	55.124	90.877	112.172	108.872	64.445
14	100	53.660	56.844	84.075	44.764	54.000	63.660	65.392	42.238	51.808
15	100	80.213	120.893	37.195	92.838	77.325	51.975	72.258	110.848	80.787
16	200	76.633	104.520	114.006	97.851	99.628	107.069	106.830	94.207	92.113
17	200	72.206	106.855	79.678	85.501	44.987	106.827	85.601	121.854	116.428
18	200	109.520	58.349	82.913	111.686	104.000	110.410	107.836	57.091	107.552
19	200	68.268	71.703	51.291	54.032	102.094	101.882	39.651	70.044	137.241
20	200	90.598	98.249	144.722	113.671	106.962	111.144	80.191	132.396	111.132
21	200	85.880	65.584	89.030	94.662	92.735	119.313	133.389	127.446	63.163
22	200	96.712	100.263	121.998	48.267	97.111	110.619	86.179	60.513	112.271
23	200	114.705	89.053	78.936	147.805	92.991	67.371	50.399	129.847	144.385
24	200	115.470	81.581	134.556	130.886	92.066	70.050	67.194	114.451	117.942
25	200	90.017	87.142	76.307	99.310	81.943	102.957	57.147	127.182	57.446
26										
27	N=	20								
28	alfa=	0.05								
29	g.l.=	1	18							
30	tc=	2.101								
31	r2=	0.3405	0.1102	0.1282	0.1610	0.0453	0.3613	0.0158	0.1480	0.1611
32	F=	9.295	2.230	2.647	3.453	0.855	10.183	0.290	3.128	3.457
33	F Sig=	0.007	0.153	0.121	0.080	0.367	0.005	0.597	0.094	0.079
34	F Sig?	Sim	Não	Não	Não	Não	Sim	Não	Não	Não
35	Sig=	593								
36										
37	b1=	40.108	60.913	53.048	50.773	69.946	36.280	67.186	55.870	62.854
38	ep(b1)=	13.456	13.457	21.523	20.247	18.388	15.976	20.937	21.293	18.332
39	LI=	11.838	32.642	7.829	8.236	31.314	2.716	23.199	11.136	24.341
40	LS=	68.379	60.913	53.048	50.773	69.946	36.280	67.186	55.870	62.854
41	beta 1 em IC=	Sim	Sim	Sim	Sim	Sim	Não	Sim	Sim	Sim
42	Sim=	469								
43	valor-p b1=	0.008	0.000	0.024	0.022	0.001	0.036	0.005	0.017	0.003
44	b1 Sig?	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
45	Sig=	741								
46										
47	b2=	0.259	0.127	0.221	0.238	0.108	0.322	0.071	0.238	0.216
48	ep(b2)=	0.085	0.085	0.136	0.128	0.116	0.101	0.132	0.135	0.116
49	LI=	0.081	-0.052	-0.065	-0.031	-0.137	0.110	-0.207	-0.045	-0.028
50	LS=	0.438	0.306	0.507	0.507	0.352	0.535	0.349	0.521	0.459
51	beta 2 em IC=	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
52	Sim=	948								
53	valor-p b2=	0.007	0.153	0.121	0.080	0.367	0.005	0.597	0.094	0.079
54	b2 Sig?	Sim	Não	Não	Não	Não	Sim	Não	Não	Não
55	Sig=	593								

Figura 3. Recorte de tela da planilha com as estimativas e estatísticas das 1.000 regressões simples simuladas

Conforme a Figura 3 mostra, depois de inseridas as fórmulas na planilha, é possível desenvolver uma série de análises a respeito do processo de estimação, no que se refere ao desempenho dos estimadores e dos testes de hipóteses realizados. Por exemplo, pode-se inserir uma modificação na planilha de modo a ampliar o tamanho da amostra para 40, e depois para 80 observações, de modo a analisar como isso afeta o desempenho dos estimadores e dos testes de hipóteses. Também é possível alterar as características do termo de erro, de modo a violar as premissas do modelo clássico de regressão linear, e consequentemente analisar o que ocorrer com os estimadores e os testes quando tais violações estão presentes. Um roteiro detalhado para aplicação da planilha aqui desenvolvida em uma disciplina de econometria básica é apresentado em arquivo suplementar.

3.2 Aplicação 2: estimação dos parâmetros na RLM usando diferentes tamanhos de amostra

Para estender os procedimentos da Aplicação 1 para o caso da regressão linear múltipla, três mudanças são necessárias. Primeiramente, é preciso introduzir uma nova variável X_2 no modelo do processo gerador de dados, tornando $Y=\beta_1+\beta_2 X_1+\beta_3 X_2$. Depois, os valores de X_2 e β_3 precisam ser definidos. E por fim, a fórmula para obtenção de Y precisa ser alterada para tornar a variável efetivamente uma função das duas variáveis, X_1 e X_2. A Tabela 4 apresenta a nova configuração do processo gerador de dados. Os valores escolhidos para os dois níveis de X_2 foram 80 e 160, e β_3 foi definido como 0,35. É importante chamar a atenção para a forma como os pares de valores (x_1,x_2) estão distribuídos, de modo a resultar em completa ausência de correlação entre os dois regressores.

Tabela 5. Configurações do processo gerador de dados para um estudo de Monte Carlo em regressão linear múltipla

	A	B	C	D	E	F
1	N=	20			sigma=	25
2	X11=	100	X21=	80	beta1=	50
3	X12=	200	X22=	160	beta2=	0.25
4					beta3=	0.35
5						
6	x1	x2	y1			
7	100	80	=$F$2+$F$3*$A7+$F$4*$B7+INV.NORM.N(ALEATÓRIO();0;$F$1)	y2	y3	y4
8	100	80				
9	100	80				
10	100	80				
11	100	80				
12	100	160				
13	100	160				
14	100	160				
15	100	160				
16	100	160	...			
17	200	80				
18	200	80				
19	200	80				
20	200	80				
21	200	80				
22	200	160				
23	200	160				
24	200	160				
25	200	160				
26	200	160	==$F$2+$F$3*$A26+$F$4*$B26+INV.NORM.N(ALEATÓRIO();0;$F$1)			

Fonte: Elaborado pelo autor.

A Figura 4 a seguir apresenta os resultados obtidos com a simulação de 1.000 amostras contendo trincas de realizações de x_1, x_2 e y, sendo que os valores de x_1 e x_2 são fixos em todas as amostras. Tais resultados foram obtidos colando as fórmulas do intervalo B7:B26 da planilha para o mesmo intervalo nas colunas C à ALN.

	A	B	C	D	E	F	G	ALJ	ALK	ALL	ALM	ALN
1	N=	20			sigma=	25						
2	X11	100	X21	80	beta1=	50						
3	X12	200	X22	160	beta2=	0.25						
4					beta3=	0.35						
5												
6	x1	x2	y1	y2	y3	y4	y5	y996	y997	y998	y999	y1000
7	100	80	102.788	125.558	98.464	68.212	111.297	109.053	68.250	104.115	113.452	144.022
8	100	80	92.332	129.185	113.236	114.757	90.861	65.501	140.797	129.573	92.996	79.112
9	100	80	109.879	96.406	58.876	96.788	112.215	105.785	78.488	42.200	89.550	92.532
10	100	80	87.185	97.043	127.558	97.776	80.679	143.689	110.133	108.695	116.217	127.407
11	100	80	130.890	147.936	93.321	125.673	142.719	43.471	110.959	101.896	114.719	113.143
12	100	160	113.413	163.984	135.050	127.103	114.029	180.031	155.460	156.441	96.473	133.383
13	100	160	126.998	121.860	86.411	113.216	120.266	163.214	126.982	139.964	145.414	163.552
14	100	160	100.896	144.763	125.710	136.625	124.839	99.160	120.333	129.533	122.701	94.848
15	100	160	136.936	91.965	158.841	158.241	125.513	68.465	130.292	92.437	106.939	141.981
16	100	160	126.196	155.254	150.331	157.895	153.426	130.454	126.223	143.406	124.408	112.526
17	200	80	130.698	114.711	132.407	72.675	126.302	147.980	144.990	98.336	109.167	113.292
18	200	80	115.296	162.544	138.694	142.621	144.826	103.492	114.219	129.631	138.775	127.880
19	200	80	73.366	133.988	136.758	151.399	95.146	97.478	113.780	137.055	102.402	132.470
20	200	80	148.740	151.470	139.433	120.718	132.825	132.947	146.757	146.394	132.508	147.956
21	200	80	111.261	123.514	116.844	170.822	134.605	164.224	171.909	101.903	156.676	133.295
22	200	160	108.092	150.159	143.847	119.433	140.841	184.197	158.449	132.536	139.393	172.734
23	200	160	132.976	126.075	146.905	135.107	165.883	167.751	161.976	194.288	130.830	191.375
24	200	160	111.982	165.071	170.274	128.370	188.924	127.494	161.210	166.106	169.666	185.502
25	200	160	174.912	180.209	188.735	181.165	114.051	111.615	128.185	124.321	148.283	147.992
26	200	160	142.135	154.446	182.289	221.258	135.624	161.944	149.007	149.991	123.568	135.971

Figura 4. Recorte de tela com os parâmetros do modelo e os valores simulados para 1.000 amostras de valores (x_1,x_2,y)

A Tabela 6 a seguir apresenta as fórmulas usadas para obter as informações das regressões múltiplas realizadas com as 1.000 amostras obtidas na simulação. Na Aplicação 2 o foco recai sobre a estimação dos parâmetros β_2 e β_3 da regressão linear múltipla, portanto a análise não inclui as informações relativas à estimação de β_1.

Tabela 6. Fórmulas para obter as estimativas dos parâmetros, estatísticas adicionais e demais informações na regressão múltipla

Célula	Fórmula	Resultado	Copiada para
B31	=INV.T.BC(0.05;D30)	t crítico (95%)	
B32	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);3;1)	R2	C32:ALN32
B33	=1-(((1-C32)*($C$28-1))/$D$30)	R2 ajustado	C33:ALN33
B34	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);4;1)	Estatística F	C34:ALN34
B35	=DIST.F.CD(C34;$C$30;$D$30)	F de significância	C35:ALN35
B36	=SE(OU(C35<0.05);"Sim";"Não")	Sim ou Não	C36:ALN36
B37	=CONT.SE(C36:ALN36;"Sim")	Número de regressões com F significante a 5%	
B39	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);1;2)	Estimativa para beta 2	C39:ALN39
B40	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);2;2)	Erro padrão da estimativa de beta 2	C40:ALM40
B41	=C40-$C$31*C41	Limite inferior (95%) para beta 2	C41:ALN41
B42	=C40+$C$31*C41	Limite superior (95%) para beta 2	C42:ALN42
B43	=SE(OU($F$3<C42;$F$3>C43);"Não";"Sim")	Reporta se o IC contém beta 2	C43:ALN43
B44	=CONT.SE(C44:ALN44;"Sim")	Número de IC que contém beta 2	
B45	=DIST.T.BC(ABS(C40/C41);$D$30)	Valor-p da estimativa de beta 2	C45:ALN45
B46	=SE(OU(B44<0.05);"Sim";"Não")	Se a estimativa beta 2 é significante ou não	C46:ALN46
B47	=CONT.SE(C47:ALN47;"Sim")	Número de regressões que produziram estimativas de beta 2 significantes a 5%	
B49	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);1;1)	Estimativa para beta 3	C49:ALN49
B50	=ÍNDICE(PROJ.LIN(C7:C26;$A$7:$B$26;1;1);2;1)	Erro padrão da estimativa de beta 3	C50:ALN50
B51	=C51-$C$31*C52	Limite inferior (95%) para beta 3	C51:ALN51
B52	=C51+$C$31*C52	Limite superior (95%) para beta 3	C52:ALN52
B53	=SE(OU($F$4<C51;$F$4>C52);"Não";"Sim")	Reporta se o IC contém beta 3	C53:ALN53
B54	=CONT.SE(C53:ALN53;"Sim")	Número de IC que contém beta 3	
B55	=DIST.T.BC(ABS(C49/C50);$D$30)	Valor-p da estimativa de beta 3	C55:ALN55
B56	=SE(OU(C55<0.05);"Sim";"Não")	Se a estimativa de beta 3 é significante ou não	C56:ALN56
B57	=CONT.SE(C56:ALN56;"Sim")	Número de regressões que produziram estimativas de beta 3 significantes a 5%	

Fonte: Elaborado pelo autor.

Deste modo, obtemos uma planilha que contém a simulação de 1.000 amostras contendo trincas de realizações de x_1, x_2 e y, e para cada amostra obtivemos também os resultados e as estatísticas da análise de regressão múltipla por mínimos quadrados ordinários. A Figura 5 apresenta um recorte de tela da planilha obtida.

	A	B	C	D	E	F	G	ALJ	ALK	ALL	ALM	ALN
1	N=	20			sigma=	25						
2	X11	100	X21	80	beta1=	50						
3	X12	200	X22	160	beta2=	0.25						
4					beta3=	0.35						
5												
6	x1	x2	y1	y2	y3	y4	y5	y996	y997	y998	y999	y1000
7	100	80	126.855	87.675	98.297	110.692	131.040	78.555	88.187	104.143	163.161	91.852
8	100	80	122.205	120.892	162.576	111.952	83.617	98.139	115.870	108.551	64.338	105.906
9	100	80	118.466	104.396	110.953	95.814	92.668	85.167	129.346	108.856	102.289	116.346
10	100	80	112.241	127.947	131.890	135.434	99.520	68.608	93.661	114.096	81.430	74.833
11	100	80	80.888	124.996	113.775	111.833	88.553	89.624	63.928	104.968	93.246	84.973
12	100	160	90.869	176.293	187.290	136.961	142.151	106.831	159.636	94.011	124.836	126.469
13	100	160	136.098	139.981	138.078	168.952	119.244	129.050	144.356	128.709	121.312	142.219
14	100	160	188.012	135.982	166.799	128.192	113.342	133.181	159.346	105.114	145.275	149.750
15	100	160	146.348	147.448	174.824	143.810	136.167	161.322	83.239	168.036	146.968	100.394
16	100	160	130.711	110.451	94.628	156.286	156.727	153.032	138.420	132.754	135.175	152.546
17	200	80	137.109	144.524	106.177	149.193	126.124	77.464	123.640	127.064	128.066	139.659
18	200	80	125.115	129.956	124.105	127.403	102.746	125.232	152.520	127.902	73.301	123.779
19	200	80	112.775	170.162	85.438	135.593	138.522	94.832	113.729	138.502	142.031	113.042
20	200	80	165.694	136.400	113.965	106.306	130.185	131.633	128.700	124.884	117.214	136.634
21	200	80	136.838	163.727	155.558	127.231	172.472	86.671	108.545	158.623	158.878	95.247
22	200	160	206.499	164.885	120.149	194.104	137.687	147.330	176.367	145.909	117.134	130.170
23	200	160	152.759	132.352	133.352	143.150	166.632	125.954	153.211	164.749	141.470	177.548
24	200	160	150.786	194.741	190.454	198.496	151.262	113.471	168.785	153.739	159.620	160.644
25	200	160	153.637	184.996	189.461	139.155	153.238	174.942	137.649	132.545	198.633	124.143
26	200	160	138.553	160.380	141.140	180.747	169.561	152.838	161.651	130.777	155.012	142.583
27												
28		N=	20									
29		alfa=	0.05									
30		g.l.=	2	17								
31		tc=	2.110									
32		r2	0.358	0.514	0.271	0.602	0.569	0.616	0.530	0.427	0.348	0.528
33		r2 ajust	0.282	0.457	0.185	0.555	0.518	0.571	0.475	0.359	0.272	0.472
34		F	4.731	9.007	3.153	12.864	11.226	13.661	9.598	6.325	4.545	9.504
35		F Sig	0.023	0.002	0.068	0.000	0.001	0.000	0.002	0.009	0.026	0.002
36		F Sig?	Sim	Sim	Não	Sim	Sim	Sim	Sim	Sim	Sim	Sim
37		Sig=	783									
38												
39		b2=	0.227	0.306	-0.019	0.201	0.285	0.127	0.249	0.235	0.213	0.198
40		ep(b2)=	0.111	0.091	0.133	0.085	0.084	0.092	0.101	0.077	0.128	0.087
41		LI=	-0.008	0.114	-0.300	0.023	0.107	-0.067	0.036	0.073	-0.057	0.014
42		LS=	0.462	0.498	0.261	0.380	0.464	0.320	0.461	0.398	0.483	0.382
43		beta 2 em IC=	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
44		Sim=	945									
45		valor-p b2	0.057	0.004	0.886	0.029	0.004	0.184	0.024	0.007	0.114	0.036
46		b2 Sig?	Não	Sim	Não	Sim	Sim	Não	Sim	Sim	Não	Sim
47		Sig=	561									
48												
49		b3=	0.320	0.296	0.417	0.473	0.351	0.578	0.456	0.173	0.402	0.405
50		ep(b3)=	0.139	0.114	0.166	0.106	0.106	0.115	0.126	0.096	0.160	0.109
51		LI=	0.027	0.056	0.066	0.250	0.128	0.336	0.190	-0.029	0.064	0.175
52		LS=	0.614	0.537	0.768	0.696	0.573	0.819	0.721	0.376	0.739	0.635
53		beta 3 em IC=	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Sim
54		Sim=	951									
55		valor-p b3	0.034	0.019	0.023	0.000	0.004	0.000	0.002	0.089	0.022	0.002
56		b3 Sig?	Sim	Sim	Sim	Sim	Sim	Sim	Sim	Não	Sim	Sim
57		Sig=	662									

Figura 5. Recorte de tela da planilha com as estimativas e estatísticas das 1.000 regressões múltiplas simuladas

Depois que a planilha é construída, é possível desenvolver uma série de análises a respeito do processo de estimação no contexto de regressão linear múltipla, no que se refere ao desempenho dos estimadores e dos testes de hipóteses quando ocorrem violações das premissas do modelo clássico de regressão linear. Por exemplo, é possível estimar os parâmetros de uma regressão de Y apenas como função de X_1, de modo a analisar o viés de omissão de variável. Também é possível introduzir correlações diferentes de zero entre os regressores, de modo a analisar como tal situação impacta no processo de estimação e nos testes de hipóteses. O arquivo suplementar contém o detalhamento dos procedimentos que podem ser utilizados para aplicação do caso como ferramenta pedagógica.

4 CONCLUSÕES

Contribuir para que os estudantes de econometria básica compreendam o conceito de distribuição amostral é um dos grandes desafios que os professores da disciplina enfrentam. A ampla maioria dos livros disponíveis para serem adotados como livro-texto em um curso introdutório em econometria trata do assunto de forma extensamente matematizada e abstrata. Apesar de vários professores-autores terem chamado a atenção para o potencial das ferramentas de simulação no ensino de econometria, seu uso ainda pode ser considerado pouco frequente.

A presente ferramenta pedagógica compreende a utilização da simulação de Monte Carlo como meio para contribuir para o aprendizado do conceito de distribuição amostral, em um contexto de estimação de parâmetros no modelo clássico de regressão linear simples e múltipla. As atividades demonstram como um software comercial de planilha eletrônica pode ser utilizado para produzir 1.000 simulações de um processo gerador de dados, representando o que seria um sorteio de 1.000 amostras com as variáveis usadas na regressão. As subsequentes estimações são obtidas a partir das amostras, e o desempenho dos estimadores e dos testes podem ser analisados, de modo a ilustrar o conceito de distribuição amostral.

O roteiro de aplicação da ferramenta, disponível em arquivo suplementar, demostra o seu potencial de aplicação como técnica de ensino do conceito de distribuição amostral de estimadores no contexto do modelo clássico de regressão linear. Por meio da aplicação das atividades propostas, os estudantes podem construir tabelas e visualizar como o desempenho dos estimadores pontuais e intervalares, e dos testes F e t, é afetado quando ocorrem violações nas premissas do modelo clássico de regressão, como heterocedasticidade, multicolinearidade entre regressores e viés de omissão de variáveis.

Deste modo, as atividades aqui apresentadas, quando aplicados em um curso de econometria de forma complementar à apresentação das definições matemáticas do conceito de distribuição amostral, tem potencial para ampliar o entendimento dos estudantes, pois fortalece a conexão entre teoria e prática. Conceitos como erro tipo I e erro tipo II são facilmente ilustrados com as atividades desenvolvidas neste caso de ensino. A simulação desenvolvida aqui é especialmente útil para viabilizar o entendimento da probabilidade do erro tipo II, que depende do verdadeiro valor do parâmetro que está sendo estimado.

REFERÊNCIAS

Barreto, H., & Howland, F. (2005). *Introductory econometrics: using Monte Carlo simulation with Microsoft excel*. Cambridge University Press.

Becker, W. E., & Greene, W. H. (2001). Teaching statistics and econometrics to undergraduates. *The Journal of Economic Perspectives*, 15(4), 169-182. DOI: https://doi.org/110.1257/jep.15.4.169.

Bekkerman, A. (2015). The role of simulations in econometrics pedagogy. *Wiley Interdisciplinary Reviews: Computational Statistics*, 7(2), 160-165. DOI: https://doi.org/10.1002/wics.1342

Briand, G., & Hill, R. C. (2013). Teaching basic econometric concepts using Monte Carlo simulations in Excel. *International Review of Economics Education*, 12, 60-79. DOI: https://doi.org/10.1016/j.iree.2013.04.001

Chance, B., Garfield, J., & del Mas, B. (1999). A model of classroom research in action: Developing simulation activities to improve students' statistical reasoning. *52nd Session of the International Statistical Institute*, Helsinki, Finland.

Chance, B., del Mas, R., & Garfield, J. (2004). *Reasoning about sampling distribitions*. In: Ben-Zvi, D. & Garfield, J., Eds. The challenge of developing statistical literacy, reasoning and thinking, pp. 295-323. Dordrecht, The Netherlands: Kluwer Academic Publishers.

Craft, R. K. (2003). Using spreadsheets to conduct Monte Carlo experiments for teaching introductory econometrics. *Southern Economic Journal*, 726-735. DOI: https://doi.org/10.2307/1061705.

Dyck, J. L., & Gee, N. R. (1998). A sweet way to teach students about the sampling distribution of the mean. *Teaching of Psychology*, 25(3), 192-195.DOI: https://doi.org/10.1207/s15328023top2503_6.

Judge, G. (1999). Simple Monte Carlo studies on a spreadsheet. *Computers in Higher Education Economics Review,* 13(2), 12-14.

Kennedy, P. E. (1998). Teaching undergraduate econometrics: A suggestion for fundamental change. *The American Economic Review*, 88(2), 487-492.

Kennedy, P. (2003). *A guide to econometrics*. MIT press.

Como citar este artigo

Pagliarussi, M. S. (2018). O ensino do modelo clássico de regressão linear por meio de simulação de Monte Carlo. *Revista de Contabilidade e Organizações*, 12:e152100. DOI: http://dx.doi.org/10.11606/issn.1982-6486.rco.2018.152100

ROTEIRO DE APLICAÇÃO DAS ATIVIDADES DE ENSINO DO MODELO CLÁSSICO DE REGRESSÃO LINAR POR MEIO DE SIMULAÇÃO DE MONTE CARLO

1 INTRODUÇÃO

Sugiro que a aplicação da atividade de simulação de Monte Carlo ocorra depois que os conteúdos relacionados a estimação, inferência estatística e análise de regressão linear simples e múltipla tenham sido trabalhados com os alunos. Deste modo, a base teórica estará pronta para ser acessada pelos alunos, e as atividades de simulação desempenharão um papel ilustrativo dos conceitos de forma mais efetiva.

Recomendo também que, antes da apresentação dos exercícios aqui apresentados, seja solicitado aos alunos a leitura da seção 2.10 do livro "*A guide to econometrics*", de Peter Kennedy (Kennedy, 1998). A apresentação que o autor faz da estrutura de um estudo de Monte Carlo é exatamente a que foi utilizada para embasar as atividades aqui apresentadas.

O professor pode começar a aplicação da simulação de Monte Carlo apresentando aos alunos as situações problema que deverão ser investigadas. Uma possibilidade é apresentar aos alunos a tarefa de buscar respostas para seis perguntas, por meio do desenvolvimento de uma série de estudos de Monte Carlo. São elas:

1. O que ocorre com a distribuição de amostragem dos estimadores β^{MQO} quando usamos amostras de tamanho 20, 40 e 80 em uma regressão linear simples?

2. Como a distribuição de amostragem do estimador β_2^{MQO} é afetada pela violação da premissa de homocedasticidade?

3. Como as distribuições de amostragem dos estimadores β^{MQO} são afetadas quando há violação da premissa de que os erros têm média zero?

4. Como as distribuições de amostragem dos estimadores β^{MQO} são afetadas quando há violação da premissa de que os erros seguem distribuição normal?

5. Como a distribuição de amostragem do estimador β_2^{MQO} é afetada quando há um viés de omissão de variável?

6. Como as distribuições de amostragem dos estimadores β_2^{MQO} e β_3^{MQO} são afetadas quando há violação de premissa de ausência de relações lineares exatas entre os regressores?

A busca das respostas para as perguntas deve se basear no processo sugerido por Kennedy (2003) para a realização de estudos de Monte Carlo. Deste modo, o professor pode resgatar com os alunos as etapas do processo, as quais são:

a. Modelar o processo gerador de dados

b. Gerar M conjuntos de dados

c. Calcular os M valores de $\hat{\beta}$

d. Obter a distribuição amostral de $\hat{\beta}$

e. Analisar as propriedades da distribuição amostral de $\hat{\beta}$

Já conhecendo o processo para construção da planilha de simulação, conforme descrito nas seções 3.1 e 3.2 do artigo, o professor pode conduzir os alunos na tarefa de, passo a passo, desenvolver as etapas do estudo de Monte Carlo. Por exemplo, o processo deve se iniciar com a definição dos valores dos parâmetros. Recomendo que o professor conduza a atividade de modo que o valor selecionado para β_2 não seja muito elevado, preferencialmente entre 0 e 0,5, para permitir a visualização dos problemas relacionados aos testes de hipóteses, notadamente o erro tipo I e erro tipo II.

Posteriormente, deverão ser escolhidos os valores de *X*. Mais uma vez, o ideal é direcionar a escolha para apenas dois níveis de valores de *X*, para facilitar a análise das violações das premissas do modelo clássico de regressão, principalmente a de homocedasticidade e a de ausência de colinearidade entre os regressores. Uma justificativa que pode ser oferecida aos alunos para a adoção de apenas dois níveis de *X* é que, no contexto de causalidade, os dois níveis podem representar a ausência ou presença do tratamento.

Definidos os valores dos parâmetros e da variável independente *X*, a próxima etapa é construir uma formula que represente o processo gerador de *Y*. Neste momento, é necessário ressaltar o papel do termo de erro como responsável pela natureza do processo estocástico, e a função ALEATÓRIO do Excel deve surgir naturalmente na discussão. Construída a primeira realização dos valores de *Y*, a obtenção das demais 999 amostras deve ressaltar a questão dos valores de *X* serem fixos nas amostras repetidas. Ao final do processo, os alunos terão desenvolvido no computador a planilha contendo as 1.000 amostras contendo pares de dados (*x,y*).

A partir do momento em que as 1.000 amostras foram obtidas, o professor pode apresentar aos alunos as funções PROJ.LIN e ÍNDICE do Excel, de modo a obter as estimativas para os parâmetros e as estatísticas adicionais da regressão.

2 PROCESSO DE CONSTRUÇÃO DAS RESPOSTAS DAS PERGUNTAS DO TRABALHO

A primeira pergunta, "*O que ocorre com a distribuição de amostragem dos estimadores* β^{MQO} *quando usamos amostras de tamanho 20, 40 e 80 em uma regressão linear simples?*", envolve a análise do valor médio e do erro padrão dos estimadores $\hat{\beta}_1$ e $\hat{\beta}_2$. Com base nas informações obtidas por meio da aplicação das fórmulas da Tabela 4 do caso de ensino, os alunos podem construir uma tabela semelhante à Tabela 1, a qual apresenta os resultados obtidos para a simulação de 1.000 regressões simples para três tamanhos de amostra.

Tabela 1. Resultados da realização de 1.000 análises de regressão linear simples com três tamanhos de amostra diferentes

	n= 20	**n= 40**	**n= 80**
F sig	560	864	990
E(b1)	51,583	48,584	49,096
ep(b1)	17,817	11,653	9,140
E(b2)	0,237	0,259	0,254
ep(b2)	0,113	0,072	0,058
b1 no IC	488	943	943
b2 no IC	955	948	944
b1 sig	758	970	1000
b2 sig	560	864	990

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+u$: β_1=50; β_2=0,25

Deste modo, os alunos podem analisar o que ocorre com o valor esperado e com o erro padrão das estimativas conforme aumenta o tamanho da amostra, de modo a observar a tendência de aproximação do valor estimado para o verdadeiro valor do parâmetro, com variância decrescente conforme aumenta o tamanho da amostra. Tal tendência pode ser observada graficamente por meio da construção de um histograma a partir dos valores estimados para $\hat{\beta}_2$, para os três tamanhos de amostra, conforme mostra a Figura 1.

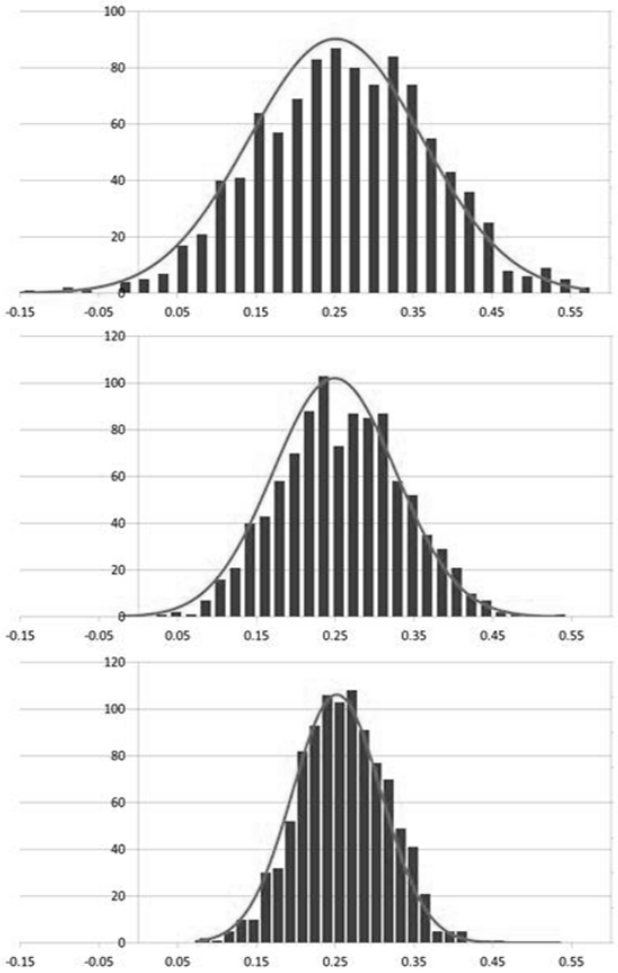

Figura 5. Histograma representativo da distribuição amostral de β_2^{MQO} para diferentes tamanhos de amostra

Adicionalmente, a Tabela 1 permite observar que os testes *F* e *t* não são capazes de detectar a existência de relação linear entre *Y* e *X* em 44% dos casos quando n=20, conforme mostram a primeira e a última linha da Tabela 1. Por outro lado, a estimativa por intervalo de confiança para β_2 apresenta resultados consistentes com o nível de confiança estabelecido, conforme mostra a antepenúltima linha da Tabela 1. Deste modo, a análise se estende para além das propriedades da distribuição amostral dos estimadores, e passa a incluir também a questão do desempenho dos testes de hipóteses no que se refere ao erro tipo II.

Para o caso da regressão múltipla, primeiramente é necessário que os alunos pensem nas mudanças que são necessárias ao processo gerador de dados para permitir tanto a obtenção das amostras tal que $Y=\beta_1+\beta_2 X_1+\beta_3 X_2+u$ como também no processo de estimação dos parâmetros $\hat{\beta}_2$ e $\hat{\beta}_3$. A seção 3.2 do artigo detalha a construção da planilha para realização da simulação das 1.000 amostras que serão utilizadas na regressão múltipla. O primeiro passo é escolher os valores de X_2 e β_3. No que se refere à distribuição dos valores de os valores de X_2 na planilha, o professor pode deixar os alunos fazerem isso livremente. É comum que os alunos distribuam os valores de X_1 e X_2 de modo a resultar em uma correlação perfeita entre as variáveis, de modo que o Excel irá retornar o valor zero para a estimativa de um dos betas. Tal situação representa uma oportunidade interessante para iniciar a discussão do problema da multicolinearidade, mesmo que esta violação seja explorada mais adiante apenas.

Conforme os alunos construírem o entendimento de que X_1 e X_2 devem ser não correlacionados, o trabalho prossegue conforme descrito na seção 3.2 do artigo, poderão ser construídas tabelas como a Tabela 2, que apresenta os resultados do processo de estimação da regressão múltipla (Aplicação 2 do artigo).

Tabela 2. Resultados da realização de 1.000 análises de regressão linear múltipla, para três tamanhos de amostra diferentes e sem correlação entre os regressores

	n= 20	**n= 40**	**n= 80**
	$\rho_{X1\,X2}$**= 0**	$\rho_{X1\,X2}$**= 0**	$\rho_{X1\,X2}$**= 0**
F sig	788	996	1000
E(b2)	0,263	0,248	0,246
ep(b2)	0,091	0,072	0,058
E(b3)	0,356	0,345	0,352
ep(b3)	0,145	0,101	0,065
b2 no IC	951	960	953
b3 no IC	954	958	951
b2 sig	587	867	993
b3 sig	639	948	998
r2 ajust	0,356	0,357	0,360

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+ \beta_3 X_2 +u$: β_1=50; β_2=0,25; β_3=0,35

Conforme se observa na primeira coluna da Tabela 2, o uso de amostras pequenas (n=20) prejudica sensivelmente a potência dos testes de hipóteses, pois em apenas 58,7% dos casos rejeita-se a hipótese nula de que β_2 é diferente de zero. Para o caso de β_3, em apenas 63,9% dos casos o teste *t* rejeitou a hipótese nula de que o valor do parâmetro é zero. Os resultados dos testes melhoram significativamente conforme o tamanho da amostra aumenta para 40 e depois para 80. De modo semelhante à Tabela 1, ao observar o valor esperado e o erro padrão das estimativas dos parâmetros é possível observar que o valor esperado se aproxima cada vez mais dos valores verdadeiros 0,25 e 0,35.

Para responder à segunda pergunta, "*Como a distribuição de amostragem do estimador* β_2^{MQO} *é afetada pela violação da premissa de homocedasticidade?*" é importante fazer os alunos refletirem a respeito de quais alterações serão necessárias no processo gerador de dados, notadamente no termo de erro. A sugestão é que o professor peça aos alunos sugestões a respeito de como a variância do termo de erro pode tornar-se heterocedástica.

Uma sugestão apresentada aqui é fazer com que o desvio padrão do erro quando *x*=200 seja duas vezes maior que quando *x*=100, na planilha de regressão simples. A análise continua sendo desenvolvida para três tamanhos de amostra, n=20, n=40 e n=80. Assim, de posse das novas estimações e resultados dos testes de hipóteses, pode ser construída uma tabela semelhante à Tabela 3.

Tabela 3. Resultados da realização de 1.000 análises de regressão linear simples, em condições de erros homocedásticos e heterocedásticos

	Erros homocedásticos			**Erros heterocedásticos**		
	n= 20	**n= 40**	**n= 80**	**n= 20**	**n= 40**	**n= 80**
F sig	560	864	990	262	503	794
E(b1)	51,583	48,584	49,096	50,746	49,713	50,064
ep(b1)	17,817	11,653	9,140	21,786	16,133	10,570
E(b2)	0,237	0,259	0,254	0,241	0,252	0,249
ep(b2)	0,113	0,072	0,058	0,173	0,127	0,085
b1 no IC	488	943	943	512	985	986
b2 no IC	955	948	944	955	952	955
b1 sig	758	970	1000	389	705	978
b2 sig	560	864	990	262	503	794

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+ u$: β_1=50; β_2=0,25

Ao analisar as linhas E(b2) e ep(b2) da Tabela 3, os alunos poderão observar que, conforme alguns livros apontam (por exemplo, Wooldridge, 2009), a heterocedasticidade não causa viés ou inconsistência nos estimadores de mínimos quadrados. Entretanto, o comportamento do erro padrão do estimador β_2^{MQO} (linha ep(b2) na Tabela 3) mostra que, na presença de heterocedasticidade, a variância do estimador passa a ser viesada, o que impacta na construção dos intervalos de confiança e na obtenção das estatísticas de teste. A estatística *t* usualmente empregada para o teste de hipótese dos parâmetros, que é baseada no erro padrão dos estimadores, não irá mais seguir a distribuição *t* na presença de heterocedasticidade, e este problema não é resolvido aumentando o tamanho da amostra (Wooldridge, 2009). De modo similar, a estatística *F* da regressão não seguirá a distribuição *F*. Em síntese, as estatísticas usadas para os testes de hipóteses não são válidas na presença de heterocedasticidade. A primeira e a última linha da Tabela 3 ilustram esta situação, pois os testes de hipóteses perdem confiabilidade.

Para responder à questão 3, "*Como as distribuições de amostragem dos estimadores* β^{MQO} *são afetadas quando há violação da premissa de que os erros têm média zero?*" também será necessário alterar os parâmetros do termo de erro na função geradora dos valores de *Y*. O último termo da função descrita na Tabela 2 do artigo pode passar a ser INV.NORM.N(ALEATÓRIO();5;$D$1) ao invés de INV.NORM.N(ALEATÓRIO();0;$D$1). Posteriormente, os alunos podem construir uma tabela semelhante à Tabela 1 aqui apresentada, e ficará evidente que quando o termo de erro tem média diferente de zero, apenas a estimação do parâmetro β_1 é prejudicada.

Para investigar questão quatro, "*Como as distribuições de amostragem dos estimadores* β^{MQO} *são afetadas quando há violação da premissa de que os erros seguem distribuição normal?*", o professor pode questionar aos alunos qual modificação deve ser feita no processo gerador de dados que possibilite a investigação. Como sugestão, recomendo modelar os erros como sendo extraídos de uma distribuição uniforme em [*a,b*]. No caso, os valores de *a* e *b* devem ser simétricos, de modo a resultar em um termo de erro com média zero, e também devem ser escolhidos de modo a produzir um termo de erro com variância não muito diferente de 625, de modo a garantir a comparabilidade.

No caso da adoção de uma termo de erro extraído de uma distribuição uniforme, o último termo da função geradora de *Y* seria *a+(b-a)**ALEATÓRIO() ao invés de INV.NORM.N(ALEATÓRIO();0;$D$1). Por exemplo, para fazer o termo de erro ser extraído de uma distribuição uniforme no intervalo [-50,50] a função geradora de *Y* descrita na Tabela 2 do artigo de ensino teria a fórmula $D$2+$D$3*A6+(-50)+(100)*ALEATÓRIO().

Conforme Wooldridgre (2009) destaca, a normalidade dos estimadores de mínimos quadrados ordinários depende crucialemente da normalidade da distribuição do termo de erro *u*. Se os erros para cada observação são extraídos a partir de outra distribuição que não a normal, as estatísticas de teste *t* e *F* não seguirão as respectivas distribuições de probabilidade. Este problema pode afetar seriamente os resultados dos valores-p que são obtidos usando tais distribuições (Wooldridge, 2009). A Tabela 4 exibe as informações obtidas com a estimação dos parâmetros β_1 e β_2 em situações em que os erros são normais (três primeiras colunas) e não normais (três ultimas colunas), para diferentes tamanhos de amostra.

Tabela 4. Resultados da realização de 1.000 análises de regressão linear simples, em condições de erros normais e não normais

	Erros normais			**Erros não normais**		
	n= 20	**n= 40**	**n= 80**	**n= 20**	**n= 40**	**n= 80**
F sig	575	861	990	429	759	972
E(b1)	50,589	50,199	50,686	49,677	49,714	50,542
ep(b1)	17,584	12,800	8,943	20,836	14,137	9,912
E(b2)	0,244	0,253	0,249	0,252	0,251	0,248
ep(b2)	0,112	0,082	0,057	0,131	0,088	0,063
b1 no IC	459	951	953	468	955	958
b2 no IC	949	946	947	950	964	952
b1 sig	744	971	1000	926	926	1000
b2 sig	575	861	990	429	759	972

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+ u$: β_1=50; β_2=0,25

Conforme a Tabela 4 permite observar, o estimador de β_2 continua não viesado e consistente (ver as linhas E(b2) e ep(b2)). Já a última linha permite observar que o desempenho dos testes de hipótese para β_2 é pior quando os erros são não normais e as amostras pequenas, mas o problema praticamente deixa de existir quando o tamanho da amostra é 80. Tal situação pode ser usada pelo professor para ilustrar a operação do Teorema do Limite Central.

A questão 5 solicita aos alunos a análise do efeito do viés de omissão de variáveis no processo de estimação. É importante conduzir o estudo da questão 5 em duas circunstâncias: (1) a correlação entre X_1 e X_2 é zero e (2) quando a correlação entre X_1 e X_2 é diferente de zero (no caso, foi escolhido o valor 0,4). As mudanças que devem ser introduzidas na planilha nesta etapa possivelmente são mais sofisticadas e mais difíceis para os alunos alcançarem sozinhos, uma vez que é necessário gerar os valores de Y em função de X_1 e X_2, mas deve-se estimar os parâmetros de um modelo de regressão simples, $Y=\beta_1+\beta_2 X_1+ u$.

Wooldridge (2009) argumenta que a omissão de uma variável relevante que seja correlacionada a qualquer um dos regressores faz com que seja violada a premissa de que o erro não tem correlação com nenhum dos regressores, e introduz um viés na estimação dos parâmetros. Entretanto, se a variável relevante omitida não for correlacionada a nenhum regressor no modelo, os estimadores de mínimos quadrados permanecem não enviesados (Wooldridge, 2009).

Deste modo, para efeito de aprendizagem, sugiro que a análise da omissão de variáveis no processo de estimação seja desenvolvida da seguinte maneira: (1) os valores de Y continuam sendo gerados pela função $\beta_1+\beta_2 X_1+\beta_3 X_2 +u$; (2) os valores de X_1 e X_2 na amostra continuam sendo x_{11}=100, x_{12}=200, x_{21}=80 e x_{22}=160; (3) uma planilha será construída com os valores de X_1 e X_2 sendo distribuídos de tal forma que a correlação entre as variáveis seja zero e outra planilha será construída com os valores de X_1 e X_2 sendo distribuídos de tal forma que a correlação entre as variáveis seja 0,4; (4) a estimação se dará, nas duas planilhas, por meio do modelo incompleto $Y=\beta_1+\beta_2 X_1+ u$.

A partir dos resultados das estimações, pode ser construída uma tabela semelhante à Tabela 5 a seguir, que exibe o resultado da estimação dos parâmetros do modelo incompleto $Y=\beta_1+\beta_2 X_1+ u$, o qual omite a variável relevante X_2.

Tabela 5. Resultados da realização de 1.000 análises de regressão linear do modelo incompleto $Y=\beta_1+\beta_2 X_1+ u$ para diferentes tamanhos de amostra e correlações entre X_1 e a variável omitida X_2

	$\rho_{X1\,X2}$= 0			$\rho_{X1\,X2}$= 0,4		
	n= 20	**n= 40**	**n= 80**	**n= 20**	**n= 40**	**n= 80**
F sig	422	791	984	793	989	1000
E(b1)	90,755	90,607	92,294	74,811	75,938	75,328
ep(b1)	17,264	12,571	8,324	16,965	14,010	9,488
E(b2)	0,254	0,259	0,248	0,363	0,351	0,359
ep(b2)	0,114	0,078	0,054	0,106	0,089	0,062
b1 no IC	509	148	7	817	632	255
b2 no IC	976	973	980	914	804	599
b1 sig	992	1000	1000	955	999	1000
b2 sig	422	791	984	793	989	1000
r2	0,187	0,175	0,166	0,309	0,307	0,297

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+ \beta_3 X_2 +u$: β_1=50; β_2=0,25; β_3=0,35 (omitido)

Por fim, a sexta questão, "*O que ocorre com a distribuição de amostragem do estimador β^{MQO} quando há violação de premissa de ausência de relações lineares exatas entre os regressores?*", foi elaborada para permitir aos alunos compreenderem na prática o impacto da multicolinearidade na análise de regressão múltipla. Wooldridge (2009) afirma que a existência de relação linear forte entre X_1 e X_2 pode levar a um grande aumento na variância dos estimadores de inclinação $\hat{\beta}_i^{MQO}$. Gujarati e Porter (2009) demonstram que a existência de relação linear forte entre X_1 e X_2 torna difícil avaliar o efeito de X_1 em Y, mantendo X_2 constante, e vice versa. Os autores também afirmam que a multicolinearidade reduz a precisão e a exatidão da estimativa dos coeficientes de inclinação, pois seu erro padrão aumenta muito em comparação ao próprio valor dos coeficientes.

Na aplicação das atividades de ensino desenvolvidas aqui, sugere-se que as análises de regressão múltipla sejam realizadas em três condições: $\rho_{X1\,X2}$= 0; $\rho_{X1\,X2}$= 0,4 e $\rho_{X1\,X2}$= 0,8. As alterações necessárias na planilha envolvem as distribuições dos valores de X_1 e X_2 nas amostras. Por exemplo, se X_1 e X_2 assumem dois valores, x_{11}=100, x_{12}=200, x_{21}=80 e x_{22}=160. A Tabela 6 a seguir ilustra as possíveis distribuições de 20 realizações de X_1 e X_2 para três níveis de correlação diferentes.

Tabela 6. Distribuições possíveis para X_1 e X_2 com diferentes níveis de correlação

$\rho_{X1\,X2}$= 0		$\rho_{X1\,X2}$= 0,4		$\rho_{X1\,X2}$= 0,8	
X_1	X_2	X_1	X_2	X_1	X_2
100	80	100	80	100	80
100	80	100	80	100	80
100	80	100	80	100	80
100	80	100	80	100	80
100	80	100	80	100	80
100	160	100	80	100	80
100	160	100	80	100	80
100	160	100	160	100	80
100	160	100	160	100	80
100	160	100	160	100	160
200	80	200	80	200	80
200	80	200	80	200	160
200	80	200	80	200	160
200	80	200	160	200	160
200	80	200	160	200	160
200	160	200	160	200	160
200	160	200	160	200	160
200	160	200	160	200	160
200	160	200	160	200	160
200	160	200	160	200	160

Fonte: Elaborado pelo autor.

A partir das distribuições apresentadas na Tabela 6, e estendendo-as para tamanhos de amostra iguais a 40 a 80 observações, o processo de análise é semelhante ao realizado anteriormente. Obtém-se as 1.000 amostras para cada condição, são estimados os parâmetros e obtidos as estatísticas de regressão e os resultados dos testes de hipóteses. Com tais resultados em mãos, pode ser construída uma tabela semelhante à Tabela 7, sintetiza os resultados obtidos.

Tabela 7. Resultados da realização de 1.000 análises de regressão linear múltipla para diferentes tamanhos de amostra e correlações entre X_1 e X_2

	$\rho_{X1\,X2}$= 0			$\rho_{X1\,X2}$= 0,4			$\rho_{X1\,X2}$= 0,8		
	n= 20	**n= 40**	**n= 80**	**n= 20**	**n= 40**	**n= 80**	**n= 20**	**n= 40**	**n= 80**
F sig	788	996	1000	899	999	1000	971	1000	1000
E(b2)	0,263	0,248	0,236	0,233	0,256	0,254	0,258	0,251	0,265
ep(b2)	0,091	0,072	0,058	0,120	0,080	0,061	0,176	0,126	0,078
E(b3)	0,356	0,345	0,362	0,355	0,360	0,353	0,342	0,355	0,343
ep(b3)	0,145	0,101	0,065	0,153	0,122	0,073	0,217	0,165	0,102
b2 no IC	951	960	953	960	958	949	948	944	947
b3 no IC	954	958	951	953	952	957	937	945	949
b2 sig	587	867	993	499	810	982	251	433	762
b3 sig	639	948	998	590	892	997	283	550	827
r2 ajust	0,356	0,357	0,360	0,436	0,437	0,439	0,504	0,503	0,505

Fonte: Elaborado pelo autor.
Nota: Parâmetros do modelo $Y=\beta_1+\beta_2 X_1+ \beta_3 X_2 +u$: β_1=50; β_2=0,25; β_3=0,35

A análise da Tabela 7 permite observar uma série de efeitos derivados da multicolinearidade. Por exemplo, o erro padrão dos estimadores aumenta com a correlação entre X_1 e X_2. As linhas b2 sig e b3 sig na tabela mostra que probabilidade de erro tipo II no teste de hipóteses relativo as coeficientes aumenta com a correlação entre X_1 e X_2. Quando n=20 e a correlação entre as variáveis é 0,8, em menos de 30% das amostras o teste rejeitou a hipótese nula de que os coeficientes são iguais a zero. Ao mesmo tempo, o teste *F* indica em 97,1% das amostras que pelo menos um dos coeficientes é diferente de zero, e o R^2 ajustado médio foi de 0,504, bem maior que o valor 0,356 observado no caso em que n=20 e a correlação entre as variáveis é 0.

Encerra-se aqui o roteiro de aplicação do artigo "*O ensino do modelo clássico de regressão linear por meio de simulação de Monte Carlo*". As seis questões apresentadas no presente roteiro e o detalhamento do uso da simulação de Monte Carlo para buscar respostas às questões demonstram o potencial da ferramenta para o ensino de Econometria Básica.

REFERÊNCIAS

Gujarati, D. N., & Porter, D. (2009). *Basic Econometrics*. McGraw-Hill International Edition.

Wooldridge, J. M. (2009) *Introductory Econometrics: A modern approach*. Canada: South-Western Cengage Learning.