Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística
Programa de Especialização em *Data Science* e *Big Data*

Rafael Afonso Monastier

Estimação de lucros operacionais correntes de empresas de investimento direto no Brasil

Curitiba
2022

Rafael Afonso Monastier

Estimação de lucros operacionais correntes de empresas de investimento direto no Brasil

Monografia apresentada ao Programa de Especialização em Data Science e Big Data da Setor de Ciências Exatas, Universidade Federal do Paraná como requisito parcial para a obtenção do grau de especialista.
Orientador: Wagner Hugo Bonat

Curitiba
2022

Especialização em Data Science e Big Data
Universidade Federal do Paraná
dsbd.leg.ufpr.br

Estimação de lucros operacionais correntes de empresas de investimento direto no Brasil

Rafael Afonso Monastier[1]
Wagner Hugo Bonat[2]

Resumo

Este trabalho descreve o ajuste e avaliação de um modelo de regressão linear múltipla para estimar lucros operacionais correntes de empresas de investimento direto no Brasil. Usou-se um painel de 877 empresas observadas ao longo de 20 trimestres. Como covariáveis, foram utilizadas o ativo contábil das empresas (uma proxy de tamanho), o setor de atividade econômica, a arrecadação de impostos federais, por setor, no trimestre (uma proxy das condições econômicas), além da própria resposta defasada. O modelo construído teve R^2 ajustado de 0,71, sem alavancagem, multicolinearidade ou autocorrelação relevantes. No entanto, os resíduos apresentaram distribuição não gaussiana, com caudas espessas. Uma estimação alternativa usando regressão robusta sugeriu que essa não normalidade não fez com que termos do modelo fossem indevidamente considerados significativos. Outra estimação alternativa, com distribuição t de 0,83 grau de liberdade para os resíduos, teve maior log-verossimilhança do que a estimação que assumia normalidade, sendo, portanto, mais confiável para fornecer intervalos de predição para a variável resposta – embora não necessariamente melhor para estimativas pontuais da média. Apesar dessa questão, a maior qualidade do modelo foi sua capacidade de acomodar empresas de diferentes tamanhos, com ativos que variam entre R$ 250 milhões a mais de R$ 100 bilhões.

Palavras-chave: Investimento direto, lucros, regressão linear múltipla, regressão robusta, distribuição t.

Abstract

This paper describes the fitting and evaluation of a multiple linear regression model that estimates current operating earnings of direct investment enterprises in Brazil. For that, a panel was used, covering 877 enterprises observed over 20 quarters. Covariables were total assets (a proxy for the size of the company), industry, federal taxes collected per industry and quarter (a proxy for economic conditions), as well as the lagged response variable. The fitted model showed an adjusted R^2 of 0,71, with no relevant leverage, multicollinearity or autocorrelation. Residuals, however, had a non-normal distribution, with thick tails. An alternative estimation with robust regression suggested that this non-normality did not lead to any terms of the model being incorrectly regarded as significant. Another alternative estimation, which fitted residuals to a t distribution with 0.83 degrees of freedom, had higher log-likelihood than the estimation that assumed normality. It was therefore more accurate in providing prediction intervals for the response, even though not necessarily for point estimates of the mean. Despite that, the greatest strength of model was its capacity to accommodate enterprises of different sizes, with assets ranging from R$ 250 million to more than R$ 100 billion.

Keywords: *Direct investment, earnings, multiple linear regression, robust regression, t distribution.*

I Introdução

O objetivo deste trabalho é desenvolver um modelo estatístico para estimar lucros operacionais correntes trimestrais de empresas de investimento direto residentes no Brasil, com base em um conjunto de dados históricos relativos à variável resposta, a características da empresa e ao estado da economia.

O investimento direto é uma das chamadas "categorias funcionais" do balanço de pagamentos (BP) e da posição de investimento internacional (PII), dois conjuntos de estatísticas macroeconômicas agregadas que apresentam, respectivamente, as transações e as posições financeiras de uma economia (em geral um país) com o resto do mundo. Os conceitos associados a esses dois conjuntos são definidos e detalhados em um

[1] Aluno do programa de Especialização em Data Science & Big Data e analista do Banco Central do Brasil, rafael.monastier@bcb.gov.br. As opiniões expressas neste trabalho são exclusivamente dos autores e não refletem, necessariamente, a visão do Banco Central do Brasil. Apenas este autor teve acesso a dados individuais de empresas reportados ao Banco Central do Brasil e protegidos por sigilo bancário e estatístico.

[2] Professor do Departamento de Estatística - DEST/UFPR, wbonat@ufpr.br.

manual para compiladores de estatísticas macroeconômicas de setor externo, o "*Balance of Payments and International Investment Position Manual, Sixth Edition*" (FMI [1]). Segundo o manual, empresas de investimento direto residentes em uma economia são aquelas que tem em seu quadro de sócios ao menos um investidor não residente que possui individualmente 10% ou mais do capital com direito a voto da empresa[3]. Ainda segundo o manual, quando há uma relação de investimento direto, devem ser reportadas no BP as transações envolvendo o capital (na forma de ações ou quotas), bem como os lucros da empresa investida. Os lucros reportados devem ser os totais, compostos pela soma dos lucros distribuídos aos investidores diretos, em geral sob a forma de dividendos, e os lucros reinvestidos na empresa investida[4]. A racionalidade para a inclusão dos lucros reinvestidos se encontra no fato de que, como o critério de 10% indica a existência de influência ou controle do investidor direto sobre a administração da empresa, então tal investidor tem participação sobre a decisão de remeter os lucros ou mantê-los na empresa investida[5].

No Brasil, o responsável pela compilação e disseminação do BP e da PII é o Banco Central do Brasil (BCB). Especificamente para a compilação dos lucros de empresas de investimento direto, o BCB tem como fonte de dados pesquisas e sistemas de registro onde as empresas de investimento direto reportam, dentre outras informações, seus lucros totais. Um desses sistemas é a Declaração Econômico-Financeira (DEF), um módulo do Registro Declaratório Eletrônico de Investimento Estrangeiro Direto (RDE-IED), que deve ser preenchida trimestralmente por todas as empresas de investimento direto com ativo ou patrimônio líquido total acima de R$ 250 milhões. Além de servir ao propósito primário de compilação de estatísticas, esta base, por armazenar dados individualizados de empresas de investimento direto, tem potencial para subsidiar análises econômicas, e constituiu a principal fonte para este trabalho.

O lucro total reportado no BP não é simplesmente o lucro contábil, mas sim o "lucro operacional corrente", ou COPC na sigla em inglês[6], que busca expurgar do lucro contábil itens que não fazem parte das operações usuais da empresa e economicamente não são consideradas lucros. Os itens expurgados são os lucros ou prejuízos causados por variação de preços de ativos, variações cambiais, operações não recorrentes como venda de subsidiárias, baixa contábil no valor de ativos ("*impairment*"), dentre outros. O desafio aqui, para os compiladores e para as empresas prestadoras de informação, é identificar corretamente esses itens e lançá-los na DEF de modo que possam ser devidamente subtraídos do lucro contábil. A variável resposta utilizada no modelo deste trabalho foi o lucro no conceito COPC, já descontado dos itens não operacionais. Além de este ser o conceito que interessa às estatísticas de setor externo, o seu uso facilitou a construção do modelo, pois os itens não operacionais têm baixa frequência, são mais difíceis de prever e por vezes tem alto impacto sobre o lucro contábil.

A construção do modelo descrita neste trabalho busca explorar os determinantes do lucro COPC do conjunto de empresas de investimento direto residentes no Brasil selecionadas para a amostra. Buscou-se entender como o lucro varia entre empresas, bem como ao longo do tempo. Além desta introdução, as seguintes seções compõem este artigo: descrição e preparo dos dados; construção do modelo; avaliação do modelo; e conclusões.

2 Descrição e preparo dos dados

O período coberto pela análise foi do início de 2017, após o módulo da DEF ser incluído no RDE-IED, ao final de 2021, perfazendo 20 trimestres. Nem todas as empresas registradas no sistema apresentaram DEF em todos estes trimestres, por variadas razões: em alguns casos, as empresas foram criadas após o início de 2017 ou extintas antes do final de 2021; outras empresas estiveram desobrigadas de entregar a declaração em um ou mais trimestres por estarem abaixo do piso mínimo de R$ 250 milhões de ativo ou patrimônio líquido; enquanto outras deixaram de ser empresas de investimento direto por incorporação ou fusão com outras companhias. Foram incluídas na análise apenas aquelas empresas que entregaram declaração em todos os 20 trimestres do período de cobertura, o que resultou em um total de 877 empresas. Desta forma, obteve-se um painel balanceado com 17.540 observações.

Como dito anteriormente, a variável resposta considerada foi o lucro COPC, equivalente ao lucro contábil das empresas subtraído de itens não operacionais. Ambos, lucro contábil e itens não operacionais, são informados na DEF e passam por monitoramento que garante a qualidade dos dados. Durante o processo de construção do modelo foram percebidos alguns valores *outliers* de lucros COPC que estavam gerando alavancagem ("*leverage*") indesejado nos coeficientes do modelo. Tais valores foram validados individualmente e estavam relacionados a itens não operacionais de difícil identificação que não haviam sido devidamente informados no sistema. Nesses casos, o valor correto foi imputado às observações.

Além do lucro COPC, também foi obtido na DEF o valor do ativo contábil, utilizado como variável explicativa por ser uma proxy do tamanho das empresas, presumivelmente relacionada ao lucro. Quanto maior o ativo, maior o tamanho da empresa e maior o seu lucro ou prejuízo potencial. Quando a empresa tinha patrimônio líquido negativo, tal patrimônio foi somado com sinal positivo ao ativo

[3] FMI [1], parágrafos 6.8 a 6.12.
[4] FMI [1], parágrafos 11.24 e 11.33.
[5] FMI [1], parágrafos 11.41.
[6] Do inglês "*Current Operating Performance Concept*", ou COPC, detalhado em OCDE [2] (um manual especializado em investimento direto, complementar a FMI [1]), parágrafo 208. Ver também parágrafo 11.44 de FMI [1].

contábil. Conceitualmente, a ideia é de que em casos assim a empresa tem um ativo contra seus acionistas, ao invés de estes terem um ativo contra ela. Na prática, fazer essa correção foi uma forma de manter a qualidade do ativo como proxy de tamanho da empresa nos casos em que o patrimônio líquido era negativo.

O setor de atividade econômica das empresas foi obtido em base de dados pública do Cadastro Nacional de Pessoas Jurídicas (CNPJ) da Receita Federal do Brasil (RFB) através do código CNPJ de cada empresa do painel. A princípio, o setor correspondeu à CNAE[7] informada na base de dados, com detalhamento de dois dígitos ("CNAE divisão"). Dois ajustes foram feitos a essa classificação inicial. O primeiro foi reclassificar setores quando a CNAE original não correspondia à atividade econômica final do grupo econômico. Por exemplo, algumas empresas de investimento direto são sedes locais de seus grupos econômicos e como tal tem códigos CNAE genéricos, como os de *holdings* ou de sedes de empresas. O segundo ajuste foi agrupar em uma categoria "Demais" aqueles códigos pouco representativos em termos de quantidade de empresas ou de valor de lucro. Ao final deste processo, restaram 26 códigos de setor de atividade econômica, que puderam ser usados como variáveis categóricas no modelo, incluindo o "Demais".

Por fim, foram feitos testes com vários indicadores do estado da economia, mas a maioria deles gerou um problema de multicolinearidade sobre o qual se comentará a seguir. Por conta disso, foi mantido na versão final do modelo apenas um dos indicadores, o relativo à arrecadação de receitas tributárias[8] pela RFB, por CNAE divisão. O dado original, mensal, foi agrupado para respeitar a frequência mensal das demais variáveis.

A Tabela 1 sintetiza as variáveis utilizadas no estudo.

Tabela 1: Dicionário do conjunto de dados.

Variável	Descrição
COPC	Lucro operacional, R$ milhões
Ativo	Ativo contábil corrigido, R$ milhões
Setor	CNAE divisão reclassificada
Arrecadação	Arrecadação da RFB, exceto previdenciária, por setor, R$ milhões

3 Construção do Modelo

Apresentamos a seguir uma sequência gradual de construção do modelo, útil para entender as relações entre as covariáveis e entre elas e a resposta. Na versão final do modelo, tais relações são numerosas devido às várias interações entre covariáveis e ao número de covariáveis categóricas de setor.

Começamos com um modelo de regressão linear simples, em que o lucro COPC em um trimestre é função apenas do ativo contábil corrigido do trimestre imediatamente anterior:

$$COPC_{i,t} = \beta_0 + \beta_1 Ativo_{i,\,t-1} \quad (1)$$

O p-valor do coeficiente para a variável de ativo foi significativo, enquanto o do intercepto não (o que é razoável, já que se espera que o lucro de empresas sem ativo deve tender a zero, a não ser em situações excepcionais), e o R^2 ajustado foi de 0,34[9]. Esses resultados dão suporte à intuição de que maiores empresas devem ter em média maiores lucros, enquanto o valor do coeficiente do ativo, de 0,015, indica que nesse painel um adicional de R$ 1 milhão em ativos de empresas de investimento direto gera, desconsiderando outras variáveis, R$ 15 mil adicionais em lucros, equivalente a um retorno de 1,5% por trimestre.

O uso do ativo defasado em um trimestre gera um componente dinâmico no modelo, permitindo que ele preveja alterações no lucro COPC derivadas de variações no tamanho das empresas relacionadas por exemplo a aportes de capital (ou descapitalizações), a fusões e aquisições de outras empresas ou a retenção de lucros para reinvestimento no negócio. Por outro lado, existem outros fatores internos às empresas, além do tamanho, que contribuem para o seu lucro, como estratégia gerencial, reação à dinâmica do mercado etc. Tais fatores podem ser de difícil identificação e mensuração. Uma forma de contornar essa dificuldade e capturar de maneira indireta a influência desses fatores é usar a própria variável resposta defasada como covariável no modelo[10]. O custo dessa estratégia é ter de desprezar os n primeiros períodos do painel, onde n é a defasagem máxima utilizada. Em nosso caso, optamos por 4 trimestres de defasagem, sob o pressuposto de que assim seria possível captar no modelo eventuais sazonalidades anuais do lucro[11]. Esse ponto sobre o número de defasagens será abordado novamente quando tratarmos de autocorrelação dos erros.

A especificação da nova versão do modelo é:

[7] Classificação Nacional de Atividades Econômicas. Para uma lista completa de códigos, acessar https://concla.ibge.gov.br/busca-online-cnae.html.

[8] Exceto receitas previdenciárias. Disponível em https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/dados-abertos/receitadata/arrecadacao/arrecadacao-por-divisao-economica-da-cnae

[9] Esses resultados foram obtidos desprezando-se os quatro primeiros trimestres do painel para que eles fossem comparáveis com os das versões posteriores do modelo, que empregaram defasagens.

[10] Sobre esse tema, Wooldridge [3], pág. 303, discute o uso de variáveis dependentes defasadas para controlar fatores não observados e inércia com um exemplo em que a resposta é a taxa de criminalidade.

[11] Tais sazonalidades no lucro podem estar relacionadas, por exemplo, ao ciclo natural de flutuação da atividade econômica ao longo do ano.

$$COPC_{i,t} = \beta_0 + \beta_1 Ativo_{i,\,t-1} + \sum_{n=1}^{4} \beta_{n+1} COPC_{i,t-n} \quad (2)$$

Nessa versão, o R^2 ajustado subiu para 0,54. Todos os coeficientes, inclusive o intercepto, tiveram p-valor significativo. Entre os coeficientes do lucro defasado, o de quatro defasagens (COPC_l4) teve valor de 0,25, igual ao coeficiente de primeira defasagem, o que parece dar suporte à hipótese de que a sazonalidade anual é um componente relevante dos lucros. Interessante notar também que, após controlarmos pelos lucros defasados, o retorno marginal do ativo corrigido caiu de 1,5% para 0,4% ao trimestre.

Tabela 2 - Coeficientes da versão 2 do modelo

Coefficients:	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3,528	1,1057	3,20	0,00142 **
Ativo_l1	0,004	0,0002	20,0	< 2e-16 ***
COPC_l1	0,246	0,0086	28,5	< 2e-16 ***
COPC_l2	0,204	0,0091	22,4	< 2e-16 ***
COPC_l3	0,097	0,0094	10,3	< 2e-16 ***
COPC_l4	0,253	0,0093	27,2	< 2e-16 ***

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

O próximo passo é incluir a variável categórica de setor de atividade econômica. O setor em que uma empresa atua pode estar relacionado de diversas formas ao seu lucro, seja via um retorno sobre ativo diferente da média de mercado, um padrão sazonal ou uma tendência de crescimento dos lucros próprios, ou simplesmente um lucro médio diferente daquele visto em outros setores. Sendo assim, parece natural incluir o setor de forma que ele interaja com as demais covariáveis, o que nos leva à seguinte especificação:

$$COPC_{i,t} = \beta_0 + \beta_s Setor_i + (\beta_1 + \beta_{1,s} Setor_i) Ativo_{i,\,t-1} + \sum_{n=1}^{4} (\beta_{n+1} + \beta_{n+1,s} Setor_i) COPC_{i,t-n} \quad (3)$$

O modelo foi estimado tomando-se como setor de referência o "demais". A interpretação dos coeficientes de setor é a seguinte:

- β_s indica quão diferente é o intercepto para o setor s, em relação ao setor de referência.
- $\beta_{1,s}$ indica quão diferente é o retorno em relação ao ativo para o setor s, em comparação com o setor de referência.
- $\beta_{2,s}...\beta_{5,s}$ indicam quão diferente, para o setor s, é a relação entre o lucro em t e o lucro defasado em 1...4 trimestres.

Cada interação envolvendo a variável categórica de setor gerou 25 covariáveis no modelo (26 níveis de setor menos o nível de referência). Assim, o modelo teve no total 156 coeficientes. O R^2 ajustado foi de 0,6, um ganho de 0,06 em relação à versão anterior. Embora o R^2 ajustado já inclua penalização para a adição de novas variáveis, é interessante avaliar também outros critérios de informação, considerando que nessa versão foram introduzidas muitas novas variáveis. O AIC[12] desta versão foi de 172.884, menor que os 174.647 da versão anterior, indicando que a quantidade de informação trazida pelas novas variáveis supera a penalização.

Por fim, temos os indicadores econômicos, que ao trazer informação sobre o ambiente onde as empresas estão operando, podem ajudar a estimar seus lucros. Tentou-se adicionar à versão 3 do modelo indicadores de atividade no comércio (indicador Serasa de atividade no comércio, índice de movimento no comércio a prazo); na indústria (sondagem de utilização da capacidade instalada na indústria da FGV, PMI industrial); indicadores de confiança (índices de confiança da indústria da CNI e FGV, índices de confiança de serviços e do consumidor da FGV); e o índice de preços de commodities do BCB (IC-Br). Isoladamente, nenhum dos indicadores teve capacidade de alterar positivamente o poder explicativo do modelo, tendo sido necessário adicioná-los em interação com i. o ativo e o setor; ou com ii. o COPC defasado e o setor. O motivo para isso é intuitivo: o impacto de uma alteração nas condições econômicas sobre o lucro depende do tamanho da empresa e do nível recente de lucro que ela vinha obtendo. O problema com esse procedimento foi que ele gerou alta multicolinearidade. Isso não ocorreu necessariamente por conta de alta correlação entre os indicadores, mas porque as interações de cada indicador são correlacionadas com outros termos do modelo. Ver por exemplo a Tabela 3, com os fatores de inflação de variância generalizados (GVIFs na sigla em inglês) e transformados para um modelo em que só foi adicionado um indicador, o nível de utilização de capacidade instalada na indústria, em uma única interação. Note-se que essa interação do indicador tem GVIF transformado acima do valor de referência de 10, indicando alta multicolinearidade. Para além disso, conforme outros indicadores são adicionados, o GVIF transformado sobe e rapidamente tende ao infinito.

[12] Akaike's Information Criterion. Ver Sakamoto, Ishiguro and Kitagawa [5].

Tabela 3 - GVIFs em modelo com indicador econômico

Coefficients	GVIF	Df	GVIF^(1/(2*Df))
Ativo_l1	3,1E+00	1	1,75
Setor	1,2E+00	25	1,00
COPC_l1	2,6E+00	1	1,60
COPC_l2	2,6E+00	1	1,60
COPC_l3	2,6E+00	1	1,60
COPC_l4	2,5E+00	1	1,59
Ativo_l1:Setor	1,9E+13	25	1,84
Setor:COPC_l1	2,2E+09	25	1,54
Setor:COPC_l2	1,4E+09	25	1,52
Setor:COPC_l3	8,6E+08	25	1,51
Setor:COPC_l4	3,3E+08	25	1,48
Ativo_l1:Setor: nuci	8,5E+64	26	17,73

Embora a presença de multicolinearidade seja esperada quando há produtos de covariáveis no modelo, e embora nessa situação ela não necessariamente gere problemas para a predição da resposta, decidiu-se não usar esses indicadores nesta versão do trabalho para tentar preservar o intervalo de confiança das estimativas dos coeficientes, bem como evitar troca de sinais de coeficientes em relação ao que seria esperado.

Entretanto, foi possível adicionar ao modelo, sem gerar multicolinearidade relevante, a arrecadação de receitas pela RFB. Um diferencial desse indicador em relação aos demais é que ele é detalhado por setor, o que pode ter contribuído para atenuar a multicolinearidade.

O indicador de arrecadação foi usado na construção de uma nova covariável, resultado da sua multiplicação pelo ativo ou lucro defasado, pelo setor e pelo trimestre (q1, ..., q4). A multiplicação pelo setor se justifica pelo fato de o indicador ser segmentado por setor, e assim queremos fazer a estimativa não em relação à arrecadação média de toda a amostra, mas à arrecadação de cada setor específico. Já a categórica de trimestre foi incluída porque a arrecadação tem um padrão sazonal anual próprio, ligado à dinâmica de apuração e recolhimento dos tributos federais. Com a inclusão da nova variável, a nova especificação foi a seguinte:

$$
\begin{aligned}
COPC_{i,t} = {} & \beta_0 + \beta_s Setor_i + (\beta_1 + \beta_{1,s} Setor_i) Ativo_{i,\,t-1} \\
& + \sum_{n=1}^{4} (\beta_{n+1} + \beta_{n+1,s} Setor_i) COPC_{i,t-n} \\
& + \beta_{5,s,q} Setor_i Ativo_{i,t-1} Q_q Arrec_{t,s} \\
& + \beta_{6,s,q} Setor_i COPC_{i,t-1} Q_q Arrec_{t,s} \\
& + \beta_{7,s,q} Setor_i COPC_{i,t-2} Q_q Arrec_{t,s} \\
& + \beta_{8,s,q} Setor_i COPC_{i,t-4} Q_q Arrec_{t,s} \quad (4)
\end{aligned}
$$

A Tabela 4 traz os GVIFs transformados para esta versão do modelo, sendo possível ver que todos os valores estão abaixo de 10.

Tabela 4 - VIFs generalizados na versão 4 do modelo

Coefficients	GVIF	Df	GVIF^ (1/(2*Df))
Ativo_l1	3,9E+00	1	1,97
Setor	1,2E+00	25	1,00
COPC_l1	3,3E+00	1	1,81
COPC_l2	3,2E+00	1	1,78
COPC_l3	3,5E+00	1	1,88
COPC_l4	3,2E+00	1	1,79
Ativo_l1:Setor	3,2E+15	25	2,04
Setor:COPC_l1	7,5E+11	25	1,73
Setor:COPC_l2	2,4E+11	25	1,69
Setor:COPC_l3	2,0E+12	25	1,76
Setor:COPC_l4	1,0E+11	25	1,66
Ativo_l1:Setor:arrecadacao:Tri	7,9E+81	104	2,48
Setor:arrecadacao:Tri:COPC_l1	1,8E+87	104	2,63
Setor:arrecadacao:Tri:COPC_l2	3,4E+85	104	2,58
Setor:arrecadacao:Tri:COPC_l4	1,5E+84	104	2,54

4 Avaliação do Modelo

A especificação do modelo na equação 4 teve um R^2 ajustado de 0,71, um ganho de 0,1 em relação à versão anterior. A diferença entre esse R^2 ajustado e o R^2 (0,72) foi de 0,01, o que reforça a hipótese de que não houve multicolinearidade relevante. O AIC foi de 168.964 (172.884 na versão anterior), indicando que o ajuste do modelo melhorou, mesmo considerando a penalização pela introdução de novas covariáveis. O gráfico de "*Residuals vs Fitted*" mostra resíduos médios próximos de zero para toda a faixa de valores estimados, sem indícios de heterocedasticidade.

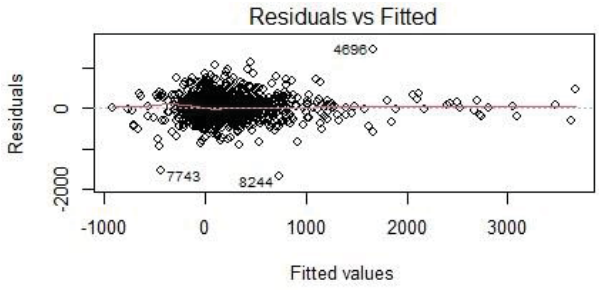

Figura 1: resíduos em relação a valores ajustados.

Nas próximas subseções, tratamos de questões específicas referentes ao ajuste do modelo.

4.1 ANOVA

Cada um dos termos envolvendo o indicador de arrecadação tem 104 graus de liberdade, resultado da multiplicação das 26 categorias de setor pelas 4 de trimestres. Como são 4 termos, temos 416 graus de liberdade, que, somados aos 156 coeficientes que já existiam na versão anterior, resultam em um total de 572 coeficientes.

Dado esse alto número, não serão analisados a seguir p-valores de coeficientes individuais, mas sim a

estatística F de cada termo, agrupado pelas variáveis categóricas individuais. A Tabela 5 apresenta a análise de variância de tipo II do modelo, em que a soma de quadrados de um termo x é aquela obtida após a inclusão de todos os demais termos, com exceção de eventuais termos de maior ordem relativos a x. Ou seja, no tipo II é medida a contribuição marginal de cada termo[13].

Tabela 5 - ANOVA tipo II para versão 4 do modelo

	Sum Sq	Df	F	P(>F)
Ativo_l1	2.022.521	2	106	***
Setor	5.597.902	45	13	***
COPC_l1	2.658.576	1	279	***
COPC_l2	2.211.309	2	116	***
COPC_l3	282.132	1	30	***
COPC_l4	2.078.230	2	109	***
Ativo_l1:Setor	4.551.547	26	18	***
Setor:COPC_l1	3.404.390	25	14	***
Setor:COPC_l2	3.134.088	26	13	***
Setor:COPC_l3	2.060.313	25	9	***
Setor:COPC_l4	1.704.992	26	7	***
Ativo_l1:Setor:Arrec:Tri	13.461.009	104	14	***
Setor:Arrec:Tri:COPC_l1	9.885.576	104	10	***
Setor:Arrec:Tri:COPC_l2	9.577.113	104	10	***
Setor:Arrec:Tri:COPC_l4	9.123.989	104	9	***
Residuals	128.426.280	13.460		

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Vemos que o termo de interação entre arrecadação e ativo é aquele com maior contribuição marginal, seguido pelos três termos de interação entre arrecadação e COPC defasado. Além disso, o teste F tem significância para todos os termos, indicando que todos contribuem de forma estatisticamente significativa para explicar a soma de quadrados da resposta. Adiante os resultados deste teste serão comparados com o mesmo teste feito para um modelo de regressão robusta às caudas espessas da distribuição dos resíduos.

4.2 Alavancagem e observações influentes

Figura 2 traz o gráfico de diagnóstico de resíduos contra alavancagem para o modelo. Há algumas observações ao redor do limiar de distância de Cook de 0,5, com uma delas levemente acima desse limiar, em 0,52. Uma inspeção mais detalhada dos "dfbetas" dessa observação mostra que a influência, quando ocorre, se dá apenas sobre os coeficientes que envolvem o setor específico daquela observação, sendo nula para os coeficientes de outros setores e os coeficientes gerais de ativo e lucro COPC defasados. Ou seja, a influência fica confinada apenas ao setor específico daquela observação.

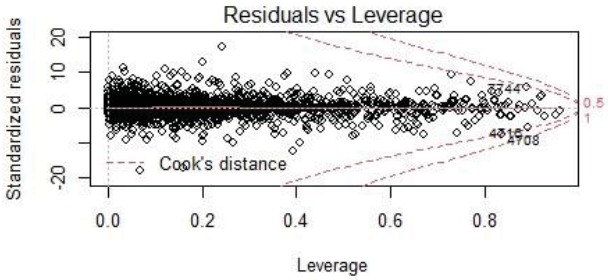

Figura 2: Resíduos em relação a alavancagem.

4.3 Correlação serial nos resíduos

A estrutura temporal embutida nos dados de painel pode gerar dependência entre o $COPC_{i,t}$ de uma empresa i e o $COPC_{i,t-n}$, onde n é um número de defasagens igual ou maior que 1. Se as covariáveis do modelo não forem capazes de explicar essa estrutura de dependência, pode ocorrer correlação serial (autocorrelação) nos resíduos, o que por sua vez implica que, condicionalmente às covariáveis, as estimativas para a resposta não são independentes entre si. A implicação é que as observações carregam menos informação do que assumido pelo modelo, fazendo com que os intervalos de confiança dos coeficientes sejam excessivamente otimistas (menores do que deveriam).

Aqui, voltamos à discussão sobre o número ideal de defasagens da variável dependente a serem incluídas no modelo. Adicionar mais defasagens pode fazer com que a estrutura de correlação serial da resposta seja mais bem capturada pelo modelo, reduzindo a autocorrelação nos resíduos. Assim, primeiro testamos a hipótese de existência de autocorrelação nos resíduos e depois, se necessário, podemos incluir mais defasagens da resposta para tentar mitigá-la.

Para fazer o teste levando em consideração que temos um painel, estimamos a seguinte equação:

$$e_{i,t} = \beta_0 + \sum_{j=1}^{n} \beta_j e_{i,t-j} \quad (5)$$

Ao estimarmos essa equação, a hipótese nula de que os coeficientes dos resíduos defasados são iguais a zero equivale a dizer que não há correlação serial dos resíduos. Caso a hipótese nula seja rejeitada, o valor do coeficiente indica a força da autocorrelação. Além disso, esperamos que o intercepto estimado tenha valor estatisticamente nulo, correspondendo a um resíduo médio igual a zero.

A equação 5 foi estimada com vários valores de n, e em nenhuma delas foi encontrada autocorrelação relevante. Por exemplo, com n = 6 tivemos:

[13] Para mais detalhes, ver Fox and Weisberg [6], seção 5.3.4.

Tabela 6 - Coeficientes do teste de autocorrelação

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1,052107	1,106735	0,95	0,341
residuals_l1	-0,017111	0,0116	-1,5	0,140
residuals_l2	0,004374	0,0120	0,4	0,714
residuals_l3	-0,02216	0,0115	-1,9	0,054 .
residuals_l4	-0,035864	0,0118	-3,0	0,002 **
residuals_l5	-0,034023	0,0127	-2,7	0,007 **
residuals_l6	0,022781	0,0131	1,7	0,081 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Notamos que os coeficientes para 4 e 5 defasagens são estatisticamente significativos a 5%, enquanto o de 6 defasagens é significativo a 10%. Mas em todos esses casos o valor em si do coeficiente é muito baixo, não superando 3,6% em módulo. Ou seja, a autocorrelação detectada é muito fraca. Já o intercepto, como esperado, não teve valor estatisticamente distinto de zero.

4.4 Normalidade dos resíduos

Uma forma de visualizar se os resíduos seguem uma distribuição normal é através de um histograma dos resíduos (figura 3).

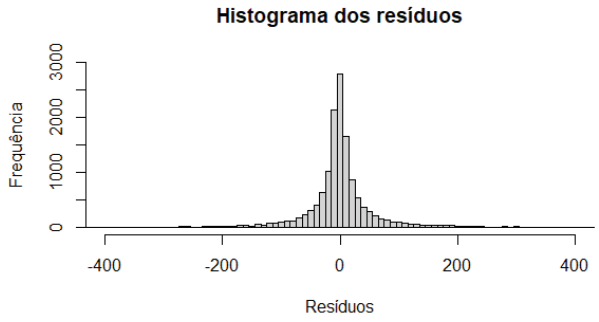

Figura 3: histograma dos resíduos indica que distribuição não é normal.

Nota-se que a distribuição tem aparência leptocúrtica, com um pico alto e fino e, principalmente, caudas espessas e longas – além de leve assimetria para a esquerda. Por questão de espaço, não foram incluídos no histograma os resíduos abaixo de R$ -400 milhões (73 observações) e acima de R$ 400 milhões (90 observações). De fato, o primeiro e o quarto quartis tem extensões bastante alongadas, enquanto a outra metade dos resíduos, entre o primeiro e o terceiro quartis, se encontra em uma faixa relativamente bem mais estreita:

Tabela 7 - Quartis dos resíduos

Mín	1Q	Mediana	3Q	Max
-1688,5	-17,9	-1,9	15,8	1452

O formato do histograma, em particular as caudas "pesadas", sugere que os resíduos não seguem uma distribuição normal. Embora o estimador de mínimos quadrados ordinários continue consistente e não viesado em uma situação assim (ou seja, continuamos com um modelo de média adequado), os intervalos de confiança para os coeficientes e de predição para as estimativas ficam prejudicados. Ao assumir uma distribuição normal para os erros, com caudas menos espessas do que a distribuição real, os testes de hipótese estimam intervalos menores, ou seja, mais otimistas, do que os verdadeiros. Isso pode levar o analista a considerar como estatisticamente significativos coeficientes que na verdade não o são, ou assumir que a resposta estimada se encontra em um intervalo de predição menor do que o verdadeiro, dado um nível de confiança.

Para contornar esse problema e avaliar seu efeito sobre o modelo, serão analisadas nas próximas seções duas alternativas: i. regressão robusta e ii. estimação assumindo erros com distribuição t de Student, com caudas mais espessas que as da normal.

4.5 Regressão robusta

A regressão robusta de modelos lineares introduz pesos na equação de estimação, que são calculados de forma iterativa em função dos resíduos. De forma geral, quanto maiores os resíduos, menores os pesos[14]. Com isso, tem-se um estimador que, embora viesado, é robusto à presença de observações com alto resíduo, como é o caso daquelas que compõem as caudas pesadas do nosso painel.

O objetivo de usar o modelo de regressão robusta é comparar o resultado dos testes F de seu quadro ANOVA com aqueles da versão original do modelo, que haviam tido resultados estatisticamente significativos para todas as covariáveis. Se algum dos testes F da estimação robusta for não significativo, tem-se um indicativo que a distorção introduzida pelos erros não normais fez com que houvesse covariáveis indevidamente classificadas como significativas.

Como se pode ver na Tabela 8, o teste F continua estatisticamente significativo para todos os termos. Assim, a regressão robusta não apresentou evidência da presença de algum termo não significativo no modelo, e que só teria parecido ser significativo por conta dos intervalos de confiança otimistas gerados pelos erros não normais.

[14] Huber [7] e Hampel, Ronchetti, Rousseeuw and Stahel [8]

Tabela 8 - ANOVA tipo II para regressão robusta

	Df	F	P(>F)
Ativo_l1	2	1198,3	***
Setor	45	92,593	***
COPC_l1	1	5352,5	***
COPC_l2	2	1313,2	***
COPC_l3	1	389,75	***
COPC_l4	2	2423,8	***
Ativo_l1:Setor	26	137,91	***
Setor:COPC_l1	25	103,41	***
Setor:COPC_l2	26	83,271	***
Setor:COPC_l3	25	60,483	***
Setor:COPC_l4	26	28,42	***
Ativo_l1:Setor:Arrec:Tri	104	89,547	***
Setor:Arrec:Tri:COPC_l1	104	71,059	***
Setor:Arrec:Tri:COPC_l2	104	91,571	***
Setor:Arrec:Tri:COPC_l4	104	66,474	***
Residuals	13460		

Signif. codes: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

4.6 Regressão com distribuição t

Diferentemente da regressão robusta, em que a função perda é ajustada para dar menos peso a erros maiores, aqui se assume explicitamente uma distribuição diferente da normal para os erros, no caso, a t de *Student*. Essa distribuição tem um parâmetro a mais que a normal, os graus de liberdade, que controla a espessura das caudas. Quanto menor esse parâmetro, mais espessas as caudas da distribuição.

Neste exercício[15], os graus de liberdade não foram fixados, mas deixados como um parâmetro a ser estimado no processo de minimização da função perda. O valor calculado nesse processo foi de 0,83, indicando caudas bastante pesadas. A log-verossimilhança do modelo foi de -71.718,8 maior que os -83.908,9 do modelo normal e os -84.802,3 da regressão robusta. Logo, há uma probabilidade maior que o processo gerador dos dados seja mais próximo de uma distribuição t do que de uma distribuição normal. Os R^2 do modelo normal e t são, entretanto, muito próximos, o que é um indício de que eles não diferem de forma relevante na estimação da média. No entanto, por ter uma distribuição mais próxima da verdadeira, o modelo t é mais adequado para fornecer intervalos de confiança dos coeficientes e de predição para a resposta.

5 Conclusões

Esse trabalho apresentou um modelo linear para estimar lucros operacionais correntes trimestrais de um painel de empresas de investimento direto residentes no Brasil, de forma a explicar tanto variações destes lucros entre as empresas quanto de uma empresa ao longo do tempo. O modelo apresentado foi capaz de colocar sob um mesmo guarda-chuva empresas bastante diferentes, especialmente em termos de porte. De fato, o ativo mediano do painel se situou ao redor de R$ 1 bilhão, com um mínimo de R$ 250 milhões[16], um terceiro quartil de R$ 2 bilhões, mas um valor máximo de R$ 132 bilhões. A própria variável resposta também apresentou distribuição incondicional com caudas espessas. Enquanto metade das respostas ficou entre um 1º quartil de R$ -3,4 milhões e um 3º quartil de R$ 40 milhões, com mediana de R$ 10 milhões, o lucro mínimo foi de R$ -2 bilhões e o máximo, de R$ 4,1 bilhões.

Embora o modelo tenha sido bem-sucedido em acomodar essas grandes diferenças entre as empresas, além de capturar as tendências de variação de seus lucros ao longo do tempo, sem apresentar heterocedasticidade, alavancagem ou autocorrelação relevantes, o modelo não foi capaz de apresentar resíduos com distribuição normal. Tal situação faz com que os intervalos de confiança dos coeficientes sejam muito otimistas. Apesar disso, os testes F da regressão robusta indicaram que não houve inversão de resultado dos testes de hipótese, ou seja, não havia termos não significativos no modelo que haviam sido incorretamente considerados como significativos. Além do efeito nos intervalos de confiança, as caudas espessas do resíduo fazem com que os intervalos de predição do modelo que assume normalidade sejam menores do que os verdadeiros. Para estimar intervalos de predição, é mais confiável usar um modelo que assuma uma distribuição para os erros mais próxima da verdadeira, com caudas mais espessas que as de uma normal, como a distribuição t avaliada na seção anterior. Por outro lado, para estimação pontual da média, espera-se que o modelo normal seja suficiente, pois a assunção de normalidade dos resíduos não é necessária para garantir consistência e ausência de viés.

O trabalho de construção do modelo mostrou empiricamente que ele tem capacidade de identificar observações *outliers* e com alavancagem, em especial de grandes empresas. Essa capacidade o torna apto a ser empregado na validação dos dados informados nos sistemas de registro pelas empresas declarantes. Além disso, o modelo pode, naturalmente, ser empregado para predição, o que aponta caminhos para trabalhos futuros. Qual a performance do modelo na predição da resposta para períodos t maiores que aqueles cobertos pela amostra original? Como os coeficientes do modelo se comportarão quando novos períodos forem incorporados? Estabilidade nos coeficientes e na performance, em particular após a incorporação na amostra de períodos de mudanças intensas nas condições econômicas e na própria resposta, indicariam resiliência do modelo. Um ponto adicional

[15] Foi usado no exercício o pacote GAMLSS do R, que permite estimar modelos lineares com distribuições não gaussianas. Ele é baseado em Rigby and Stasinopoulos [9].

[16] Este é o limite mínimo para que as empresas tenham de reportar suas informações contábeis trimestrais no sistema de registro.

é conseguir fazer a inclusão de novos indicadores econômicos sem incorrer no problema de multicolinearidade reportado neste trabalho. Uma possibilidade é recorrer a indicadores mais granulares do que os utilizados, em especial aqueles segmentados por setor, como o indicador de arrecadação.

Referências

[1] Fundo Monetário Internacional (FMI). *Balance of Payments and International Investment Position Manual*. International Monetary Fund, 2009.

[2] Organização para a Cooperação e Desenvolvimento Econômico (OCDE). *OECD Benchmark Definition of Foreign Direct Investment*. OECD, 2008.

[3] Jeffrey M. Wooldridge. *Introductory Econometrics: A Modern Approach*. Cengage, 2019.

[4] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2010.

[5] Y. Sakamoto, M. Ishiguro, and G. Kitagawa. *Akaike Information Criterion Statistics*. D. Reidel Publishing Company, 1986.

[6] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Sage Publications, 2019.

[7] P. J. Huber. *Robust Statistics*. Wiley, 1981.

[8] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw and W. A. Stahel. *Robust Statistics: The Approach based on Influence Functions*. Wiley, 1986.

[9] R. A. Rigby and D. M. Stasinopoulos. *Generalized additive models for location, scale and shape*. Journal of the Royal Statistical Society: Series C (Applied Statistics), volume 54, issue 3, 2005.