



Universidade Federal de Lavras  
Superintendência de Governança  
Coordenadoria de Inteligência e Governança de Dados

# **Análise do Aumento das Despesas com Terceirização de Mão de Obra na UFLA de 2009 a 2022**

Flávio Lopes de Moraes  
João Chrysostomo de Resende Júnior  
Adriano Higino Freire  
Márcio Machado Ladeira

**Reitor**

João Chrysostomo de Resende Júnior

**Vice-Reitor**

Valter Carvalho de Andrade Júnior

**Chefe de Gabinete**

Cinthia Divino Bustamante Murad

**Superintendente de Governança**

Adriano Higino Freire

**Superintendente de Integridade e Correição**

Débora Cristina de Carvalho

**Pró-Reitores**

Pró-Reitora de Assuntos Estudantis e Comunitários

Elisângela Elena Nunes Carvalho

Pró-Reitora de Extensão e Cultura

Christiane Maria Barcellos Magalhães da Rocha

Pró-Reitora de Gestão e Desenvolvimento de Pessoas

Viviane Naves de Azevedo

Pró-Reitor de Graduação

Ronei Ximenes Martins

Pró-Reitor de Infraestrutura e Logística

João Cândido de Souza

Pró-Reitor de Pesquisa

Luciano José Pereira

Pró-Reitor de Planejamento e Gestão

Márcio Machado Ladeira

Pró-Reitora de Pós-Graduação

Adelir Aparecida Saczk

**Diretor do Núcleo de Inovação Tecnológica**

Márcio André Stefanelli Lara

**Coordenador de Inteligência e Governança de Dados**

Flávio Lopes de Moraes

## Resumo

Este estudo consistiu em investigar a existência de relação entre o aumento proporcional dos gastos com a contratação de mão de obra terceirizada na UFLA e a defasagem na quantidade de técnicos administrativos em relação ao número de docentes e estudantes de graduação. Para atingir esse objetivo, foram coletados dados do SIAFI e do INEP, bem como foram aplicadas análises e testes estatísticos para avaliar a possível relação entre essas variáveis. De acordo com os resultados da análise, foi observada uma forte correlação inversa entre a variável que representa a proporção percentual de técnicos em relação aos docentes ("tae\_docente") e a variável que representa os gastos com terceirização ("terceirizacao"). Além disso, também foi identificada uma correlação inversa significativa, porém menos intensa, entre a variável que representa a proporção percentual de técnicos em relação aos estudantes de graduação ("tae\_estudante") e a variável "terceirizacao". Buscando quantificar essas correlações e entender o comportamento da variável "terceirizacao", foram desenvolvidos três modelos de regressão linear. No Modelo 1 utilizou-se a variável "tae\_docente" como variável explicativa, no Modelo 2 utilizou-se a variável "tae\_estudante" como variável explicativa e, no Modelo 3, utilizou-se ambas as variáveis como explicativas. As análises de regressão confirmaram que ambas as variáveis independentes têm capacidade para explicar o comportamento da variável dependente com um nível de confiança de 95%. Foi constatado que a variável "tae\_docente", sozinha, é capaz de explicar 85,4% da variação no percentual de gastos com terceirização na UFLA, a um nível de confiança de 99%. A inclusão da variável "tae\_aluno" aumenta a capacidade explicativa do modelo em 1,6 pontos percentuais. Contudo, esse aumento vem acompanhado de uma maior complexidade e da diminuição da significância estatística do modelo.

# 1. Introdução

Ao longo dos anos, a UFLA vem se consolidando como uma das mais importantes instituições de ensino superior do Brasil, fato que pode ser comprovado por meio do Índice Geral de Cursos (IGC), avaliado pelo Ministério da Educação. De 2010 a 2021, a UFLA ficou sempre entre as 10 primeiras Universidades Federais do Brasil e as 3 primeiras de Minas Gerais. Tal desempenho reflete o trabalho que tem sido desenvolvido no âmbito estrutural e pedagógico da Instituição, fazendo com que a UFLA venha se mantendo no seleto grupo de Universidades do Brasil que receberam o conceito máximo (nota 5).

Esses índices atestam a eficiência da Instituição e a qualidade de seus cursos. Assim, deve-se ressaltar que mesmo com a importante expansão da UFLA, o desafio de manter a qualidade desta IFES tem sido alcançado, garantindo uma prestação de serviços públicos de acordo com os anseios da sociedade.

De todo o sistema de universidades federais, a UFLA é a instituição que, percentualmente, mais depende de recursos de custeio para para pagamento de mão de obra terceirizada, pois além da deficiência de Técnicos Administrativos em Educação (TAES), causa principal desse fenômeno, temos um câmpus em Lavras com cerca de 500 hectares, três fazendas experimentais, área do hospital universitário, além de um campus em São Sebastião do Paraíso, o que eleva enormemente os custos das despesas correntes, como manutenção predial, jardinagem, limpeza, vigilância, dentre outros. Essa deficiência de TAES, apesar de ser um processo histórico, foi agravado consideravelmente pelo não cumprimento de pactuações realizadas com o MEC com vistas ao aumento de vagas nos cursos de graduação. Nas expansões de cursos desde o ano de 2014, onde a UFLA implantou cinco cursos de engenharias, curso de medicina e pedagogia, bem como o novo campus de São Sebastião do Paraíso, com quatro cursos, as contrapartidas do MEC com vagas de TAES, representam apenas 1/3 do total pactuado.

Esses recursos de custeio, extremamente pressionados pela necessidade de contratação de mão de obra, colocam em risco a qualidade do ensino, o apoio institucional à permanência dos estudantes, entre outras atividades, já que restam poucos recursos para apoio direto às atividades de ensino, pesquisa e extensão.

Diante desse grande desafio, e de um orçamento aprovado na LOA de 2023 que não é suficiente para o funcionamento do ano, temos defendido a necessidade de que o financiamento das Ifes precisa estar vinculado a um indicador que possibilite a constância de orçamento suficiente, à semelhança do que ocorre com as universidades estaduais paulistas, para que possa haver planejamento e uma prestação de serviços cada vez melhor à sociedade.

A UFLA tem enfrentado desafios significativos na gestão do seu quadro de pessoal, especialmente quando se trata de técnicos administrativos. A limitação de recursos financeiros e a dificuldade na realização de concursos públicos para a contratação de novos técnicos administrativos têm agravado esse problema. Consequentemente, observa-se uma diminuição na proporção de técnicos administrativos por docente e estudante de graduação ao longo dos anos, acompanhada por um aumento relativo nos gastos com a contratação de mão de obra terceirizada.

Objetivamos com este trabalho investigar se há relação entre o aumento relativo dos gastos com a contratação de mão de obra terceirizada na UFLA e a defasagem no número de técnicos administrativos comparada ao número de docentes e ao número de estudantes de graduação.

O restante deste trabalho está organizado da seguinte forma: Na seção 2, descrevemos o universo de estudo e a metodologia utilizada no desenvolvimento desta análise. A seção 3 apresenta uma comparação da terceirização na UFLA com dados de outras universidades públicas federais, seguida pela análise estatística dos resultados que buscam explicar o aumento da terceirização na UFLA devido à defasagem no quantitativo de técnicos administrativos. Finalmente, a seção 4 traz as conclusões finais desta análise.

## 2. Materiais e Métodos

Nesta análise, o universo de estudo é formado pelos dados dos valores orçamentários empenhados e dos quantitativos de TAEs, docentes e estudantes de graduação de todas as universidades públicas federais brasileiras. A seguir, são apresentadas as ferramentas, as técnicas, as fontes de dados e as regras utilizadas na extração, tratamento e análise dos dados utilizados.

### 2.1. Extração e tratamento dos dados

Abaixo são descritas as fontes de dados, o escopo temporal considerado e as regras aplicadas no tratamento dos dados.

#### 2.1.1. Dados de valores orçamentários empenhados

2.1.1.1. **Fonte dos dados:** Os dados de valores empenhados foram obtidos do sistema SIAFI, a partir do Tesouro Gerencial, disponível em <https://tesourogerencial.tesouro.gov.br/>, acessado em 10 de março de 2023.

2.1.1.2. **Escopo temporal:** Foram extraídos dados de 2013 a 2022. À princípio, havíamos incluído dados desde 2008, no entanto, de acordo com o Diretor de Contabilidade da UFLA, dados anteriores a 2013 não eram confiáveis para utilização nesta análise devido à mudança no plano de contas desta universidade ocorrida em tal ano.

2.1.1.3. **Regras aplicadas na extração de dados do Tesouro Gerencial:**

- Item informação igual a “29:DESPESAS EMPENHADAS (CONTROLE EMPENHO)”.
- Órgão UGE - Órgão Máximo (Código) igual a “26000”.
- Órgão UGE (Código) na lista: 26457;26456;26455;26454;26453;26452;26450;26449;26448;26447;26442;26441;26440;26352;26351;26230;26231;26350;26286;26285;26284;26283;26282;26281;26280;26279;26278;26277;26276;26275;26274;26273;26272;26271;26270;26269;26268;26267;26266;26264;26263;26262;262

61;26260;26258;26255;26254;26253;26252;26251;26250;  
26249;26248;26247;26246;26245;26244;26243;26242;262  
41;26240;26239;26238;26237;26236;26235;26234;26233;  
26232.<sup>1</sup>

2.1.1.4. **Regras aplicadas no tratamento dos dados:** Após a extração dos dados, as seguintes regras foram aplicadas:

- **Regra para definição de Orçamento Discricionário:** São filtrados apenas os empenhos de recursos orçamentários, cujo código da ação de governo esteja dentre as seguintes: 4572, 20GK, 20RK, 216H, 00PW, 8282 e 4002.
- **Regra para definição de Orçamento de Custeio:** São filtrados apenas os empenhos de recursos orçamentários cujo código do grupo de despesa esteja entre os seguintes: 1, 2 ou 3.
- **Regra para definir o que é gasto orçamentário com Terceirização:** É considerada uma despesa com terceirização todo empenho cujo código da natureza de despesa seja igual a '339037', ou o código da natureza de despesa detalhada (sub elemento de despesa) seja '33903916', '33903977', '33903978' ou '33903979'.

2.1.2. Dados relativos a TAEs, docentes e estudantes de graduação

Abaixo são descritas a fonte de dados, o escopo temporal e as regras aplicadas na extração dos quantitativos de TAEs, docentes e estudantes de graduação.

2.1.2.1. **Fonte dos dados:** Os dados foram extraídos a partir do portal de Dados Abertos do Censo da Educação Superior, providos pelo INEP no seguinte endereço: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/censo-da-educacao-superior>, acessado em 10 de março de 2023. Nesse portal é disponibilizado um arquivo compactado (.ZIP) contendo os arquivos de dados, em formato .CSV, relativos aos cursos e às instituições de ensino superior brasileiras, tanto públicas quanto privadas. Para o escopo deste trabalho, foram extraídos apenas os dados relativos às universidades públicas federais.

2.1.2.2. **Escopo temporal:** Foram extraídos dados de 2009 a 2021. Embora no portal mencionado acima constem dados anteriores a 2009, estes estão em formato incompatível com os dados de 2009 em diante. Os dados relativos ao Censo de 2022 ainda não estavam disponíveis no momento da extração dos dados.

---

<sup>1</sup> Códigos SIAFI das 69 universidades públicas federais.

2.1.2.3. **Regras aplicadas no tratamento dos dados:** Após a extração dos dados, as seguintes regras foram aplicadas:

- Os quantitativos de TAEs e docentes em exercício foram extraídos respectivamente a partir dos campos QT\_TEC\_TOTAL e QT\_DOC\_EXE dos arquivos MICRODADOS\_CADASTRO\_IES\_[[ANO]].CSV.
- Os quantitativos de estudantes de graduação foram extraídos a partir da soma dos valores do campo QT\_MAT dos arquivos MICRODADOS\_CADASTRO\_CURSOS\_[[ANO]].CSV de todos os cursos pertencentes à instituição. O Campo QT\_MAT possui descrição no dicionário de dados do INEP igual a “Quantidade de matrículas”, com a seguinte observação: “Cálculo de matrículas: soma do número de estudantes com situação de vínculo ao curso igual a: Cursando e/ou Formado”.

Os dados coletados passaram por um processo de limpeza e tratamento. Além disso, foram criadas novas variáveis, como o percentual individual e médio gastos com terceirização e o percentual individual e médio da relação entre a quantidade de técnicos administrativos e docentes. Após essa etapa, os dados foram armazenados no *Data Warehouse* da UFLA.

## 2.2. Análise dos dados

A análise dos dados foi realizada com o auxílio do *software* Metabase<sup>2</sup> e da linguagem de programação Python. Inicialmente, foi realizada uma análise descritiva dos dados, utilizando as bibliotecas Pandas<sup>3</sup> e Seaborn<sup>4</sup>.

Em seguida, foi realizada a análise de regressão utilizando as bibliotecas statsmodels<sup>5</sup> e SciPy<sup>6</sup> do Python. Foram construídos três modelos de regressão linear, sendo que cada um deles foi ajustado para explicar a relação entre a variável dependente e as variáveis independentes.

Para avaliar a qualidade dos modelos, foram calculados os coeficientes de determinação ( $R^2$ ) e realizados testes de hipóteses para verificar a significância estatística dos coeficientes estimados.

Finalmente, foi realizada uma análise quanto à validade dos modelos diante dos pressupostos requeridos para modelos de regressão do tipo *Ordinary Least Squares regression* (OLS).

---

<sup>2</sup> <https://www.metabase.com/>

<sup>3</sup> <https://pandas.pydata.org/>

<sup>4</sup> <https://seaborn.pydata.org/>

<sup>5</sup> <https://www.statsmodels.org/>

<sup>6</sup> <https://scipy.org/>

A partir dos resultados obtidos, foi possível avaliar a capacidade dos modelos de explicar a variação da taxa de terceirização de mão de obra na UFLA, bem como sua validade estatística.

### 2.3. Escolha do modelo de regressão

A escolha por utilizar a regressão se deu pelo fato de que esta é uma técnica estatística que permite modelar e investigar a relação entre uma variável dependente e uma ou mais variáveis independentes (Fávero e Belfiore, 2017).

É fundamental destacar que as variáveis examinadas nesta análise representam séries temporais, abrangendo um período que compreende desde o ano de 2009 até o ano de 2022. De acordo com Fávero e Belfiore (2017), quando se trata de análise de séries temporais, é possível utilizar a regressão para entender o comportamento de uma série em relação a outra. Portanto, uma das principais justificativas para a utilização da análise de regressão nesse contexto é a possibilidade de identificar o grau de associação entre as séries temporais e, assim, entender a relação entre elas.

Ainda de acordo com Fávero e Belfiore (2017), a escolha do modelo de regressão deve levar em consideração a análise da característica e do padrão de distribuição da variável dependente, juntamente com as características das variáveis independentes. É importante escolher um modelo que seja apropriado para os dados e para os objetivos da análise.

Neste sentido, após analisar as características das variáveis dependentes e independentes, foi selecionado o modelo de regressão linear por mínimos quadrados ordinários (*Ordinary Least Squares regression - OLS*) -- Técnica para estimar os coeficientes de uma equação de regressão linear que descreve o relacionamento entre uma ou mais variáveis independentes quantitativas e uma variável dependente, também quantitativa (Fávero e Belfiore, 2017).

Na análise de regressão simples, cada variável independente foi analisada separadamente em relação à variável dependente, ou seja, "tae\_docente" e "tae\_estudante" foram analisadas individualmente para avaliar sua influência sobre "terceirizacao". Já na análise de regressão múltipla, ambas as variáveis independentes foram incluídas no modelo para avaliar a contribuição conjunta delas sobre a variável dependente.

A seguinte nomenclatura foi utilizada para facilitar a identificação dos modelos utilizados:

- Modelo 1: "terceirizacao" como variável dependente, e "tae\_docente" como variável independente.
- Modelo 2: "terceirizacao" como variável dependente, e "tae\_estudante" como variável independente.
- Modelo 3: "terceirizacao" como variável dependente e "tae\_docente" juntamente com "tae\_estudante" como variáveis independentes.



### 3. Resultados e Discussão

Nesta seção, são apresentados os resultados e as discussões da presente análise.

#### 3.1. Comparação do nível de terceirização na UFLA em relação às outras universidades públicas federais

Com o objetivo de melhor compreender os gastos da UFLA com mão de obra terceirizada em relação às outras universidades públicas federais, foi elaborado um gráfico (figura 1)<sup>7</sup> que apresenta o valor anual do percentual gasto com terceirização pela UFLA, o percentual médio considerando todas as instituições e o valor percentual da instituição que mais gastou com terceirização naquele ano. O cálculo do percentual de gastos com terceirização é obtido por meio da razão entre o valor gasto com terceirização e o valor gasto com orçamento discricionário de custeio.

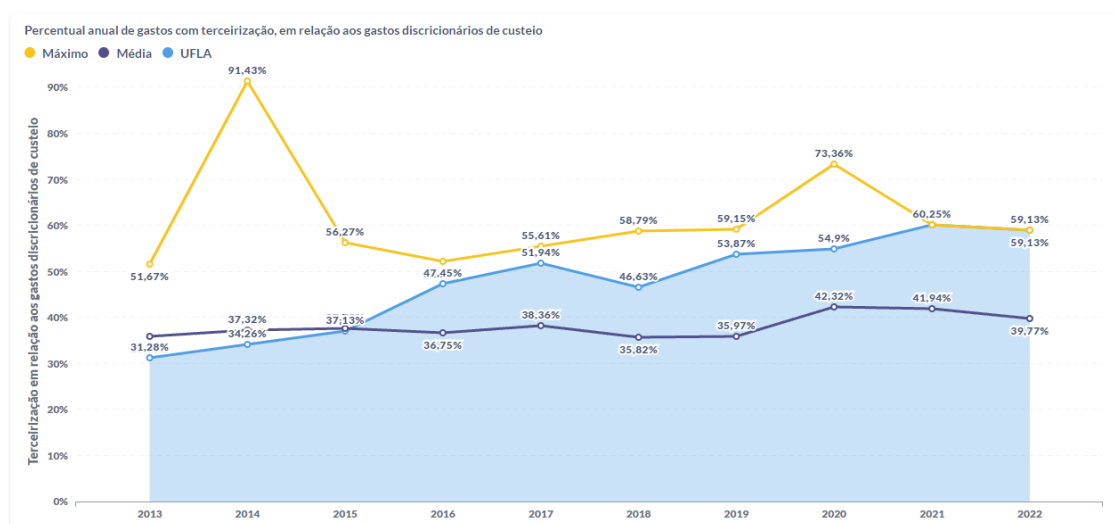


Figura 1: Percentual anual de gastos com terceirização, em relação aos gastos discricionários de custeio.

Analisando a figura 1, é possível notar que os gastos relativos da UFLA com terceirização de mão de obra foram ascendentes, atingindo a média em 2015 e superando a média a partir de 2016. A partir de 2021, a UFLA se tornou a instituição com o maior gasto relativo com terceirização, atingindo 60,25% em 2021 e 59,13% em 2022.

Objetivando compreender as razões que levaram ao aumento da terceirização nesta instituição, foram elaborados os gráficos apresentados nas figuras 2 e 3. A figura 2<sup>8</sup> mostra a proporção, por ano, da quantidade de servidores TAEs ativos em relação à quantidade de docentes em exercício, e a figura 3<sup>8</sup> mostra a proporção, por ano, da quantidade de servidores TAEs ativos em relação à quantidade de estudantes de

<sup>7</sup> Para mais detalhes, acesse:

- [Percentual anual de gastos com terceirização, em relação aos gastos discricionários de custeio.](#)
- [Painel detalhado de gastos com terceirização.](#)

<sup>8</sup> Para mais detalhes, acesse:

- [Relação percentual entre TAEs, Estudantes de Graduação e Docentes.](#)
- [Evolução quantitativa de Técnicos, Estudantes e Docentes.](#)

graduação. Ambos os gráficos exibem ainda as proporções médias, considerando todas as instituições utilizadas nesta análise.

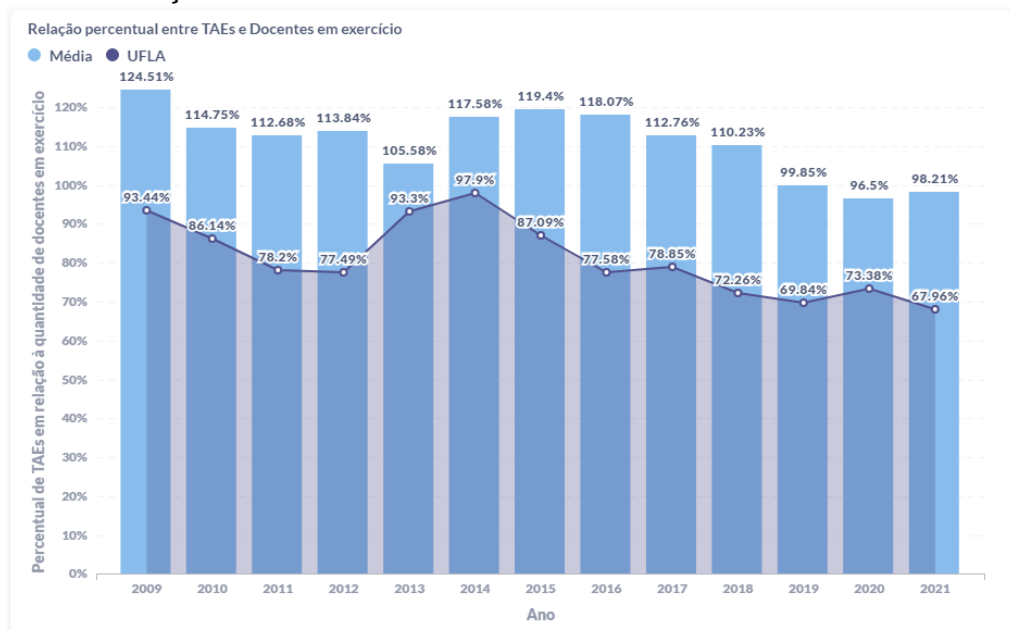


Figura 2: Relação percentual entre TAEs e docentes da UFLA em relação à média das universidades públicas federais.

O gráfico representado na figura 2 evidencia uma queda contínua na proporção entre TAEs e docentes na UFLA de 2009 a 2012, passando de 93.44% em 2009 para 77.49% em 2012. Houve um aumento nessa proporção em 2013 e 2014, refletindo a expansão ocorrida pelo Programa de Apoio a Planos de Expansão e Reestruturação das Universidades Federais (REUNI), que foi a última pactuação de TAEs inteiramente cumprida pelo MEC. A partir de 2014, onde se iniciou o segundo grande ciclo de expansão, com os cursos de engenharias, medicina e pedagogia, a taxa tem diminuído progressivamente, partindo de 97.9% em 2014 e atingindo 67.96% em 2021.

É Possível também notar que a relação entre TAEs e docentes na UFLA permanece significativamente abaixo da média ao longo de toda a série histórica avaliada, refletindo defasagem dessa relação em comparação com outras instituições.

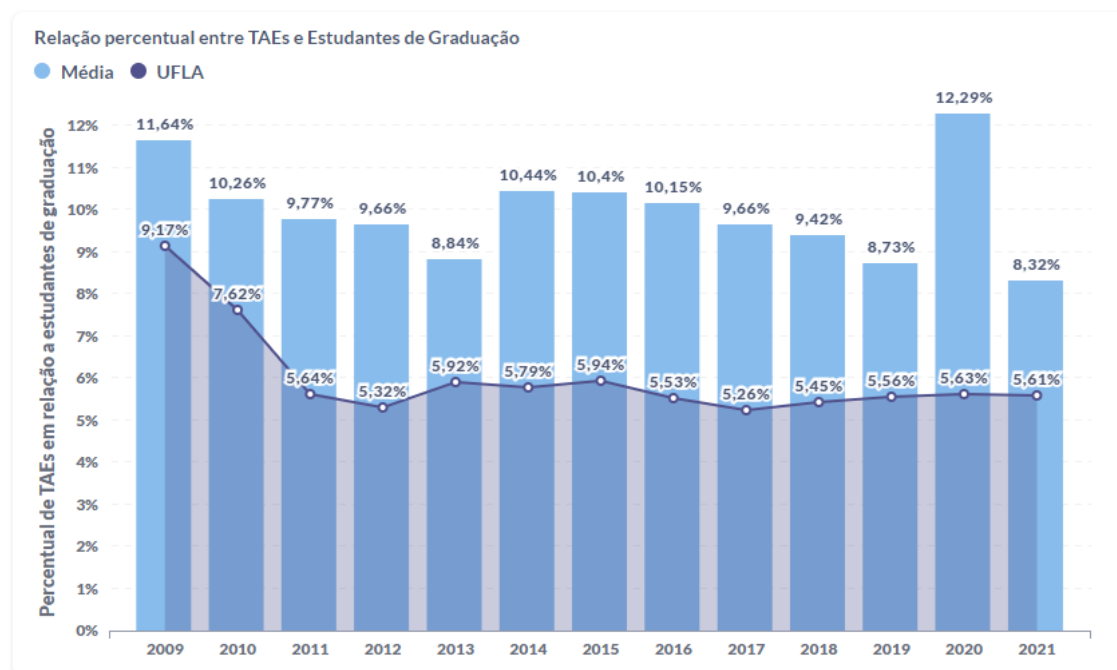


Figura 3: Relação percentual entre TAEs e estudantes de Graduação da UFLA em relação à média das universidades públicas federais.

Conforme ilustrado na figura 3, a proporção de servidores TAEs em relação aos estudantes de graduação na UFLA apresentou uma queda contínua entre 2009 e 2012, caindo de 9.17% em 2009 para 5.32% em 2012. Após esse período de declínio constante, o indicador se estabilizou, variando de 5.26% a 5.94% até 2021.

Nota-se também que a relação entre servidores TAEs e estudantes de graduação na UFLA permanece significativamente abaixo da média ao longo de toda a série histórica avaliada, também sugerindo uma defasagem dessa relação em comparação com outras instituições.

Os dados da figura 2 e da figura 3 podem ser interpretados da seguinte maneira: Quanto menor a relação entre servidores TAEs e docentes, ou entre servidores TAEs e estudantes de graduação, maior poderia ser a carga de trabalho sobre tais servidores, e, portanto, essas poderiam ser possíveis justificativas para o aumento de investimentos com terceirização de mão-de-obra na UFLA, ao longo do tempo.

Os dados apresentados nas figuras 2 e 3 sugerem que a diminuição da proporção de servidores técnico-administrativos em relação aos docentes ou estudantes de graduação levam ao aumento da carga de trabalho desses servidores, justificando o aumento dos investimentos em terceirização de mão-de-obra ao longo do tempo na UFLA.

O restante deste trabalho buscou examinar se as proporções de TAEs em relação aos docentes (variável *tae\_docente*) e aos estudantes de graduação (variável *tae\_estudante*) podem explicar o aumento da terceirização na UFLA (variável *terceirizacao*). Para tal, foi realizada uma análise descritiva dos dados, seguida de uma análise de regressão.

### 3.2. Análise estatística sobre o aumento da terceirização na UFLA

Nesta seção, foram realizadas análises para verificar se as mudanças nas proporções de TAEs em relação aos docentes e aos estudantes de graduação podem explicar as alterações nas taxas de terceirização na UFLA.

#### 3.2.1. Conjunto de dados utilizado

Com o objetivo de investigar se as variáveis “tae\_docente” e “tae\_estudante” têm influência sobre a variável “terceirizacao”, foram utilizados apenas os dados referentes à UFLA. O conjunto de dados utilizado pode ser encontrado na tabela 1.

Ano	tae_docente	tae_estudante	terceirizacao
2009	93,44	9,17	N/D
2010	86,14	7,62	N/D
2011	78,20	5,64	N/D
2012	77,49	5,32	N/D
2013	93,30	5,92	31,28
2014	97,90	5,79	34,26
2015	87,09	5,94	37,13
2016	77,58	5,53	47,45
2017	78,85	5,26	51,94
2018	72,26	5,45	46,63
2019	69,84	5,56	53,87
2020	73,38	5,63	54,90
2021	67,96	5,61	60,25
2022	N/D	N/D	59,13

Tabela 1: Base de dados com os valores percentuais das variáveis tae\_docente, tae\_estudante e terceirizacao relativos à UFLA.

A tabela 1 revela que as variáveis "tae\_docente" e "tae\_estudante" contêm valores não definidos (N/D) para o ano de 2022, enquanto a variável "terceirizacao" apresenta valores não definidos (N/D) para os anos 2009, 2010, 2011 e 2012. Portanto, optamos por excluir os registros dos anos que têm valores ausentes, uma vez que esses dados não estão disponíveis devido a restrições das próprias fontes de dados, conforme explicado na seção 2. A base de dados resultante após a exclusão dos registros com valores faltantes pode ser visualizada por meio da tabela 2.

Ano	tae_docente	tae_estudante	terceirizacao
2013	93,30	5,92	31,28
2014	97,90	5,79	34,26
2015	87,09	5,94	37,13
2016	77,58	5,53	47,45
2017	78,85	5,26	51,94
2018	72,26	5,45	46,63
2019	69,84	5,56	53,87
2020	73,38	5,63	54,90
2021	67,96	5,61	60,25

Tabela 2: Base de dados com os valores percentuais das variáveis tae\_docente, tae\_estudante e terceirizacao relativos à UFLA, após a exclusão dos registros com valores faltantes.

### 3.2.2. Análise descritiva dos dados

O conjunto de dados analisado é composto por informações relativas a 9 anos consecutivos, de 2013 a 2021, e se refere à proporção de TAEs em relação a docentes (variável tae\_docente), à proporção de TAEs em relação a estudantes de graduação (variável tae\_estudante) e à proporção de terceirização em relação ao total de gastos discricionários de custeio (variável terceirizacao) na UFLA.

A tabela 3 contém as medidas descritivas para as variáveis "tae\_docente", "tae\_estudante" e "terceirizacao". Ao examinar essa tabela, é possível concluir que a média da relação entre TAEs e docentes na UFLA é de 79,79%, a média da relação entre TAEs e estudantes de graduação é de 5,63% e o percentual médio de terceirização é de 46,41%. Além disso, as medidas de desvio padrão, média e mediana (representada pelo segundo quartil - Q2) evidenciam que a variável "terceirizacao" possui uma variação relativa superior às demais variáveis.

	tae_docente	tae_estudante	terceirizacao
<b>Quantidade</b>	9	9	9
<b>Média</b>	79,79	5,63	46,41
<b>Desvio Padrão</b>	10,64	0,22	10,08
<b>Mínimo</b>	67,96	5,26	31,28
<b>Q1</b>	72,26	5,53	37,12
<b>Q2</b>	77,58	5,61	47,44
<b>Q3</b>	87,09	5,79	53,86
<b>Máximo</b>	97,90	5,94	60,25

Tabela 3: Medidas descritivas das variáveis tae\_docente, tae\_estudante e terceirizacao.

O gráfico da figura 4 exibe os diagramas de dispersão dos pares de variáveis utilizados na análise. A diagonal principal retrata o diagrama de uma variável em relação a ela mesma, o que resulta em um padrão linear perfeitamente definido.

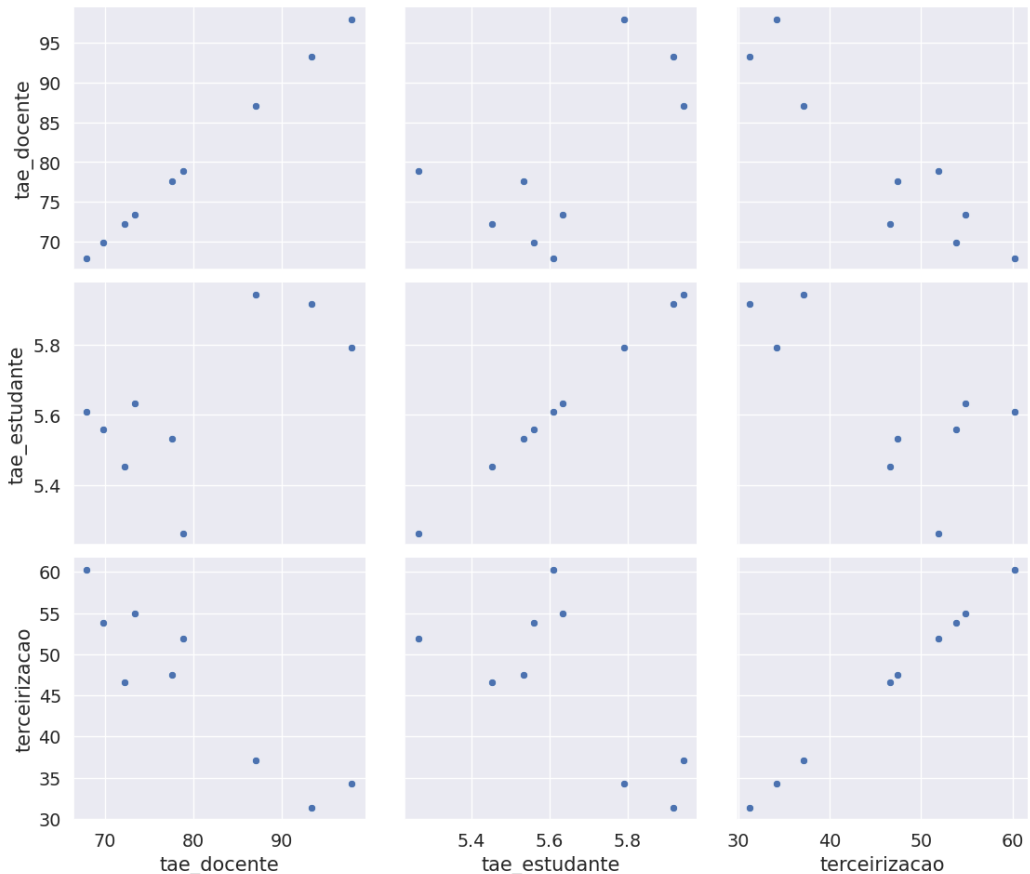


Figura 4: Diagramas de dispersão das variáveis tae\_docente, tae\_estudante e terceirizacao.

Por meio dos diagramas da figura 4 é possível observar um comportamento linear inverso entre as variáveis "tae\_docente" e "terceirizacao", sugerindo que quanto menor a proporção de TAEs em relação aos docentes, maiores são os gastos com terceirização de mão de obra. Embora os gráficos também indiquem uma tendência de comportamento linear inverso entre as variáveis "tae\_estudante" e "terceirizacao", nota-se uma exceção quando os valores de "tae\_estudante" estão próximos de 5,6. Por sua vez, o diagrama de dispersão que relaciona "tae\_docente" e "tae\_estudante" apresenta uma relação menos linear e mais positiva, indicando uma tendência para que o aumento em uma variável seja acompanhado pelo aumento da outra.

A análise das figuras 5, 6 e 7 obtidas com a técnica *Kernel Density Estimator* (KDE) permite avaliar a distribuição das variáveis "terceirizacao", "tae\_docente" e "tae\_estudante". Com base nessas figuras, é possível obter uma noção de quão próximas as distribuições das variáveis "terceirizacao" (*terceirizacao density*), "tae\_docente" (*densidade de TAE docente*) e "tae\_estudante" (*tae\_estudante density*) estão de uma distribuição normal (*Normal distribution density*).

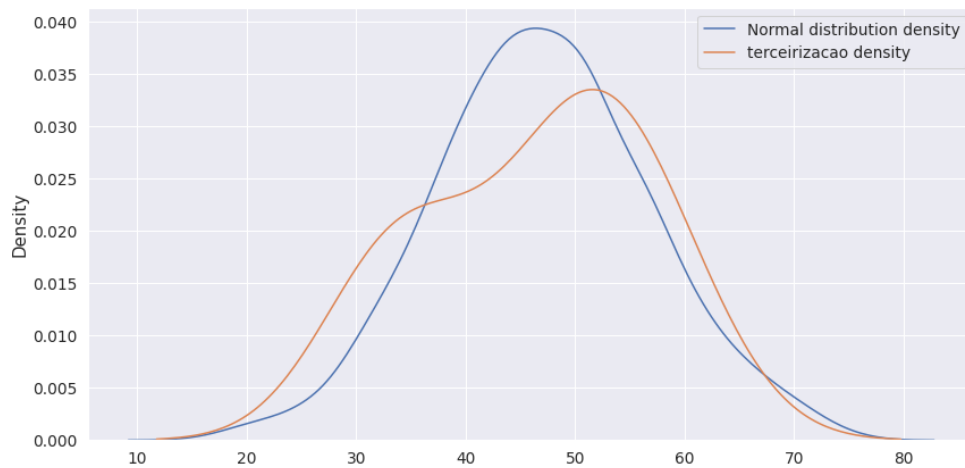


Figura 5: Aderência entre a distribuição dos valores da variável “terceirizacao” e a distribuição normal.

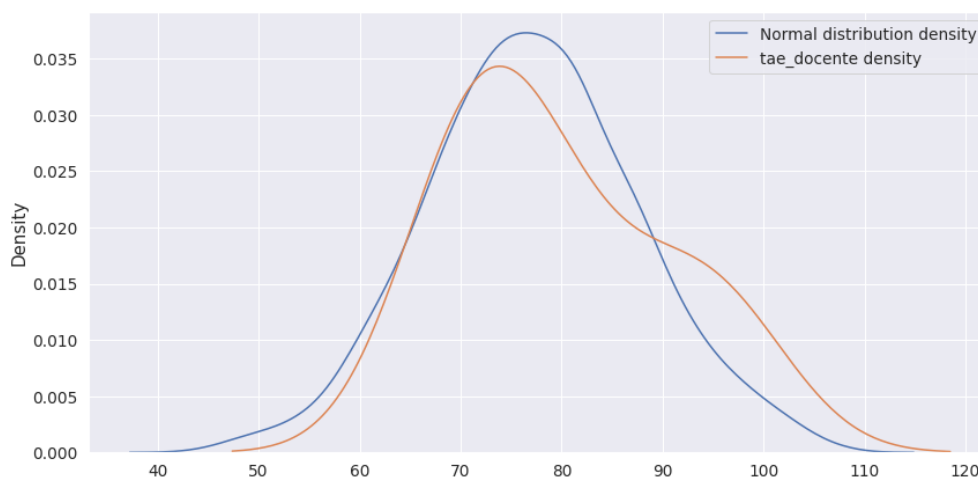


Figura 6: Aderência entre a distribuição dos valores da variável “tae\_docente” e a distribuição normal.

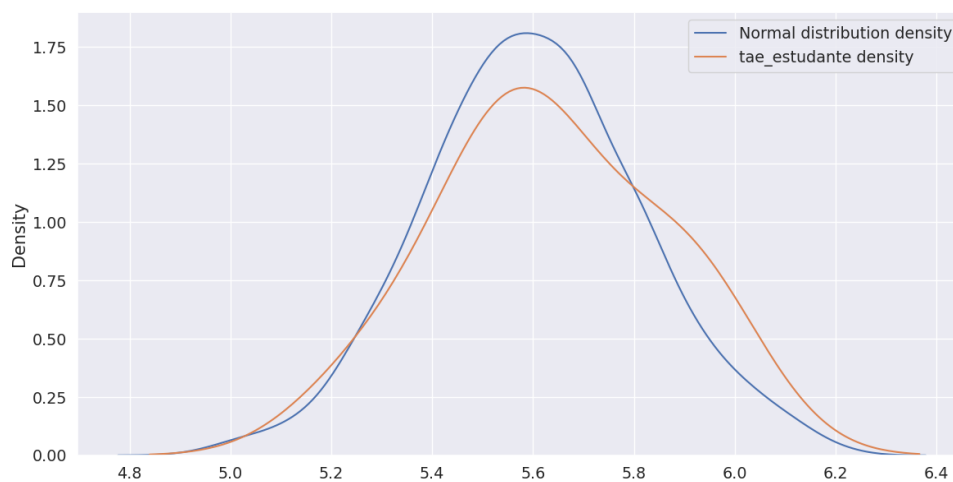


Figura 7: Aderência entre a distribuição dos valores da variável “tae\_estudante” e a distribuição normal.

Buscando confirmar a aderência das variáveis à normalidade, foi executado o teste de Shapiro-Wilk -- teste de normalidade univariada, aplicável em amostras

com tamanho entre 4 e 2000 elementos (Fávero e Belfiore, 2017) --, cujos resultados são apresentados pela figura 8.

```
print('terceirizacao: ',stats.shapiro(df.terceirizacao.values))
print('tae_docente: ',stats.shapiro(df.tae_docente.values))
print('tae_estudante: ',stats.shapiro(df.tae_estudante.values))

terceirizacao: ShapiroResult(statistic=0.9342262744903564, pvalue=0.5225975513458252)
tae_docente: ShapiroResult(statistic=0.9083102941513062, pvalue=0.30422720313072205)
tae_estudante: ShapiroResult(statistic=0.9573407769203186, pvalue=0.7701181173324585)
```

Figura 8: Resultado da aplicação do teste de Shapiro-Wilk sobre as variáveis em análise.

O teste de Shapiro-Wilk testa a hipótese nula de que a amostra provém de uma população com distribuição normal. A figura 8 exibe os valores-P das três variáveis, os quais estão acima do valor crítico de 0,05. Com base nesses resultados, pode-se concluir que não há evidências suficientes para rejeitar a hipótese nula. Portanto, é possível afirmar, com um nível de confiança de 95%, que as amostras das variáveis foram extraídas de uma distribuição normal.

Dando continuidade à análise descritiva dos dados, foi realizado o cálculo do coeficiente de correlação de Pearson entre as variáveis "tae\_docente", "tae\_estudante" e "terceirizacao". O coeficiente de correlação de Pearson (*ccp*) é uma medida que varia de -1 a 1 e permite verificar a relação linear entre duas variáveis analisadas. À medida que os valores das variáveis se aproximam dos extremos, a correlação entre elas se torna mais forte (Fávero e Belfiore, 2017).

- Se *ccp* for positivo, há uma relação diretamente proporcional entre as variáveis;
- Quando *ccp* é igual a 1, a correlação linear entre elas é perfeita e positiva;
- Se *ccp* for negativo, há uma relação inversamente proporcional entre as variáveis;
- Quando *ccp* é igual a -1, a correlação linear entre elas é perfeita e negativa;
- Se *ccp* for zero, não há correlação entre as variáveis.

A figura 9 apresenta o resultado do cálculo do coeficiente de correlação de Pearson sobre as variáveis da análise. Os resultados indicaram uma correlação negativa elevada entre as variáveis "tae\_docente" e "terceirização", com um valor de -0,92, e uma correlação negativa um pouco menor, porém ainda considerável, entre as variáveis "tae\_estudante" e "terceirizacao", com um valor de -0,69. Esses achados indicam que os valores de "tae\_docente" e "tae\_estudante" estão negativamente relacionados aos valores de "terceirizacao". Um valor negativo de *ccp* significa que quando uma variável aumenta, a outra tende a diminuir, e vice-versa.



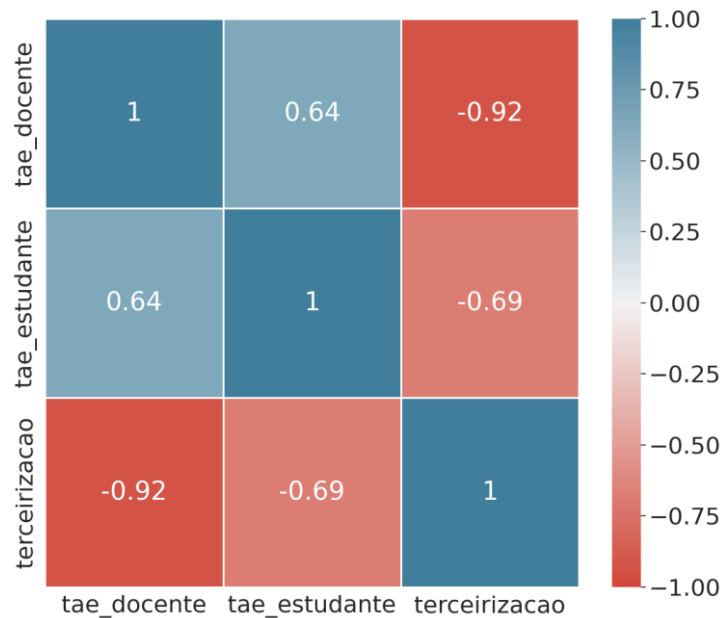


Figura 9: Resultado do cálculo do *Coefficiente de Correlação de Pearson* para as variáveis tae\_docente, tae\_estudante e terceirizacao.

A figura 9 também mostra uma correlação positiva de 0,64 entre as variáveis "tae\_docente" e "tae\_estudante", indicando que essas variáveis tendem a apresentar comportamentos semelhantes. Vale ressaltar que as correlações com valor 1 na horizontal correspondem às correlações de uma variável consigo mesma, ou seja, a diagonal da matriz de correlação.

De acordo com o que foi observado na figura 10, todas as correlações apresentaram valor-P abaixo de 0,05, exceto pela correlação entre as variáveis "tae\_docente" e "tae\_estudante". Isso significa que essas correlações são estatisticamente significativas com um nível de confiança de 95%. É importante destacar que a correlação entre "terceirizacao" e "tae\_estudante" teve um valor-P abaixo de 0,01, indicando significância estatística a um nível de confiança de 99%.

```
print('p-value tae_docente X terceirizacao:', pearsonr(df.tae_docente.values, df.terceirizacao.values)[1])
print('p-value tae_aluno X terceirizacao:', pearsonr(df.tae_estudante.values, df.terceirizacao.values)[1])
print('p-value tae_docente X tae_estudante:', pearsonr(df.tae_docente.values, df.tae_estudante.values)[1])

p-value tae_docente X terceirizacao: 0.00036580628414076077
p-value tae_aluno X terceirizacao: 0.04134104634614107
p-value tae_docente X tae_estudante: 0.0653259031258409
```

Figura 10: Resultado do cálculo do valor-P das correlações entre as variáveis "tae\_docente", "tae\_estudante" e "terceirizacao".

É importante destacar que, embora a correlação entre duas variáveis indique o grau de associação entre elas, essa relação não implica em causalidade. Em outras palavras, não se pode afirmar que uma variável causa a outra apenas com base em uma correlação. Para estabelecer uma relação de causalidade, é preciso realizar estudos mais aprofundados e considerar outros fatores que possam estar envolvidos. Portanto, é essencial ter cautela ao interpretar correlações e não assumir uma relação causal sem evidências adicionais (Fávero e Belfiore, 2017).

### 3.2.3. Análise de Regressão

Com o objetivo de aprofundar a análise sobre o relacionamento das variáveis "tae\_docente" e "tae\_estudante" com a variável "terceirizacao", são apresentados nesta seção os resultados de análises de regressão sobre os dados utilizados neste estudo.

Os resultados das análises dos três modelos são apresentados a seguir.

#### 3.2.3.1. Modelo 1: terceiraizacao em função de tae\_docente

A figura 11 apresenta o resumo dos resultados obtidos a partir do modelo de regressão simples desenvolvido com a variável independente "tae\_docente" e a variável dependente "terceirizacao".

Model:	OLS	Adj. R-squared:	0.833			
Dependent Variable:	terceirizacao	AIC:	52.7497			
Date:	2023-03-22 12:02	BIC:	53.1441			
No. Observations:	9	Log-Likelihood:	-24.375			
Df Model:	1	F-statistic:	41.01			
Df Residuals:	7	Prob (F-statistic):	0.000366			
R-squared:	0.854	Scale:	16.947			
	<b>Coef.</b>	<b>Std.Err.</b>	<b>t</b>	<b>P&gt; t </b>	<b>[0.025</b>	<b>0.975]</b>
<b>Intercept</b>	116.2881	10.9968	10.5748	0.0000	90.2850	142.2913
<b>tae_docente</b>	-0.8757	0.1367	-6.4042	0.0004	-1.1990	-0.5524

Figura 11: Resumo dos resultados obtidos a partir do modelo de regressão linear. Variável "terceirizacao" em função da variável "tae\_docente".

A partir da figura 11 é possível ver que o modelo gerou um coeficiente de determinação (*R-squared*), ou  $R^2$ , de 0,854. O  $R^2$  é uma medida estatística que indica o grau de ajuste do modelo de regressão aos dados. Ele varia de 0 a 1 e representa a proporção da variabilidade na variável dependente que é explicada pelas variáveis independentes incluídas no modelo (Montgomery et al., 2012). Em outras palavras, a variável que representa a proporção de TAEs em relação à quantidade de docentes explica 85,4% da variação no percentual de gastos com terceirização, enquanto os 14,6% restantes são atribuídos a outras variáveis que não foram consideradas no modelo. No entanto, de acordo com Stock e Watson (2004) e Fávero et al. (2009), o  $R^2$  não diz se determinada variável explicativa é estatisticamente significativa e se esta variável é a causa verdadeira da alteração de comportamento da variável depende.

A avaliação da significância estatística de um modelo de regressão linear simples pode ser feita por meio do teste t. O teste t é utilizado para testar a hipótese nula de que o coeficiente da variável explicativa (*beta*) no modelo de regressão é igual a zero, o que significa que essa variável não tem efeito significativo na variável dependente. Se o valor-P do teste *t* for menor que um nível de significância pré-determinado (por exemplo, 0,05), então rejeitamos a hipótese nula e concluímos que a variável explicativa é estatisticamente

significativa para explicar a variabilidade da variável dependente (Fávero e Belfiore, 2017).

A figura 11 mostra que o parâmetro estimado do *beta*, referente à variável “tae\_docente”, mostrou-se estatisticamente diferente de zero ao nível de significância de 1%, uma vez que a magnitude do seu erro-padrão resultou em um valor-P ( $P > |t| < 0,01$ ). Em outras palavras, o resultado do teste *t* dos resultados apresentados pela figura 11 nos permite afirmar que a variável explicativa (“tae\_docente”) é estatisticamente significativa para explicar o comportamento da variável dependente (“terceirizacao”). Como o nível de significância observado é 0,0004, valor inferior a 0,01, o teste *t* apresenta rejeição da hipótese nula, o que nos permite concluir, ao nível de confiança de 99%, que a variável “tae\_docente” tem efeito significativo sobre a variável “terceirizacao”.

A figura 12 mostra o gráfico do modelo de regressão linear gerado para a variável “terceirizacao” em função da variável “tae\_docente”. Cada ponto no gráfico (cor azul) representa um par de valores das duas variáveis. A linha vermelha é a linha de regressão, que representa a melhor estimativa da relação entre as variáveis. Essa linha foi determinada através do método estatístico OLS, que minimiza a distância entre a linha de regressão e todos os pontos do gráfico.

O gráfico da figura 12 mostra ainda os intervalos de confiança para a linha de regressão (Mean\_ci) e para os dados observados (Obs\_ci).

- As linhas Mean\_ci representam o intervalo de confiança para a linha de regressão. Esses intervalos representam a faixa de valores em que a linha de regressão pode estar, com um grau de confiança de 99%, para um determinado valor da variável independente. Em outras palavras, esses intervalos representam a incerteza na estimativa da relação entre as variáveis e indicam a faixa dentro da qual se espera que a média populacional das previsões esteja, com o nível de confiança especificado.
- As linhas Obs\_ci representam o intervalo de confiança para a média dos valores observados da variável dependente (“terceirizacao”) para um determinado valor da variável independente (“tae\_docente”). Esses limites indicam a faixa dentro da qual se espera que a previsão populacional esteja com um certo nível de confiança, no caso, 99%. Em outras palavras, se a mesma relação entre as duas variáveis fosse observada várias vezes em diferentes amostras, a previsão dentro dessa faixa seria correta em uma certa porcentagem das vezes, de acordo com o nível de confiança especificado.

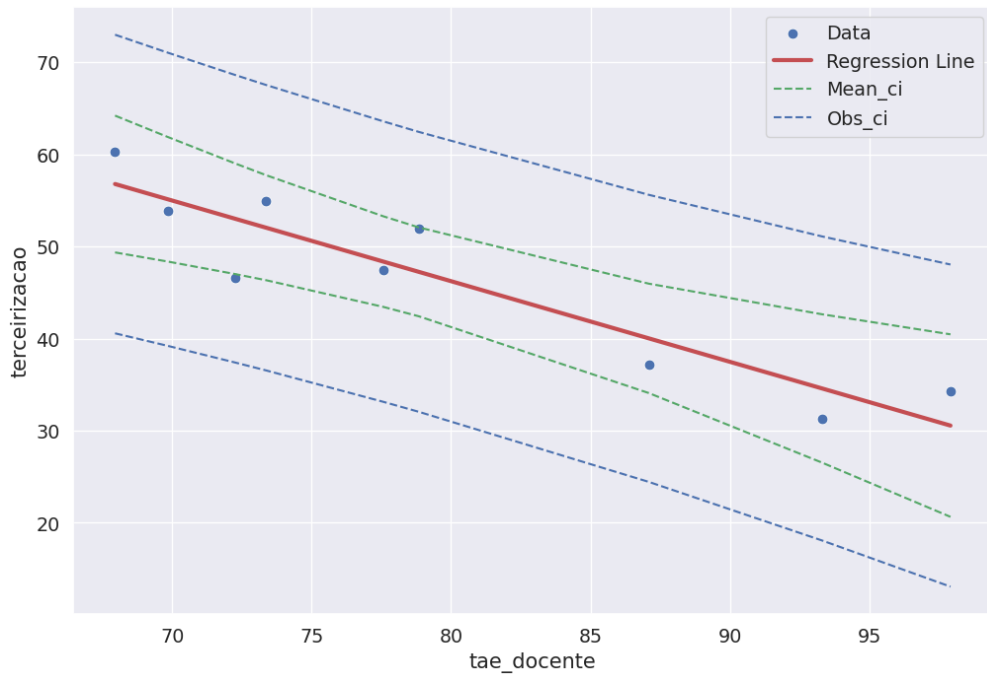


figura 12: Gráfico de regressão linear da variável “terceirizacao” em função da variável “tae\_docente”.

Os intervalos de confiança são úteis para entender a incerteza inerente à análise de dados e para avaliar a qualidade da linha de regressão e das estimativas derivadas dela. Eles podem ajudar a determinar se os dados são suficientes para fazer uma conclusão significativa ou se mais dados são necessários. Quando os intervalos são mais estreitos, é uma indicação que a relação entre as duas variáveis é mais confiável e que as previsões são mais precisas.

### 3.2.3.2. Modelo 2: Terceirizacao em função de tae\_estudante

A figura 13 apresenta o resumo dos resultados obtidos a partir do modelo de regressão linear desenvolvido com a variável independente "tae\_estudante" e a variável dependente "terceirizacao".

Model:	OLS	Adj. R-squared:	0.395
Dependent Variable:	terceirizacao	AIC:	64.3568
Date:	2023-04-17 23:13	BIC:	64.7513
No. Observations:	9	Log-Likelihood:	-30.178
Df Model:	1	F-statistic:	6.221
Df Residuals:	7	Prob (F-statistic):	0.0413
R-squared:	0.471	Scale:	61.544

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
<b>Intercept</b>	223.9242	71.2174	3.1442	0.0163	55.5218	392.3266
<b>tae_estudante</b>	-31.5098	12.6332	-2.4942	0.0413	-61.3826	-1.6371

Omnibus:	1.468	Durbin-Watson:	0.527
Prob(Omnibus):	0.480	Jarque-Bera (JB):	1.016
Skew:	0.659	Prob(JB):	0.602
Kurtosis:	2.013	Condition No.:	158

Figura 13: Resumo dos resultados obtidos a partir do modelo de regressão simples. Variável “terceirizacao” em função da variável “tae\_estudante”.

A partir da figura 13 é possível ver que o modelo gerou um coeficiente de determinação  $R^2$ , de 0,471. Ou seja, a variável que representa a proporção de TAEs em relação à quantidade de estudantes de graduação explica 47,1% da variação no percentual de gastos com terceirização.

Ao analisar a figura 13 percebe-se que o parâmetro estimado do beta, referente à variável “tae\_estudante”, mostrou-se estatisticamente diferente de zero ao nível de significância de 5%, uma vez que a magnitude do seu erro-padrão resultou em um valor-P ( $P > |t| < 0,05$ ). Em outras palavras, o resultado do teste  $t$  dos resultados apresentados pela figura 13 nos permite afirmar que a variável explicativa (“tae\_estudante”) é estatisticamente significativa para explicar o comportamento da variável dependente (“terceirizacao”). Como o nível de significância observado é 0,0413, valor inferior a 0,05, o teste  $t$  apresenta rejeição da hipótese nula, o que nos permite concluir, ao nível de confiança de 95%, que a variável “tae\_estudante” tem efeito significativo sobre a variável “terceirizacao”.

A figura 14 mostra o gráfico do modelo de regressão linear gerado para a variável “terceirizacao” em função da variável “tae\_estudante”. Cada ponto no gráfico (cor azul) representa um par de valores das duas variáveis. A linha vermelha é a linha de regressão, que representa a melhor estimativa da relação entre as variáveis. Essa linha foi determinada através do método estatístico OLS, que minimiza a distância entre a linha de regressão e todos os pontos do gráfico.

O gráfico da figura 14 mostra ainda os intervalos de confiança para a linha de regressão (Mean\_ci) e para os dados observados (Obs\_ci). Indicando uma incerteza muito maior deste modelo com a variável independente “tae\_estudante” que o modelo gerado com a variável independente “tae\_docente”.

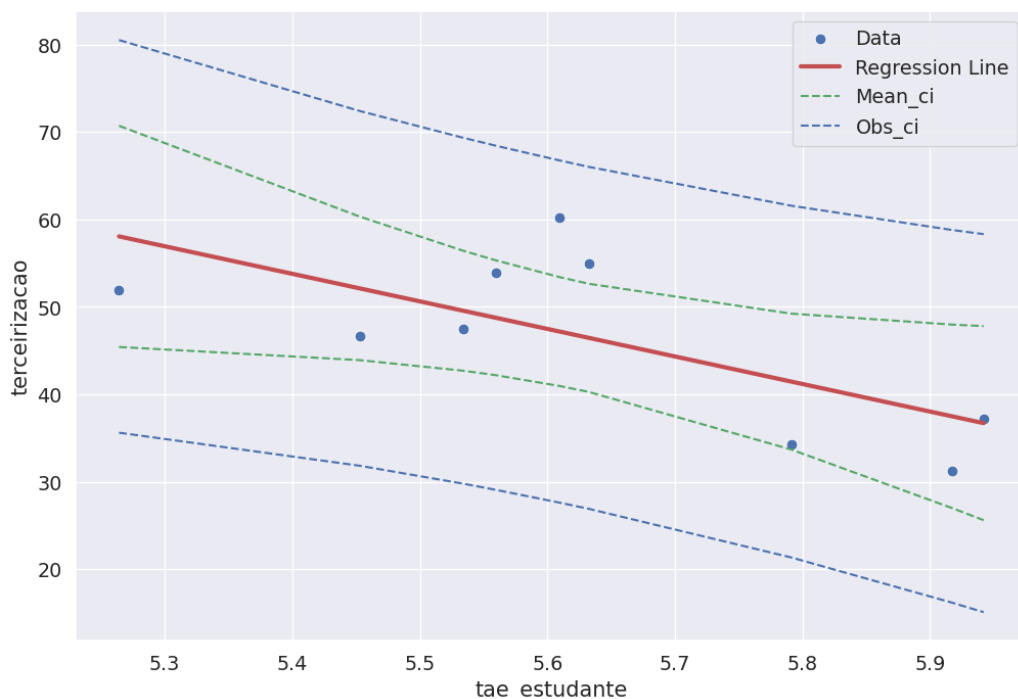


Figura 14: Gráfico de regressão linear da variável “terceirizacao” em função da variável “tae\_estudante”.

Ao comparar os resultados dos modelos de regressão linear simples para a variável "terceirizacao" em relação a "tae\_docente" e "tae\_estudante", observa-se que a variável "tae\_docente" tem uma capacidade explicativa muito maior sobre o comportamento da variável "terceirizacao" do que a variável "tae\_estudante". A justificativa para essa conclusão baseia-se nas comparações entre os valores de  $R^2$ , valores-P ( $P < 0,01$  e  $< 0,05$ , respectivamente) e na análise gráfica das figuras 13 e 14, que demonstram uma incerteza significativamente menor para o modelo com a variável “tae\_docente” como variável independente.

### 3.2.3.3. Modelo 3: Terceirizacao em função de tae\_docente e tae\_estudante

Após as análises de regressão simples apresentadas acima, construiu-se um modelo de regressão múltipla com a variável "terceirizacao" como dependente e as variáveis "tae\_docente" e "tae\_estudante" como independentes. O resumo dos resultados obtidos a partir desse modelo de regressão múltipla é apresentado na figura 15.

Model:	OLS	Adj. R-squared:	0.827
Dependent Variable:	terceirizacao	AIC:	53.7000
Date:	2023-04-17 23:16	BIC:	54.2917
No. Observations:	9	Log-Likelihood:	-23.850
Df Model:	2	F-statistic:	20.12
Df Residuals:	6	Prob (F-statistic):	0.00218
R-squared:	0.870	Scale:	17.595

	Coef.	Std.Err.	t	P> t	[0.025	0.975]
<b>Intercept</b>	150.8884	41.6961	3.6188	0.0111	48.8618	252.9149
<b>tae_docente</b>	-0.7767	0.1806	-4.2994	0.0051	-1.2187	-0.3346
<b>tae_estudante</b>	-7.5448	8.7576	-0.8615	0.4220	-28.9739	13.8844

Figura 15: Resumo dos resultados obtidos a partir do modelo de regressão múltipla. Variável “terceirizacao” em função das variáveis “tae\_docente” e “tae\_estudante”.

A partir da figura 15 é possível ver que o modelo gerou um coeficiente de determinação  $R^2$ , de 0,870. Ou seja, as variáveis “tae\_docente” e “tae\_estudante” juntas explicam 87% da variação no percentual de gastos com terceirização.

Em análise de regressão múltipla, a estatística F (*F-statistic*) é uma medida utilizada para avaliar a significância global do modelo de regressão, ou seja, se o grupo de variáveis independentes possui significância estatística para explicar o comportamento da variável dependente (Hair et al., 2009).

A hipótese nula da estatística F é que não há relação linear significativa entre as variáveis independentes e a variável dependente, ou seja, todos os coeficientes de regressão (*beta*) das variáveis explicativas são iguais a zero. Em outras palavras, a hipótese nula afirma que o modelo de regressão não explica a variação na variável dependente, e qualquer relação observada nos dados é devida ao acaso. Se o valor da estatística F for menor que um nível de significância pré-determinado (por exemplo, 0,05), então rejeitamos a hipótese nula e concluímos que ao menos uma variável explicativa é estatisticamente significativa para explicar a variabilidade da variável dependente.

É importante ressaltar que a estatística F deve ser analisada em conjunto com os valores-P ao decidir sobre a significância do modelo como um todo (Hair et al., 2009).

Se o valor da estatística F for significativo e alguns ou todos os valores-P das variáveis individuais também forem significativas, pode-se concluir que o modelo geral é significativo e que as variáveis individuais estão contribuindo para o poder explicativo do modelo. Por outro lado, se a estatística F for significativa, mas nenhum dos valores-P das variáveis individuais for significativo, isso pode sugerir que o modelo está mal especificado ou que as variáveis apresentam multicolinearidade -- correlações elevadas entre as

variáveis explicativas --, e pode ser necessário investigar mais a fundo (Hair et al., 2009).

De acordo com a figura 15, o valor da estatística F do modelo é estatisticamente significativo a um nível de significância de 5%, o que leva à rejeição da hipótese nula, uma vez que o valor 0,00218 é menor do que 0,05. Isso significa que, de acordo com a estatística F, o modelo é capaz de explicar a variação da variável dependente. Além disso, a figura 15 mostra que a variável "tae\_estudante" tem um valor-P estatisticamente significativo de 0,0051, enquanto o valor-P da outra variável não é significativo, com um valor de 0,4220. Deste modo, pode-se concluir que o modelo geral possui significância estatística para explicar o comportamento da variável terceirização.

Ao contrário da regressão linear simples, que ajusta uma reta aos dados, a regressão linear múltipla ajusta um hiperplano aos dados. Isso ocorre porque, enquanto a regressão linear simples lida com duas variáveis - uma variável dependente e uma variável independente - a regressão linear múltipla lida com duas ou mais variáveis independentes, e uma variável dependente. Nesse caso, a relação entre as variáveis é modelada por um hiperplano, que é uma superfície de dimensões maiores que uma reta.

Devido ao fato de o Modelo 3 ser composto por três variáveis - duas independentes e uma dependente - é possível exibir o gráfico do modelo de regressão por meio de uma visualização em 3 dimensões.

A figura 16 mostra o gráfico do modelo de regressão linear do Modelo 3. Cada ponto no gráfico (cor azul) representa um ponto tridimensional das variáveis "tae\_docente", "tae\_estudante" e "terceirizacao". O hiperplano vermelho representa a melhor estimativa da relação entre as variáveis, obtido através do método estatístico OLS, que minimiza a distância entre o hiperplano de regressão e os pontos do espaço tridimensional. A superfície verde mostra os intervalos de confiança para o hiperplano de regressão e a superfície azul mostra os intervalos de confiança para os dados observados.

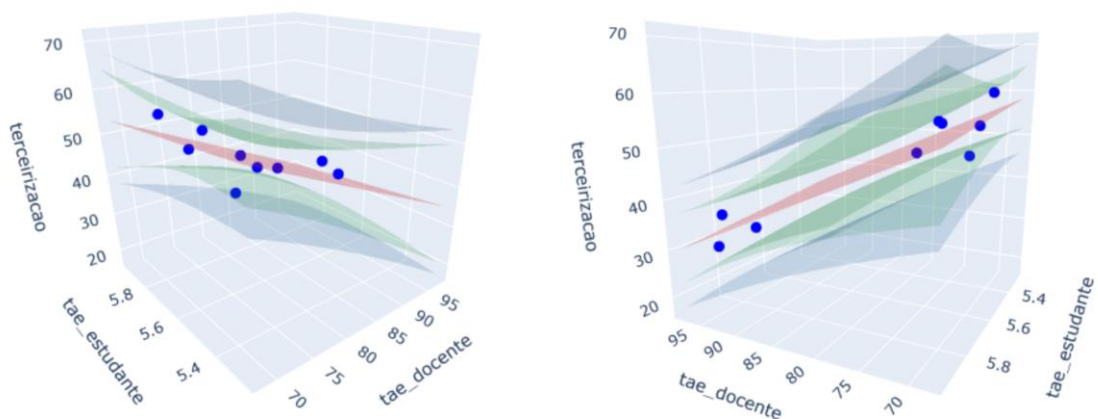




Figura 16: Visualizações do gráfico tridimensional de regressão linear múltipla da variável “terceirizacao” em função das variáveis “tae\_docente” e “tae\_estudante”.

De acordo com Fávero e Belfiore (2017), a avaliação da significância conjunta das variáveis explicativas pela estatística F não permite a identificação das variáveis específicas que apresentam significância estatística no modelo. Portanto, de acordo com os autores, é necessário que o pesquisador verifique individualmente se cada parâmetro do modelo de regressão é estatisticamente diferente de zero, para determinar se a variável correspondente deve ser incluída no modelo final proposto. Fávero e Belfiore (2017) complementam que a não rejeição da hipótese nula para o parâmetro *beta* a determinado nível de significância, deve indicar que a correspondente variável independente não se correlaciona com a variável dependente e, portanto, deve ser excluída do modelo final.

Fávero e Belfiore (2017) indicam a aplicação do procedimento *Stepwise* sobre modelos de regressão múltipla. Esse procedimento apresenta a capacidade de excluir ou manter automaticamente os parâmetros *beta* no modelo com base nos critérios estabelecidos, resultando em um modelo final com apenas os parâmetros *beta* estatisticamente significativos para um determinado nível de significância.

A figura 17 apresenta o resultado da aplicação do procedimento *Stepwise*, configurado para um valor-P limite de 0,05, sobre o modelo de regressão múltipla “terceirizacao” em função de “tae\_docente” + “tae\_estudante”.

```
Regression type: OLS

Estimating model...:
terceirizacao ~ tae_docente + tae_estudante

Discarding attribute "tae_estudante" with p-value equal to 0.4220462608387328

Estimating model...:
terceirizacao ~ tae_docente

No more attributes with p-value higher than 0.05

Attributes discarded on the process...:

{'attribute': 'tae_estudante', 'p-value': 0.4220462608387328}

Model after stepwise process...:
terceirizacao ~ tae_docente
```

Figura 17: Resultado da aplicação do procedimento *Stepwise* sobre o modelo de regressão múltipla “terceirizacao” em função de “tae\_docente” + “tae\_estudante”.

Como pode ser observado na figura 17, o procedimento *Stepwise* removeu o atributo “tae\_estudante” do modelo de regressão, visto que este apresentava um valor-P de 0,422, valor superior a 0,05. Após a aplicação do procedimento *Stepwise*, conclui-se que o modelo de regressão simples com a variável independente “tae\_docente” e a variável dependente “terceirizacao” deve ser

mantido, mesmo que esse modelo apresente um valor de  $R^2$  menor do que o modelo de regressão múltipla. Isso se deve ao fato de que o modelo de regressão final apresenta maior significância estatística (menor valor de estatística F do modelo e menor valor-P da variável explicativa “tae\_docente”), apesar de apresentar um coeficiente de ajuste ( $R^2$ ) menor.

Segundo Fávero e Belfiore (2017), a eliminação de parâmetros que não sejam estatisticamente significativos através do método *Stepwise* pode levar a problemas de especificação do modelo devido à exclusão de uma variável que poderia ser relevante para explicar o comportamento da variável dependente, caso não houvesse outras variáveis explicativas no modelo final. Fávero e Belfiore (2017) complementam sobre a importância da aplicação de técnicas para verificação de possíveis erros de especificação do modelo final. Para verificar a existência deste problema, o modelo foi submetido ao teste de Breusch-Pagan para verificação de heterocedasticidade. O resultado desse teste, juntamente com outros necessários para garantir a validade da significância estatística do modelo final, são apresentados na seção seguinte.

### 3.3. Verificação dos pressupostos para validade dos modelos de regressão

A validade dos modelos de regressão por OLS depende do atendimento aos seguintes pressupostos (Fávero e Belfiore, 2017; Wooldridge, 2019):

- Linearidade: a relação entre as variáveis deve ser linear.
- Normalidade: os erros (ou resíduos) devem apresentar uma distribuição normal.
- Ausência de multicolinearidade: as variáveis independentes não devem estar altamente correlacionadas.
- Homocedasticidade: a variância dos erros deve ser constante para todas as observações.
- Independência: os resíduos devem ser independentes uns dos outros, para modelos temporais.

A não observância desses pressupostos pode resultar em modelos de regressão viesados e imprecisos (Fávero e Belfiore, 2017; Wooldridge, 2019). Por isso, para que sejam considerados válidos, é necessário verificar o cumprimento de cada um dos pressupostos mencionados.

#### 3.3.1. Linearidade

A análise de regressão pressupõe a existência de uma relação linear entre as variáveis independentes e a variável dependente, ou seja, uma relação que pode ser representada por uma reta de ajuste. Caso a relação seja não linear, outras técnicas de modelagem devem ser utilizadas (Montgomery et al., 2012).

Os resultados obtidos por meio da análise visual dos diagramas de dispersão e os valores do coeficiente de correlação de Pearson, apresentados na seção de análise descritiva dos dados, evidenciam que a relação entre as variáveis

independentes ("tae\_docente" e "tae\_estudante") e a variável dependente ("terceirizacao") é consideravelmente linear. Os coeficientes de correlação de Pearson entre "tae\_docente" e "terceirizacao", e entre "tae\_estudante" e "terceirizacao", foram de -0,92 e -0,69, respectivamente. Esses valores indicam uma relação linear negativa entre as variáveis, com o primeiro par apresentando uma correlação bem mais significativa do que o segundo.

### 3.3.2. Normalidade

De acordo com Montgomery et al. (2012), a análise de regressão requer que os resíduos da análise apresentem uma distribuição normal. Neste sentido, Fávero e Belfiore (2017) afirmam que a normalidade dos resíduos é requerida para que sejam validados os testes de hipótese dos modelos de regressão, ou seja, o pressuposto da normalidade assegura que o valor-P dos testes do teste F sejam válidos.

Fávero e Belfiore (2017) afirmam que o teste de Shapiro-Wilk é um teste de normalidade univariada, aplicável em amostras com tamanho entre 4 e 2000 elementos, sendo assim adequado para a presente análise, que conta com 9 registros.

O teste de Shapiro-Wilk assume as seguintes hipóteses (Fávero e Belfiore, 2017):

- Hipótese nula: a amostra provém de uma população com distribuição normal.
- Hipótese alternativa: a amostra não provém de uma população com distribuição normal.

Para a avaliação da normalidade do resíduos foi considerado um nível de significância de 5%. A figura 18 mostra os resultados dos testes de Shapiro-Wilk para os três modelos de regressão construídos nesta análise.

```
print('terceirizacao ~ tae_docente: ',stats.shapiro(m_tae_docente.resid))
print('terceirizacao ~ tae_estudante: ',stats.shapiro(m_tae_estudante.resid))
print('terceirizacao ~ tae_docente + tae_estudante: ',stats.shapiro(m_multiplo.resid))

terceirizacao ~ tae_docente: ShapiroResult(statistic=0.9212020039558411, pvalue=0.4022851586341858)
terceirizacao ~ tae_estudante: ShapiroResult(statistic=0.8797516226768494, pvalue=0.15607845783233643)
terceirizacao ~ tae_docente + tae_estudante: ShapiroResult(statistic=0.9304954409599304, pvalue=0.4860783815383911)
```

Figura 18: Resultado da aplicação do teste de Shapiro-Wilk sobre os modelos de regressão.

A figura 18 exibe os valores-P de 0,402, 0,156 e 0,486 para os modelos 1, 2 e 3, respectivamente. Uma vez que os valores-P estão acima do valor crítico de 0,05, a hipótese nula do teste não foi rejeitada. Ou seja, os testes mostraram que os resíduos dos três modelos de regressão possuem distribuição compatíveis com a distribuição normal.

### 3.3.3. Ausência de multicolinearidade

A multicolinearidade ocorre quando há alta correlação entre duas ou mais variáveis independentes em um modelo de regressão, o que pode dificultar a identificação das contribuições individuais dessas variáveis para a variável dependente. Este é um problema para a regressão linear porque ela pode

levar a resultados imprecisos ou até mesmo enganosos, visto que as estimativas dos coeficientes de regressão podem ser instáveis e altamente sensíveis às mudanças nos dados de entrada (Wooldridge, 2019).

O método mais básico e comum para detectar a multicolinearidade é verificar a presença de altas correlações entre as variáveis independentes, por meio da análise da matriz de correlação. No entanto, embora seja fácil de aplicar, esse método é limitado em sua capacidade de detectar relações simultâneas entre mais de duas variáveis (Fávero e Belfiore, 2017).

Segundo Fávero e Belfiore (2017), a técnica estatística conhecida como *Variance Inflation Factor* (VIF) pode ser utilizada para diagnosticar a presença de multicolinearidade, por meio da realização de regressões auxiliares. O cálculo do VIF é baseado no coeficiente de determinação ( $R^2$ ) dessas regressões auxiliares, permitindo obter informações sobre as correlações simultâneas entre as variáveis analisadas. Não há um consenso absoluto sobre o valor ideal do VIF para se considerar a multicolinearidade em um conjunto de dados. No entanto, muitos autores sugerem que valores de VIF acima de 5 ou 10 indicam problemas significativos de multicolinearidade.

A Figura 9 apresenta o coeficiente de correlação de Pearson entre as variáveis "tae\_docente" e "tae\_estudante" (0,68), enquanto a Figura 10 exibe o valor-P para essa correlação (0,065). Observa-se uma correlação positiva entre as variáveis, no entanto, não há significância estatística, uma vez que o valor-P está acima do nível de significância estabelecido de 0,05. Na figura 19, é apresentado o valor da estatística VIF, que é de 1,68 para as variáveis.

```
X = df[['tae_docente', 'tae_estudante']]
X = add_constant(X)
vif = pd.Series([variance_inflation_factor(X.values, i) for i in range(X.shape[1])], index=X.columns)
print(vif)
```

const	889.311286
tae_docente	1.680944
tae_estudante	1.680944

Figura 19: Valores da estatística VIF para as variáveis "tae\_docente" e "tae\_estudante".

Com base nos resultados apresentados pelo valor da estatística VIF, pelo coeficiente de correlação de Pearson e pela falta de significância estatística dessa correlação, podemos concluir que não há uma multicolinearidade significativa entre as variáveis "tae\_docente" e "tae\_estudante". Portanto, podemos considerar que o Modelo 3 não é afetado por problemas de multicolinearidade. Os outros modelos desta análise não apresentam problemas de multicolinearidade, visto que possuem apenas uma variável explicativa.

#### 3.3.4. Homocedasticidade

Homocedasticidade é uma propriedade desejada em modelos estatísticos que indica que a variância dos erros de um modelo é constante para todos os valores das variáveis explicativas. Em outras palavras, os erros do modelo apresentam a mesma dispersão em torno da linha de regressão para todos

os níveis das variáveis independentes. A presença de homocedasticidade é importante para garantir a precisão e validade dos resultados de análise estatística e inferência (Fávero e Belfiore, 2017; Wooldridge, 2019).

De acordo com Fávero e Belfiore (2017) o teste de Breusch-Pagan é um teste de heterocedasticidade, ou seja, é utilizado para verificar se os erros do modelo de regressão são heterocedásticos, ou seja, se a variância dos erros não é constante ao longo dos valores da variável independente. Se os erros forem heterocedásticos, isso pode indicar que há variáveis importantes faltando no modelo. No entanto, é importante notar que a presença de heterocedasticidade não significa necessariamente que há omissão de variáveis. Outros fatores, como erros de medida, outliers ou não linearidades nas relações entre as variáveis, podem levar à heterocedasticidade (Wooldridge, 2019).

O teste de Breusch-Pagan assume as seguintes hipóteses (Fávero e Belfiore, 2017):

- Hipótese nula: a variância dos termos de erro é constante (erros homocedásticos).
- Hipótese alternativa: a variância dos termos de erro não é constante (erros heterocedásticos).

```
print('terceirizacao ~ tae_docente: ',breusch_pagan_test(m_tae_docente))
print('terceirizacao ~ tae_estudante: ',breusch_pagan_test(m_tae_estudante))
print('terceirizacao ~ tae_docente + tae_estudante: ',breusch_pagan_test(m_multiplo))
```

```
chisq: 0.02239770042211959
p-value: 5.271992052481361
terceirizacao ~ tae_docente: (0.02239770042211959, 5.271992052481361)
chisq: 0.025973008067250662
p-value: 4.886962080789545
terceirizacao ~ tae_estudante: (0.025973008067250662, 4.886962080789545)
chisq: 0.7715586802744775
p-value: 0.6176091791700338
terceirizacao ~ tae_docente + tae_estudante: (0.7715586802744775, 0.6176091791700338)
```

Figura 20: Valores do teste de Breusch-Pagan para os modelos de regressão.

A figura 20 exibe os valores-P de 5,27, 4,88 e 0,61 para os modelos 1, 2 e 3, respectivamente. Uma vez que os valores-P estão acima do valor crítico de 0,05, a hipótese nula do teste não foi rejeitada. Portanto, os testes mostraram que a variância dos termos de erro dos modelos é constante, portanto, homocedásticos. Deste modo, pode-se concluir que os modelos não apresentam falhas de especificação e não há evidências quanto à omissão de variáveis explicativas relevantes.

### 3.3.5. Independência

Os termos de erro em uma análise de regressão para uma série temporal devem ser independentes para garantir que a análise seja confiável e que os resultados obtidos sejam precisos (Fávero e Belfiore, 2017). A independência dos erros é importante porque se os erros forem correlacionados, isso pode

levar a uma análise incorreta dos dados. Por exemplo, se os erros estiverem correlacionados, a estimativa dos parâmetros da regressão pode ser tendenciosa e os intervalos de confiança podem ser muito estreitos ou muito amplos.

De acordo com Fávero e Belfiore (2017) o teste de Durbin-Watson é útil para verificar a independência dos erros em uma análise de regressão, sendo o teste mais utilizado por pesquisadores para esta finalidade.

O teste de Durbin-Watson assume as seguintes hipóteses (Fávero e Belfiore, 2017):

- Hipótese nula: Não há autocorrelação nos resíduos da regressão, o que significa que os erros são independentes.
- Hipótese alternativa: Existe autocorrelação nos resíduos da regressão, o que significa que os erros não são independentes.

```
print('terceirizacao ~ tae_docente: ',durbin_watson(m_tae_docente.resid))
print('terceirizacao ~ tae_estudante: ',durbin_watson(m_tae_estudante.resid))
print('terceirizacao ~ tae_docente + tae_estudante: ',durbin_watson(m_multiplo.resid))

terceirizacao ~ tae_docente: 2.4838336341429605
terceirizacao ~ tae_estudante: 0.5266305571769105
terceirizacao ~ tae_docente + tae_estudante: 1.908597482792241
```

Figura 21: Valores do teste de Durbin-Watson para os modelos de regressão.

A figura 21 exibe os valores-P de 2.48, 0.52 e 1.90 para os modelos 1, 2 e 3, respectivamente. Uma vez que os valores-P estão acima do valor crítico de 0,05, a hipótese nula do teste não foi rejeitada. Portanto, os testes mostraram que não há autocorrelação nos resíduos da regressão, o que significa que os erros são independentes nos três modelos testados.

## 4. Conclusão

Com base nas análises realizadas, é possível concluir que o aumento relativo dos gastos com terceirização de mão de obra na UFLA está relacionado à defasagem histórica no quantitativo de técnicos administrativos, tanto em relação ao número de docentes quanto ao número de estudantes de graduação. Esta conclusão é baseada nas descobertas resultantes da análise descritiva dos dados e das análises de regressão. Quando a proporção de TAEs sobre o número de docentes ou de TAEs sobre o número de estudantes diminui, os valores orçamentários empenhados para pagar mão de obra terceirizada aumentam. Entretanto, a variável que mais fortemente se relacionou com aumento de valores orçamentários empenhados para pagamento de mão de obra terceirizada foi a diminuição proporcional do número de TAEs sobre o número de docentes. Essa descoberta indica que, quando os valores de "tae\_docente" ou "tae\_estudante" diminuem, o valor de "terceirizacao" aumenta, com uma tendência muito mais forte para as variações ocorridas em "tae\_docente".

Tais relacionamentos lineares também ficaram evidentes ao analisar visualmente os diagramas de dispersão. Tais diagramas mostram um comportamento com tendência linear inversa das variáveis explicativas em relação à variável dependente.

Foram construídos três modelos utilizando técnicas de regressão linear para expressar de maneira quantitativa a relação entre as variáveis. Esses modelos buscaram explicar o comportamento da variável "terceirizacao", sendo que um deles utilizou a variável "tae\_docente" como explicativa (Modelo 1), outro utilizou a variável "tae\_estudante" como explicativa (Modelo 2) e um terceiro modelo utilizou ambas as variáveis como explicativas (Modelo 3), por meio da regressão múltipla.

Ao analisar os resultados dos modelos de regressão, verificou-se que as variáveis independentes exercem influência sobre o comportamento da variável dependente. Os Modelos 1, 2 e 3 apresentaram coeficientes de determinação ( $R^2$ ) de 85,5%, 47,1% e 87%, respectivamente. Além disso, todos os três modelos foram estatisticamente significativos e cumpriram todos os pressupostos necessários para validar os resultados da técnica de regressão empregada.

Embora o Modelo 3 tenha apresentado um  $R^2$  superior, após a aplicação do procedimento *Stepwise* a variável "tae\_estudante" foi descartada por não apresentar significância estatística na presença da variável "tae\_docente", regredindo o modelo de regressão múltipla ao Modelo 1. Deste modo, embora o Modelo 3 possa ser utilizado por atender aos pressupostos da técnica de regressão, o Modelo 1 apresenta-se mais estatisticamente significativo para explicar o comportamento da variável "terceirizacao".

Isso significa que a variável "tae\_docente" por si só é capaz de explicar 85,4% da variação no percentual de gastos com terceirização na UFLA, a um nível de confiança de 99%. A inclusão da variável "tae\_estudante" no modelo aumenta a capacidade explicativa da variável "terceirizacao" em 1,6 pontos percentuais. No entanto, esse aumento vem acompanhado de uma maior complexidade do modelo e da diminuição da significância estatística.

## 5. Referências

Fávero, L.P.L; Belfiore, P.P. **Manual de análise de dados: estatística e modelagem multivariada com excel, SPSS e stata**. Rio de Janeiro: Elsevier, 2017.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. **Multivariate data analysis (7th ed.)**. Prentice Hall. 2009.

Montgomery, D. C., Peck, E. A., & Vining, G. G. **Introduction to Linear Regression Analysis**. John Wiley & Sons, 2012.

Wooldridge, J. M. **Introductory econometrics: A modern approach (9th)**. Cengage Learning. 2019.