escola superior de tecnologia e gestão
instituto politécnico de leiria

Efeitos do *mismatch* educacional nos salários no mercado de trabalho português: uma análise dos Quadros de Pessoal (2010–2023)

Mestrado em Ciência de Dados

David Santos Oliveira

Leiria, setembro de 2025

Efeitos do *mismatch* educacional nos salários no mercado de trabalho português: uma análise dos Quadros de Pessoal (2010–2023)

Mestrado em Ciência de Dados

David Santos Oliveira

Trabalho de Projeto realizado sob a orientação do Professor Doutor Rui Filipe Vargas de Sousa Santos e da Professora Doutora Ana Sofia Patrício Pinto Lopes.

Leiria, setembro de 2025

Originalidade e Direitos de Autor

O presente relatório de projeto é original, elaborado unicamente para este fim, tendo sido devidamente citados todos os autores cujos estudos e publicações contribuíram para o elaborar.

Resumo

O presente trabalho examina a relação entre a qualificação dos trabalhadores e os rendimentos reais no mercado de trabalho português, centrando-se no fenómeno do *mismatch* educacional, nomeadamente a discrepância entre o nível de escolaridade do indivíduo e o requerido pela sua profissão, e analisa as suas implicações salariais. Utilizando dados longitudinais dos Quadros de Pessoal entre 2010 e 2023, a investigação combina duas abordagens metodológicas complementares: regressões lineares anuais com base em dados *cross-sectional*, focando mais nas diferenças salariais entre trabalhadores, e modelos de dados em painel com efeitos fixos, permitindo captar variações temporais e controlar a heterogeneidade não observada dos indivíduos.

As variáveis analisadas incluem fatores demográficos, de capital humano, empresariais, regionais, setoriais e contratuais, beneficiando da utilização dos Quadros de Pessoal, uma base de dados administrativa que cobre todos os trabalhadores do setor privado em Portugal. A dimensão desta base permite trabalhar com milhões de observações anuais, assegurando representatividade, bem como uma grande riqueza em termos de diversidade de variáveis disponíveis.

Os resultados evidenciam que a escolaridade, o *mismatch* educacional, a produtividade e a dimensão da empresa são determinantes salariais robustos, enquanto características como género apresentam efeitos mais complexos e dependentes do modelo estimado. A comparação entre métodos mostra que o modelo em painel tende a reduzir a magnitude dos coeficientes. Esta diferença decorre não apenas do controlo de heterogeneidade não observada através dos efeitos fixos, mas também do facto de o painel captar a evolução do rendimento do mesmo indivíduo ao longo do tempo, proporcionando uma análise dinâmica das relações entre variáveis.

Este trabalho contribui para o aprofundamento do conhecimento sobre desigualdades salariais associadas ao *mismatch* educacional em Portugal, fornecendo evidência empírica útil para o desenho de políticas públicas que promovam uma melhor adequação entre qualificações e funções laborais.

Palavras-chave: qualificação, salários, *mismatch* educacional, dados em painel, regressão linear, mercado de trabalho português.

Abstract

This study examines the relationship between workers' qualifications and real earnings in the Portuguese labor market, focusing on the phenomenon of educational mismatch, namely the discrepancy between an individual's level of education and that required for their occupation, and analyses its wage implications. Using longitudinal data from "Quadros de Pessoal" between 2010 and 2023, the research combines two complementary methodological approaches: annual linear regressions based on cross-sectional data, focusing more on wage differences between workers, and panel data models with fixed effects, which capture temporal variations and control for unobserved individual heterogeneity.

The variables analyzed include individual, firm, regional and sectoral-level factors, benefiting from the use of "Quadros de Pessoal", an administrative dataset that covers virtually all employees in the private sector in Portugal. The large size of this dataset allows for the use of millions of annual observations, ensuring representativeness, as well as a wide diversity of available variables.

The results show that education, educational mismatch, productivity and firm size are robust determinants of wages, while characteristics such as gender present more complex effects that depend on the model estimated. The comparison between methods shows that the panel model tends to reduce the magnitude of the coefficients. This difference arises not only from controlling unobserved heterogeneity through fixed effects, but also from the fact that the panel setting captures the evolution of the same individual's earnings over time, providing a dynamic analysis of the relationships between variables.

This study contributes to a deeper understanding of wage inequalities associated with educational mismatch in Portugal, providing empirical evidence that can inform public policies aimed at improving the alignment between qualifications and job requirements.

Keywords: qualification, earnings, educational mismatch, panel data, linear regression, Portuguese labor market.

Índice

Lista de figuras

Lista de tabelas

Lista de siglas e acrónimos

CAE	Classificação das Atividades Económicas
CPP	Classificação Portuguesa das Profissões
DGEEC	Direção-Geral de Estatísticas da Educação e Ciência
ESTG	Escola Superior de Tecnologia e Gestão
FCT	Fundação para a Ciência e a Tecnologia
GEP	Gabinete de Estratégia e Planeamento
GVIF	Generalized Variance Inflation Factor (Fator Generalizado de Inflação da Variância)
INE	Instituto Nacional de Estatística
IPC	Índice de Preços no Consumidor
MEA	Modelo de Efeitos Aleatórios
MEF	Modelo de Efeitos Fixos
MTSSS	Ministério do Trabalho, Solidariedade e Segurança Social
NA	Not Available / Dados omissos
NUTS	Nomenclatura das Unidades Territoriais para Fins Estatísticos
OLS	Ordinary Least Squares (Mínimos Quadrados Ordinários)
QP	Quadros de Pessoal
RGPD	Regulamento Geral sobre a Proteção de Dados
SPSS	Statistical Package for the Social Sciences
TeSP	Cursos Técnicos Superiores Profissionais
VIF	Variance Inflation Factor (Fator de Inflação da Variância)

1. Introdução

O presente estudo analisa a relação entre a qualificação dos trabalhadores e os seus rendimentos no mercado de trabalho português, com particular foco no fenómeno do *mismatch* educacional, nomeadamente a discrepância entre o nível de escolaridade do indivíduo e o requerido pela sua profissão. A investigação recorre a dados administrativos provenientes dos Quadros de Pessoal, disponibilizados pelo Instituto Nacional de Estatística (INE) e recolhidos e tratados pelo Gabinete de Estratégia e Planeamento do Ministério do Trabalho, Solidariedade e Segurança Social (MTSSS), abrangendo o período de 2010 a 2023. Esta base de dados, não só possibilita obter informação acerca de todos os trabalhadores afetos ao setor privado em Portugal, como permite uma análise longitudinal e detalhada da evolução dos salários face às qualificações e a um vasto conjunto de características demográficas, profissionais e organizacionais.

Este estudo centra-se na temática do *mismatch* educacional, entendido como a ausência de correspondência entre o nível de escolaridade do trabalhador e o nível de escolaridade exigido para o exercício da sua profissão. Podem então verificar-se duas situações distintas: a *overeducation* (sobre-educação), quando o trabalhador possui um nível de escolaridade superior ao requerido, e a *undereducation* (sub-educação), quando o trabalhador apresenta um nível de escolaridade inferior ao associado à sua profissão.

A pertinência do tema decorre da relevância do *mismatch* educacional no contexto das transformações económicas, tecnológicas e demográficas, que alteram o equilíbrio entre oferta e procura de qualificações. Estudos anteriores indicam que tanto a *overeducation* como a *undereducation* afetam significativamente os rendimentos dos trabalhadores e a produtividade das empresas. Assim, este trabalho procura atualizar a evidência existente, explorando a realidade do mercado de trabalho português com um horizonte temporal alargado e métodos econométricos robustos.

O objetivo geral é identificar e quantificar o impacto do *mismatch* educacional, no salário mensal dos trabalhadores. Para alcançar este propósito, os objetivos específicos incluem:

- Construir uma base de dados longitudinal com dimensão e cobertura significativas, a partir dos Quadros de Pessoal entre 2010 e 2023, garantindo representatividade do mercado de trabalho português;
- Controlar fatores que podem influenciar a relação entre o *mismatch* educacional e os salários, incluindo características demográficas (idade, antiguidade, nacionalidade e género), fatores organizacionais (produtividade e dimensão da empresa), bem como variáveis regionais, setoriais e contratuais;
- Aplicar metodologias econométricas, nomeadamente regressões lineares anuais e modelos de dados em painel com efeitos fixos, de forma a comparar as estimativas obtidas e identificar padrões consistentes;
- Examinar a evolução temporal dos efeitos estimados e testar a robustez dos resultados, através da análise da estabilidade dos coeficientes e da aplicação de correções para heterocedasticidade.

A metodologia combina duas abordagens:

(i) Modelos de dados em painel, com efeitos fixos e aleatórios, selecionando o modelo apropriado via teste de Hausman, acompanhados de diagnóstico de multicolinearidade (VIF) e aplicando correção para heterocedasticidade com erros padrão robustos;

(ii) Regressões lineares (OLS) anuais, com diagnóstico de multicolinearidade (VIF) e heterocedasticidade (Breusch-Pagan), seguidas de correção robusta.

Este relatório de projeto está organizado da seguinte forma:

- **Capítulo 1** introduz o tema e enquadra a sua relevância;
- **Capítulo 2** apresenta o enquadramento teórico e a revisão da literatura;
- **Capítulo 3** descreve a preparação, limpeza e normalização dos dados;
- **Capítulo 4** apresenta a estatística exploratória dos dados com a descrição de algumas das variáveis mais relevantes;
- **Capítulo 5** desenvolve a análise com dados em painel;
- **Capítulo 6** aborda a análise *cross-section*, através de regressões lineares anuais;
- **Capítulo 7** compara os resultados das duas metodologias;
- **Capítulo 8** apresenta as conclusões, limitações e sugestões para investigação futura.

2. Enquadramento teórico e revisão de literatura

Este capítulo contextualiza e revê o fenómeno do *mismatch* educacional no mercado de trabalho. Com a crescente qualificação da mão de obra, impulsionada pela expansão do ensino superior, a análise da divergência entre as habilitações dos trabalhadores e as exigências dos seus empregos torna-se fundamental. Para isso, nesta secção abordam-se as principais teorias explicativas, os métodos de medição e as consequências deste fenómeno em domínios como os salários, a produtividade e a mobilidade.

A discussão abrange tanto a literatura internacional como a investigação específica sobre o mercado de trabalho português, servindo de base teórica e conceptual para a metodologia que será desenvolvida e aplicada nas seções subsequentes deste trabalho.

2.1. Fatores determinantes dos salários

A teoria do capital humano diz que os rendimentos do trabalho refletem a produtividade adquirida por meio de educação e experiência [1]. O modelo seminal de Mincer (1974) formalizou esta relação através da equação de rendimentos, na qual o salário (tipicamente em logaritmo) é expresso como função linear dos anos de escolaridade e de uma componente quadrática para a experiência profissional [2, 3]. Esta especificação simples tornou-se padrão na literatura acerca da análise de diferenças salariais, evidenciando de forma robusta que a escolaridade tem impacto positivo e significativo nos ganhos, enquanto a experiência contribui positivamente, mas com retornos decrescentes ao longo da vida laboral [3]. Em termos práticos, diversos estudos estimam que cada ano adicional de estudo se traduz, em média, num prémio salarial apreciável. Por exemplo, para a economia portuguesa, Sousa et al. (2015) estimaram um retorno de cerca de 10% por ano de estudo para homens e ligeiramente superior para mulheres (10,5%) no período 1986–2009 [4, 5]. No caso da experiência, Mincer demonstrou que os ganhos gerados pelos primeiros anos de carreira são maiores do que os proporcionados pelos anos adicionais [6]. Este efeito de salário-experiência alinha-se com a ideia de acumulação de capital humano: os trabalhadores aumentam a produtividade (e salário) com a idade e antiguidade no emprego, mas a um ritmo que abranda à medida que se aproxima o final da vida ativa [6].

Para além da educação e da experiência, uma série de outros fatores individuais e contextuais afetam o nível de vencimento. Do ponto de vista demográfico, destaca-se o efeito do género, onde é consistente a evidência de que, mesmo comparando trabalhadores com características semelhantes, as mulheres auferem salários inferiores aos homens. Esse diferencial de género (*gap*) tem sido estimado em cerca de 15% em desfavor das trabalhadoras [6], refletindo fatores como discriminação ou segregação ocupacional.

A nacionalidade também surge como fator relevante, onde trabalhadores imigrantes tendem a enfrentar penalizações salariais face aos nacionais. Em Portugal, por exemplo, estimou-se que, em média, os imigrantes ganhavam cerca de 14% menos do que os portugueses no período 2002–2008, sem controlar outros fatores [7]. Grande parte desta diferença resulta da concentração de estrangeiros em empregos de menor remuneração, bem como da menor "portabilidade" inicial do seu capital humano (educação e experiência obtidas no seu país de origem) no mercado de trabalho local [7, 8].

As características laborais e empresariais igualmente influenciam os salários, como é o caso do tipo de contrato de trabalho e a dimensão da empresa. Trabalhadores com contrato sem termo (efetivos) auferem, em média, um ligeiro prémio salarial em comparação com colegas em contratos a termo ou temporários (na ordem de +3% de vencimento) [6, 9], possivelmente refletindo maior segurança e acumulação de capital humano específico. As empresas de maior porte tendem a pagar remunerações mais elevadas do que as de menor dimensão, evidência essa consistente tanto em Portugal como noutros países. Estimativas para o mercado português confirmam este efeito, onde um trabalhador numa empresa média ganha aproximadamente 11% mais do que um trabalhador afeto a uma pequena empresas, diferença que sobe para cerca de +18% no caso de grandes empresas vs. microempresas [6]. Este prémio de dimensão pode estar associado a fatores como maior produtividade nas grandes firmas, estruturas ocupacionais complexas (com mais cargos bem remunerados) ou práticas salariais internas mais vantajosas.

Diferenciais intersetoriais e regionais estão também documentados. Por um lado, certos setores de atividade pagam consistentemente mais do que outros, tais como os setores da energia, petróleo e química, o setor financeiro e setores das áreas da tecnologia da informação. Por outro, setores mais tradicionais, tais como, têxteis, vestuário e calçado, madeira, hotelaria/restauração ou comércio a retalho, apresentam os salários mais baixos [6,

10]. A nível regional, observam-se disparidades regionais persistentes, com as áreas metropolitanas a oferecer salários médios superiores. Por exemplo, Lisboa e Vale do Tejo apresenta prémios salariais significativos em relação a outras regiões, cerca de +8% em média comparada com a região Centro, ao passo que no Norte os salários são cerca de 2% inferiores aos apresentados pela região Centro [6].

2.2. Definição de *mismatch* educacional (o*vereducation* versus *undereducation*) e teorias explicativas

O *mismatch* educacional ocorre quando há divergência entre a escolaridade do trabalhador e aquela que está associada ao desempenho da sua profissão/emprego. Especificamente, fala-se em *overeducation* (sobre-educação) quando o indivíduo tem um nível de escolaridade superior ao nível requerido para desempenhar adequadamente o seu trabalho, e em *undereducation* (sub-educação) no caso em que o nível de escolaridade do trabalhador é inferior ao que se associa à sua profissão [11]. Estudos estimam que entre 10% e 40% dos trabalhadores podem estar *overeducated* em diferentes países, dependendo do contexto e do método de medida [11].

É importante notar que o *mismatch* educacional aqui referido diz respeito a desajuste "vertical" entre anos de escolaridade e a escolaridade exigida na função/profissão desempenhada. De notar que existe ainda a possibilidade de *mismatch* "horizontal" que ocorrem quando há uma desadequação entre a área de formação do trabalhador e a que melhor se associa à profissão desempenhada, e o *skill mismatch.* É ainda possível falar de *over* e *underqualification* que correspondem a um desajuste de competências, ou seja, quando as competências de um indivíduo, incluindo a educação formal, mas também competências informais, como experiência de trabalho e formação contínua, são superiores ou inferiores às exigidas para a função [11]. Na presente investigação, o foco recai sobre o desajuste vertical com base em diferenças associadas ao nível de escolaridade, o que permite uma maior objetividade, sendo também o foco para a revisão da literatura.

Várias teorias económicas foram desenvolvidas no sentido de procurar explicar por que ocorrem situações de *overeducation* no mercado de trabalho, muitas vezes oferecendo perspetivas complementares. As quatro abordagens clássicas são: Teoria do Capital

Humano, Teoria da Concorrência por Emprego, Teoria da Afetação (*assignment*) e Teoria da Mobilidade Ocupacional/Carreira. A literatura agrupa-as frequentemente conforme encarem estes *mismatches* como fenómenos de curto ou longo prazo [11]. Abaixo resumem-se os pressupostos centrais de cada teoria.

2.2.1. Teoria do capital humano

A Teoria do Capital Humano de Becker (1964) assume que a escolaridade aumenta a produtividade e é plenamente utilizada pelas empresas [1]. Assim, eventuais situações de *overeducation* seriam transitórias e pouco relevantes, resultantes de informação imperfeita ou atritos na procura e oferta de trabalho [11]. Nesse enquadramento, trabalhadores inicialmente *overeducated* ajustam-se rapidamente: eles procuram empregos alinhados com as suas qualificações ou as empresas acabam por aproveitar as suas competências, de modo a maximizar a produtividade [11]. Em suma, para a literatura mais tradicional referente ao capital humano, espera-se que no equilíbrio de longo prazo não haja grandes desfasamentos, onde cada trabalhador tenderá a usar plenamente os seus conhecimentos adquiridos na sua formação, dado que tanto indivíduos quanto empresas tendem a beneficiar deste ajustamento que resulta na maximização de salários e produção [11].

2.2.2. Teoria da concorrência por emprego

O Modelo da Concorrência por Emprego de Thurow (1975) oferece uma visão distinta, destacando a alocação por filas de candidatos e vagas [12]. Nesta teoria, os empregos formam uma hierarquia de requisitos (filas de vagas ordenadas do mais ao menos qualificado) e os trabalhadores formam outra fila segundo as suas qualificações [11]. A educação funciona como mecanismo de classificação, onde indivíduos investem em educação não apenas pela produtividade, mas para se posicionarem mais à frente na fila por empregos mais bem remunerados [11]. Se a oferta de candidatos com maior qualificação ou escolaridade exceder a disponibilidade de cargos de alto nível, mesmo candidatos "no topo da fila" podem acabar em empregos de nível inferior (*overeducation*).

Assim, a *overeducation* pode persistir no longo prazo numa situação de excesso de oferta de habilitações, especialmente na ausência de criação de empregos altamente qualificados em número suficiente [11]. Diferente da visão do capital humano, aqui a responsabilidade recai na estrutura de emprego oferecido pelas empresas. Neste contexto, é importante que o

aumento da escolaridade que se tem verificado em Portugal nos últimos tempos seja acompanhado pelo aumento de ofertas de emprego associado a maiores níveis de qualificação.

2.2.3. Teoria da afetação (*assignment*)

A Teoria da Afetação de Sattinger (1993) insere-se num ponto intermédio entre as visões anteriores, reconhecendo que tanto as características dos trabalhadores quanto as dos empregos influenciam conjuntamente a alocação no mercado [13]. Nesta abordagem, assume-se que os trabalhadores têm diferentes preferências e habilidades, e as empresas oferecem vagas com diferentes exigências. A correspondência (*matching*) resulta de um processo que considera ambos os lados. Em primeiro lugar, os indivíduos escolhem setores ou profissões de acordo com preferências e incentivos e, em seguida, são alocados a empregos específicos conforme suas qualificações e outras características [11].

Os *mismatches* educacionais podem ser solucionados de duas formas [11]:

(a) Ajuste do trabalhador: O trabalhador pode procurar um novo emprego ou aceitar um salário que não seja totalmente ajustado ao seu nível de escolaridade.
(b) Ajuste da empresa: A empresa pode alterar as tarefas do trabalhador ou promovê-lo, para que a sua função se ajuste melhor às suas qualificações.

Ao contrário da teoria concorrente, aqui admite-se heterogeneidade de preferências entre indivíduos e diferenças entre setores e ocupações na possibilidade de *mismatch* [11, 12]. Segundo Dieter Verhaest e Eddy Omey (2010), esta teoria tende a explicar melhor a realidade, por incorporar características dos empregos e dos trabalhadores simultaneamente [11].

2.2.4. Teoria da mobilidade ocupacional (carreira)

Por fim, a Teoria da Mobilidade de Carreira (ou *Job Mobility*) de Sicherman e Galor (1990) propõe que a *overeducation* pode ser parte de uma estratégia ou etapa natural de progressão profissional [14]. Inicialmente, trabalhadores jovens ou sem experiência podem aceitar postos abaixo da sua qualificação como forma de adquirir experiência, sinalizar as suas competências aos empregadores ou "ganhar uma porta de entrada" na carreira [11]. Inspirada também pela ideia de sinalização de Spence (1973), esta teoria sugere que a informação imperfeita no recrutamento leva indivíduos qualificados a ocuparem temporariamente

posições aquém do seu nível, até conseguirem demonstrar plenamente as suas capacidades ou até surgir uma vaga apropriada [15]. Assim, a *overeducation* tende a ser um fenómeno de curto ou médio prazo concentrado em certas fases da carreira (principalmente no início). Com o acumular de experiência e a confirmação das habilidades, espera-se que o indivíduo progrida para um posto condizente, reduzindo o *mismatch* ao longo do tempo [16]. Ao contrário das abordagens anteriores, aqui o foco está no indivíduo como responsável por superar o *mismatch* (pela aquisição de experiência ou melhor comunicação das suas competências), não atribuindo tanto peso às características dos empregos disponíveis [11].

2.3. Métodos de medição do *mismatch* educacional

A forma como o *mismatch* é medido é crucial, pois diferentes métodos de medição podem levar a estimativas e conclusões distintas [11]. Há três abordagens principais usadas para determinar a escolaridade "requerida" de um emprego e classificar trabalhadores como *over* ou *undereducated.*

2.3.1. Método da análise do trabalho (*job analysis*)

Este método utiliza classificações ocupacionais para atribuir a cada profissão um nível de escolaridade padrão. Por exemplo, cada profissão é identificada e codificada através da classificação nacional de profissões elaborada pelo INE, sendo depois definida objetivamente a escolaridade típica de cada ocupação [16, 17]. Um trabalhador é, nesta abordagem, considerado *overeducated* se sua escolaridade exceder a atribuída à sua ocupação, e *undereducated* se estiver aquém.

O método da análise do trabalho é considerado "objetivo" por não depender da perceção dos envolvidos, mas pode ser limitado por generalizar requisitos dentro de ocupações amplas e por possivelmente estar desatualizado em relação às evoluções do trabalho.

2.3.2. Métodos de autoavaliação (*self-assessment*)

Baseiam-se na perceção dos próprios trabalhadores sobre as exigências de seus empregos. Há duas variantes:

(a) Direta: perguntando-se ao trabalhador qual o nível de escolaridade que considera necessário para desempenhar o seu trabalho.

(b) Indireta: perguntando ao trabalhador se se considera *overeducated* ou *undereducated* para suas funções atuais [16].

Este método tem a desvantagem de ser bastante subjetivo uma vez que a perceção de cada individuo pode naturalmente ser bastante diferente da dos restantes e por isso difícil de comparar.

2.3.3. Método empírico/estatístico (*realized matches*)

Aqui a escolaridade "requerida" é inferida através dos dados efetivos de escolaridade por ocupação numa amostra. Uma abordagem comum define o nível requerido como a escolaridade modal entre trabalhadores de determinada ocupação, classificando como *over* ou *undereducated* aqueles cuja escolaridade ultrapassava ou ficava abaixo desse nível [16]. Por exemplo, se na ocupação "técnico de laboratório" a modalidade modal de escolaridade é licenciatura, assume-se licenciatura como o requisito. Logo, técnicos com mestrado seriam *overeducated* e com o ensino secundário seriam *undereducated* [16]. Em Portugal este método foi utilizado, por exemplo, no estudo de Kiker, Santos e de Oliveira (1997) [18].

Outra variante, proposta por Verdugo & Verdugo (1989), utiliza a média de anos de escolaridade na ocupação e um desvio padrão, onde se considera existência de *match* educacional se a escolaridade de um determinado trabalhador estiver contida no seguinte intervalo: [média – desvio padrão, média + desvio padrão], considerando a média e o desvio padrão por ocupação. Se se encontrar abaixo do intervalo é classificado como *undereducated* e acima corresponde a *overeducated* [19].

2.3.4. Abordagem adotada

Cada método apresenta vantagens e limitações. As autoavaliações captam nuances do posto de trabalho e competências efetivamente usadas, mas sofrem de subjetividade e viés de resposta [20]. A análise do trabalho é objetiva e consistente para comparações amplas, mas pode desatualizar-se frente a mudanças tecnológicas e não reflete variações entre empresas ou dentro da mesma ocupação. Já os métodos empíricos refletem a realidade corrente do mercado, combinando oferta e procura, mas por isso mesmo alguns autores criticam-nos, argumentando que a escolaridade modal ou média de uma ocupação é resultado de um equilíbrio de mercado, não necessariamente um "requisito técnico" [21]. Assim, pode subestimar o fenómeno caso todo um setor esteja a influenciar habilitações, ou sobrestimar

caso haja escassez generalizada de qualificados. Além disso, o corte em 1 desvio padrão na abordagem média é arbitrário e a moda pode ignorar variações dentro da ocupação [21].

Neste trabalho, dada a disponibilidade dos Quadros de Pessoal, é viável aplicar uma abordagem de *realized match* empírico. A própria base de dados contém a escolaridade de cada trabalhador e a profissão exercida, o que permite calcular a média (e o desvio padrão) de anos de escolaridade para cada profissão. A nossa análise, no entanto, adota uma variação deste método para classificar o *mismatch* ao nível da profissão. Em vez de usar o desvio padrão da escolaridade dentro de cada profissão, como abordado por Verdugo et al. (1989), utilizámos os valores do 1.º e do 3.º quartis da distribuição dos desvios da escolaridade (anos de escolaridade) face à média de anos de escolaridade da respetiva profissão para definir os limiares [19]. Assim, um trabalhador é considerado *overeducated* se a diferença entre a sua escolaridade (número de anos de escolaridade) e a média de anos de escolaridade do seu grupo profissional for superior ao 3.º quartil dos desvios, *undereducated* se for inferior ao 1.º quartil dos desvios, e *match* (adequadamente qualificado) se estiver entre estes dois valores.

A opção por utilizar os quartis da distribuição em vez da média e do desvio padrão justifica-se por duas razões principais. Em primeiro lugar, os quartis são menos sensíveis a valores extremos, permitindo uma classificação mais robusta quando existem distribuições assimétricas de escolaridade dentro dos grupos profissionais [18, 20]. Em segundo lugar, esta abordagem assegura que a definição de qualificação é consistente e comparável entre profissão, reduzindo a probabilidade de enviesamento causado por variabilidade excessiva em profissões com menor dispersão de níveis de escolaridade [18]. Como consequência, a utilização desta metodologia tende a produzir resultados mais estáveis e representativos, evitando que casos atípicos distorçam a identificação de situações de *mismatch* educacional [18, 20].

2.4. Revisão da literatura sobre os efeitos do *mismatch*

Nas últimas décadas, uma vasta literatura empírica tem investigado as consequências do *mismatch* educacional para diversos fatores, incluindo salários, produtividade das empresas, mobilidade laboral e satisfação no trabalho, entre outros. A seguir sintetizam-se as principais pesquisas, dando conta do estado atual do conhecimento.

2.4.1. Impacto do *mismatch* nos salários

O impacto do *mismatch* educacional nos salários é um tema amplamente estudado, com evidências robustas desde os trabalhos pioneiros de Duncan & Hoffman (1981) [22]. A literatura converge para os seguintes pontos:

- *Overeducation*: Trabalhadores com mais escolaridade do que o exigido para a sua função (*overeducated*) recebem salários mais altos que os colegas com menos escolaridade na mesma posição. Contudo, ganham significativamente menos do que colegas com a mesma qualificação mas que estão em posições adequadas ao seu nível de escolaridade [14]. A "penalização salarial" pela *overeducation* é consistente, com o retorno de cada ano de escolaridade excedente sendo tipicamente metade do retorno de um ano de escolaridade exigido pelo posto [14]. Ou seja, embora os trabalhadores *overeducated* ganhem mais do que os seus colegas menos qualificados no mesmo posto, eles não são recompensados financeiramente de forma total pelo seu nível de educação. Por outras palavras, o seu salário é "penalizado" por não estarem a trabalhar numa posição que exija a sua formação completa.
- *Undereducation*: O efeito inverso ocorre com trabalhadores *undereducated*. Eles tendem a ganhar menos do que colegas adequadamente qualificados na mesma função, mas ainda assim auferem um prémio salarial por estarem em posições que exigem mais do que sua qualificação formal [23].

Estas evidências desafiam tanto a teoria do capital humano, que prevê retornos iguais para cada ano de escolaridade, quanto o modelo de Thurow, que sugere que a escolaridade excedente não traz ganhos salariais [12]. O padrão observado sugere que a determinação salarial é influenciada tanto pelo capital humano quanto pela alocação imperfeita dos trabalhadores.

Uma vertente de pesquisa importante foca a causalidade na relação entre o *mismatch* e os salários: o *mismatch* realmente causa uma penalização salarial ou a *overeducation* é um reflexo de características não observadas que já levariam a salários mais baixos? Estudos que usam dados em painel e controlam para heterogeneidade individual, como o de Bauer (2002) na Alemanha, demonstram que uma parte significativa da penalização salarial se deve a estas características não observadas [24]. No entanto, mesmo após este controlo, permanece um efeito causal estatisticamente significativo do *mismatch* educacional nos

salários. Isso indica que trabalhadores com o mesmo nível de escolaridade ganham menos quando estão em posições abaixo da sua qualificação, confirmando que o desajuste em si impõe um custo salarial real.

2.4.2. O *mismatch* educacional e a produtividade

A relação entre o *mismatch* educacional e a produtividade do trabalho a nível das empresas e da economia é um campo de estudo em crescimento. A suposição inicial é que um mau alinhamento de qualificações diminui a eficiência. No entanto, estudos recentes mostram que a realidade é mais complexa.

- *Overeducation*: Um estudo com dados relativos à economia belga, por Mahy, Rycx & Vermeylen (2015), revelou que a presença de trabalhadores *overeducated* pode, em média, aumentar a produtividade ao nível da empresa, enquanto a *undereducation* tende a reduzi-la [16]. Esse efeito positivo da *overeducation* é mais acentuado em empresas de alta tecnologia, que dependem de mão-de-obra qualificada ou que operam em ambientes de grande incerteza. Isso sugere que a escolaridade extra desses trabalhadores pode aumentar a adaptabilidade e o potencial de inovação da empresa. Pelo contrário, outras investigações admitem a possibilidade de se observar um impacto nulo ou negativo da *overeducation*, quando associada à desmotivação ou a uma alocação ineficiente de recursos humanos [25].
- *Undereducation*: A literatura mostra que *undereducation* tem um impacto negativo mais consistente na produtividade, especialmente em situação económica desfavorável [25]. Em contexto português, Rocha et al. (2025) encontram evidência de que o *mismatch* educacional compromete os resultados produtivos das empresas, concluindo que melhores alocações entre qualificações e funções poderiam elevar o desempenho agregado [26].

A literatura da última década sugere que os trabalhadores *overeducated* não são "desperdício de talento", mas pelo contrário, geram alguma produtividade adicional. No entanto, o mercado pode estar a incorrer num custo de oportunidade, já que esse potencial poderia ser mais bem aproveitado em posições mais alinhadas. Ao mesmo tempo, o *undereducation* consistente indica obstáculos de qualificação que afetam negativamente a produção das empresas [16].

2.4.3. Efeitos do *mismatch* educacional na mobilidade laboral

A relação entre o *mismatch* educacional e a mobilidade no mercado de trabalho é complexa e manifesta-se de diferentes formas, incluindo mudanças de função, de profissão, organização ou até de país.

- *Overeducation*: A evidência sugere que a *overeducation* está ligada a uma maior mobilidade laboral. Trabalhadores *overeducated* tendem a procurar novos empregos com mais frequência e apresentam taxas de rotatividade mais altas do que os trabalhadores bem alinhados [27]. Isso ocorre porque a *overeducation* gera geralmente insatisfação, levando os profissionais a procurarem oportunidades que melhor correspondam às suas qualificações [28]. Por exemplo, o estudo de Groeneveld & Hartog (2004) mostra que muitos trabalhadores *overeducated* conseguem, com o tempo, mudar para empregos mais adequados às suas qualificações e conhecimentos adquiridos, utilizando a posição inicial como um "degrau" para ascender na carreira [28]. De modo semelhante, alguns estudos longitudinais concluíram que uma fração substancial dos jovens *overeducated* consegue sair dessa situação após alguns anos de experiência, embora outra fração permaneça persistentemente *overeducated*, indicando heterogeneidade [29].
- *Undereducation*: A mobilidade de trabalhadores *undereducated* tende a ser diferente. Eles podem ter uma menor propensão a mudar de emprego, especialmente se já tiverem construído uma carreira estável. Em vez de mudarem de empresa ou ocupação, podem investir em educação adicional para preencher a lacuna de qualificações, como frequentar cursos ou formações [29].

A relação entre o *mismatch* e a emigração é um tema emergente. A evidência é mista, mas alguns estudos sugerem que a *overeducation* pode aumentar a probabilidade de emigração, pois trabalhadores mais qualificados podem procurar oportunidades no exterior que melhor correspondam ao seu perfil ou ofereçam salários mais altos [30]. No entanto, outros fatores, como as condições gerais do mercado de trabalho, parecem ter um peso ainda maior na decisão de emigrar.

A maioria dos estudos concorda que a *overeducation* está associada a uma maior movimentação no mercado de trabalho. Essa mobilidade é um mecanismo importante que

ajuda a corrigir o desajuste ao longo do tempo, embora não seja uma solução para todos, e alguns indivíduos possam ficar presos em situações de *overeducation* de longo prazo.

2.5.*Mismatch* educacional em Portugal: A perspetiva da investigação

A investigação sobre o *mismatch* educacional em Portugal, embora menos abundante do que a literatura internacional, reflete as suas principais conclusões e revela dinâmicas próprias do mercado de trabalho português.

2.5.1. Primeiros estudos e a evolução do fenómeno

O estudo pioneiro de Kiker, Santos e de Oliveira (1997) já confirmava a presença de *overeducation* e *undereducation* em Portugal no início dos anos 90, com impactos salariais significativos [18]. Um trabalhador *overeducated* ganhava menos do que um colega com o mesmo nível de escolaridade que estivesse num emprego com as qualificações adequadas. Por outro lado, um *undereducated* ganhava mais do que se estivesse num emprego compatível com o seu nível de escolaridade, quando comparado com trabalhadores com o mesmo nível de ensino, mas que desempenhavam funções adequadas à sua formação.

Nas décadas seguintes, Portugal assistiu a um grande aumento na escolarização da sua força de trabalho. Esta evolução, de acordo com o estudo de Pimenta & Pereira (2019), resultou numa redução drástica da *undereducation*, que era um problema histórico do país [31]. A *overeducation*, por outro lado, manteve-se relativamente controlada, com taxas abaixo da média europeia, sugerindo que a economia portuguesa conseguiu, em grande parte, absorver o aumento de trabalhadores qualificados.

2.5.2. Efeitos na mobilidade e nos salários

A evidência portuguesa também apoia a teoria de que o *mismatch* influencia a mobilidade laboral. Estudos como o de Cerejeira et al. (2007) mostraram que muitos jovens *overeducated* conseguem, ao longo do tempo, mudar de emprego ou ser promovidos, reduzindo o seu desajuste inicial [32].

No que toca aos salários, trabalhos mais recentes, como o de Araújo & Carneiro (2017), confirmam que o *mismatch* tem um efeito intrínseco nos salários, mesmo quando se controlam fatores individuais [32]. A penalização salarial para a *overeducation,* quando comparando indivíduos com a mesma escolaridade, é uma realidade em Portugal.

Estes estudos também indicam que crises económicas, como a crise financeira, podem agravar temporariamente a *overeducation*, levando trabalhadores qualificados a aceitar empregos abaixo das suas competências.

2.5.3. Disparidades regionais e setoriais

A investigação aponta para uma grande diversidade setorial e regional em Portugal no que respeita à existência de *mismatch* educacional. A *overeducation* tende a ser mais acentuada em regiões menos desenvolvidas ou em setores económicos tradicionais, que não oferecem tantos empregos que exijam qualificações elevadas. Por outro lado, as grandes áreas metropolitanas de Lisboa e do Porto e os setores mais modernos (tecnologia, saúde, educação) conseguem absorver melhor os trabalhadores qualificados [33].

Em suma, embora Portugal tenha feito progressos significativos na redução da *undereducation*, persistem bolsas de *mismatch* que têm efeitos negativos nos salários e na satisfação dos trabalhadores.

2.6. Modelos econométricos e a causalidade do *mismatch* educacional

A investigação recente sobre o *mismatch* educacional tem-se focado em melhorar a metodologia para garantir que as conclusões sejam robustas e reflitam uma aproximação real. A principal preocupação é controlar as características individuais que podem influenciar os resultados, nomeadamente as observáveis, tais como características demográficas, de capital humano, empresariais e contratuais, e as não observáveis (como o talento ou a motivação), através de modelos de efeitos fixos.

A solução mais comum para este problema é usar dados em painel, que acompanham os mesmos indivíduos ao longo do tempo. Através de modelos de efeitos fixos, é possível isolar o efeito do *mismatch* do efeito de características fixas do indivíduo. Um exemplo notório é o estudo de Bauer (2002) na Alemanha, que mostrou que o impacto da *overeducation* sobre os salários diminuiu quando se usam modelos de efeitos fixos, mas ainda assim permanece significativa [24]. Isso mostra que, embora parte da diferença salarial se deva a características individuais, o *mismatch* em si tem um impacto real.

Para além dos modelos de efeitos fixos, a literatura também recorre a outras técnicas para garantir a robustez das conclusões. Métodos como variáveis instrumentais ou modelos de

correção de seleção são usados para resolver problemas onde a causa e o efeito estão correlacionados com fatores não observados, o que distorceria os resultados. Adicionalmente, são realizados testes de robustez que incluem a utilização de diferentes medidas de *mismatch* e a análise de subgrupos, para confirmar que os resultados são consistentes [32].

A utilização destas metodologias mais sofisticadas alinha-se com as melhores práticas internacionais. Ao usar modelos em painel e técnicas de robustez, a investigação consegue obter estimativas mais sólidas sobre o impacto do *mismatch* nos salários e noutros resultados. A literatura confirma que o *mismatch* educacional é um fenómeno complexo, mas mensurável, com efeitos reais e causais que justificam a atenção de políticas públicas.

3. Preparação, limpeza e normalização dos dados

O presente capítulo descreve o processo de preparação da base de dados utilizada neste projeto, desde a aquisição até à sua consolidação e normalização, resultando numa estrutura única, coesa e apta para análise empírica. O objetivo é apresentar este processo de forma suficientemente detalhada para que possa ser replicado por outro investigador, assegurando transparência e reprodutibilidade científica.

O capítulo inicia-se com a caracterização da origem institucional dos dados, o processo formal de obtenção e as condições legais associadas à sua utilização, dada a sua natureza confidencial. Seguem-se as etapas técnicas e metodológicas adotadas para garantir a consistência e integridade dos dados ao longo dos catorze anos disponíveis (2010–2023). Incluem-se a padronização de nomes de ficheiros e variáveis, importação eficiente, gestão de memória, remoção de duplicados, harmonização de categorias e transformação de variáveis-chave.

3.1. Metodologia e ferramentas

A abordagem proposta neste projeto baseia-se na exploração da base de dados dos Quadros de Pessoal e utiliza uma combinação de modelos estatísticos para analisar a relação entre a qualificação dos trabalhadores, a sua profissão e os rendimentos no mercado de trabalho português. Apresentam-se, de seguida, as variáveis utilizadas, com particular foco para a definição rigorosa de *mismatch* educacional, que servirá como principal variável de interesse na nossa análise empírica.

3.1.1. Dados e amostra

O estudo é baseado nos dados dos Quadros de Pessoal, abrangendo o período entre 2010 e 2023. A base de dados, muito rica em informação acerca das características demográficas, profissionais, organizacionais e regionais relativas aos trabalhadores, tem uma enorme dimensão e é uma mais-valia relevante, pois permite análises representativas do mercado de trabalho português, tanto em perspetiva longitudinal como *cross-section*. No presente estudo, possibilita seguir mais de 930 mil trabalhadores por 13 ou mais anos, assegurando uma visão única sobre trajetórias individuais e dinâmicas salariais no setor privado.

3.1.2. Definição das variáveis

Com base na literatura apresentada anteriormente e nos dados disponíveis, o presente estudo utiliza as seguintes variáveis:

- Variável dependente: logaritmo do rendimento real mensal (*log_rganho_real*).
- Variáveis independentes:
    - Ao nível do trabalhador: sexo, nacionalidade, idade, anos de escolaridade, *match/mismatch* educacional (dividida nas categorias *undereducation*, *overeducation* e *match*), antiguidade na empresa, tipo de contrato;
    - Ao nível da empresa: setor de atividade (CAE), dimensão e localização (NUTS II);

Algumas destas variáveis, nomeadamente a produtividade da empresa e o rendimento médio por profissão, não resultam diretamente da literatura ou nos dados disponíveis. A sua inclusão justifica-se pelo papel potencial que desempenham na explicação das diferenças salariais:

- Ao nível da empresa: a produtividade permite captar a eficiência económica das empresas.
- Ao nível da profissão: o rendimento médio por profissão controla as disparidades salariais das diferentes ocupações, permitindo que a variável dependente seja interpretada como um desvio face ao rendimento médio da profissão.

3.1.3. Metodologia para a determinação do *mismatch educacional*

A determinação da variável que mede *mismatch* educacional segue o método *realized match*, previamente apresentado, adaptado de forma a ser robusto a valores atípicos. Primeiro, foi calculada a média de anos de escolaridade de todos os trabalhadores de cada profissão em Portugal (considerando a definição de cada profissão conforme a classificação das profissões portuguesas com 4 dígitos, INE, 2011). Em seguida, para cada trabalhador foi determinado o desvio de escolaridade, definido como o número de anos de escolaridade do trabalhador menos o número médio de anos de escolaridade da sua profissão. Posteriormente, considerando todos os trabalhadores (de todas as profissões), foram determinados os 1.º e 3.º quartis da distribuição dos desvios de escolaridade. Por fim, estabeleceu-se a seguinte classificação:

- *Overeducated*: quando a diferença entre o nível de escolaridade do trabalhador e o nível de escolaridade médio da sua profissão se encontra acima do 3.º quartil da distribuição global dos desvios de escolaridade.
- *Undereducated*: quando a diferença entre o nível de escolaridade do trabalhador e o nível de escolaridade médio da sua profissão está abaixo do 1.º quartil.
- *Match*: quando o desvio de escolaridade do trabalhador se situa entre o 1.º e o 3.º quartil.

Esta abordagem, baseada no método de *realized match* via média dos anos de escolaridade em cada profissão, segue práticas reconhecidas na literatura para medir *educational mismatch*, permitindo comparabilidade com estudos prévios e garantindo maior estabilidade na classificação ao longo do tempo.

3.1.4. Análise estatística e diagnósticos

Para analisar a relação entre o *mismatch* educacional e os rendimentos, o estudo recorreu a modelos de dados em painel (efeitos fixos e aleatórios), com a seleção do modelo mais adequado feita através do teste de Hausman. Adicionalmente, foram realizadas regressões lineares por ano. A robustez dos modelos foi assegurada por meio de diagnósticos como o VIF (para a identificação de problemas de multicolinearidade) e o teste de Breusch-Pagan (para heteroscedasticidade), com a aplicação de erros padrão robustos para corrigir a heteroscedasticidade em todas as estimações.

3.1.5. Ferramentas de análise

Todo o processo de limpeza e de análise estatística foi realizada utilizando a linguagem R (version 4.5.0) no ambiente R Studio [34, 35], com os pacotes indicados na Tabela 1.

Tabela 1 - Lista de pacotes R Studio

Categoria	Pacote	Funções principais usadas
Manipulação e transformação de dados	*dplyr [36]*	*filter(), mutate(), summarise(), group_by()*
	tidyr [36]	*pivot_wider(), pivot_longer()*
	data.table [37]	*Filtragem e junções rápidas*
Leitura e escrita de ficheiros	*readr [36]*	*read_csv()*
	openxlsx [38]	*write.xlsx(), read.xlsx()*
	haven [36]	*read_sav() (ficheiros SPSS)*
Modelação estatística	*stats (base) [35]*	*lm(), predict(), residuals()*
	lmtest [39]	*bptest(), dwtest(), coeftest()*
	car [40]	*vif()*
	sandwich [41]	*vcovHC()*
	nortest [42]	*ad.test()*
	plm [43]	Modelos FE/RE, *phtest()*
Visualização	*ggplot2 [36]*	Gráficos customizados
	scales [36]	Escalas e formatação de eixos

3.2. Origem e processo de aquisição dos dados

Os dados utilizados neste projeto foram disponibilizados pelo Instituto Nacional de Estatística (INE) no âmbito de um pedido formal de acesso a microdados confidenciais. Esta cedência integra-se no Protocolo INE–FCT–DGEEC (FCT – Fundação para a Ciência e a Tecnologia; DGEEC – Direção-Geral de Estatísticas da Educação e Ciência), que regula o acesso a dados protegidos para fins de investigação científica.

A informação tem como fonte o inquérito anual Quadros de Pessoal, cuja recolha é da responsabilidade do Ministério do Trabalho, Solidariedade e Segurança Social (MTSSS), através do seu Gabinete de Estratégia e Planeamento (GEP). O INE, enquanto entidade

coordenadora do Sistema Estatístico Nacional, procede à anonimização, harmonização e disponibilização dos dados para efeitos de investigação.

O pedido de acesso foi submetido a 11 de setembro de 2024 através do sistema de acreditação *online*. O aluno e os respetivos orientadores assinaram os formulários de cedência e as declarações de compromisso com o Código de Conduta. A credencial atribuída restringe a utilização dos dados à realização do presente projeto, vinculando os envolvidos ao cumprimento do segredo estatístico, bem como às obrigações de armazenamento seguro e destruição final dos ficheiros após o termo da investigação.

3.3. Descrição inicial dos dados fornecidos

Os ficheiros disponibilizados pelo INE estão totalmente anonimizados, não contêm identificadores pessoais diretos e continuam abrangidos pelo regime de confidencialidade previsto no Regulamento Geral de Proteção de Dados (RGPD) e na Lei do Sistema Estatístico Nacional.

No total, foram fornecidos 42 ficheiros, totalizando cerca de 11.5 GB, no formato SPSS (Statistical Package for the Social Sciences), com a extensão .sav, correspondentes a:

- 14 anos de dados sobre empresas (informação económica, dimensão, localização, setor de atividade, entre outros);
- 14 anos de dados sobre estabelecimentos (setor de atividade, número de trabalhadores, localização, entre outros);
- 14 anos de dados sobre trabalhadores (remuneração, idade, sexo, escolaridade, vínculo contratual, entre outros).

No seu conjunto, os dados representam cerca de 42.5 milhões de registos individuais. Cada trabalhador está ligado a uma empresa através do identificador comum EMP_ID, o que permite cruzar características do empregador com o perfil do trabalhador. Importa salientar que os Quadros de Pessoal abrangem todos os trabalhadores por conta de outrem do setor privado em Portugal, o que corresponde cerca de 3 milhões de trabalhadores em cada ano analisado (varia entre aproximadamente 2.650.000 e 3.650.000 entre os 14 anos analisados). A dimensão e riqueza da base permitem uma análise profunda da evolução do mercado de trabalho em Portugal ao longo do período 2010–2023.

3.3.1. Estrutura dos dados

O INE disponibilizou 42 ficheiros SPSS (.sav), cuja distribuição se apresenta na Tabela 2. Cada ano do período 2010-2023 é constituído por três ficheiros:

Tabela 2 - Distribuição dos ficheiros entregues pela INE

Universo	Nº de ficheiros	Anos	Variáveis por ficheiro
Empresas	14	2010-2023	30
Estabelecimentos	14	2010-2023	21
Trabalhadores	14	2010-2023	42

Os ficheiros com os dados das Empresas (Tabela 3) contêm métricas financeiras anuais, setor de atividade, classe de dimensão, região, ano de constituição e outros, permitindo caracterizar a empresa.

Tabela 3 - Ficheiro de Empresas - Anonimizado

Nome da variável	Designação da variável	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
ANO	Ano de referência	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NPC_FIC	NPC_FIC	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NUEMP	Número da empresa (ligação com a série QP até 2009)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
EMP_ID	ID da empresa (ano 2010 e seguintes)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ancon	Ano/mês de constituição	x	x	x	x	x	x	x							
ANO_CONSTITUICAO	Ano de constituição								x	x	x	x	x	x	x
antiguidade	Antiguidade da empresa (anos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
esc_antig_emp	Escalão de antiguidade da empresa (anos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
natju	Natureza jurídica	x	x	x	x	x	x	x	x	x	x	x	x	x	x
csoc	Capital Social (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
cspri	Capital Social privado nacional (%)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
cspub	Capital Social público nacional (%)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
csest	Capital Social estrangeiro (%)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
csocesc	Escalão do Capital Social	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nut1_emp2013	NUT I da empresa em 31 de outubro (NUT2013)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nut2_emp2013	NUT II da empresa em 31 de outubro (NUT2013)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nut1_emp	NUT I da empresa em 31 de outubro (NUT2024)													x	x
nut2_emp	NUT II da empresa em 31 de outubro (NUT2024)													x	x
caem1	Actividade Económica da empresa (CAE_Rev.3 - 1 letra)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CAE2	Actividade Económica da empresa (CAE_Rev.3 - 2 dígitos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
CAE4_COD	Actividade Económica da empresa (CAE_Rev.3)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
vn	Volume de Negócios (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
vn_ano	Ano de referência do Volume de Negócios	x	x	x	x	x	x	x	x	x	x	x	x	x	x
vndesc1	Escalão de Volume de Negócios	x	x	x	x	x	x	x	x	x	x	x	x	x	x
vndesc2	Escalão de Volume de Negócios	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nest	Número de estabelecimentos	x	x	x	x	x	x	x	x	x	x	x	x	x	x
pemp	Número de pessoas ao serviço da empresa (31 de outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
trabsind	Número de trabalhadore sindicalizados (31 de outubro)													x	x
escdim1	Escalão de dimensão da empresa (31 de outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
escdim2	Escalão de dimensão da empresa (31 de outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
pempl	Número de pessoas ao serviço da empresa (outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
escdim_linhas1	Escalão de dimensão da empresa (outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
escdim_linhas2	Escalão de dimensão da empresa (outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
tcoemp	Número de trabalhdores por conta de outrem (TCO) (outubro)	x	x	x	x	x	x	x	x	x	x	x	x	x	x

No caso dos dados nos ficheiros dos trabalhadores (Tabela 4) reportam, para cada indivíduo, variáveis como sexo, idade, habilitações literárias, grupo profissional, tipo de contrato, remuneração mensal bruta (entre outros).

Tabela 4 - Ficheiro de Trabalhadores - Anonimizado

Nome da variável	Designação da variável	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
ANO	Ano de referência	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NPC_FIC	Número de Identificação Ficticio da empresa	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NUEMP	Número da empresa (ligação com a série QP até 2009)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
EMP_ID	ID da empresa (ano 2010 e seguintes)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
NUEST	Número do estabelecimento (ligação com a série QP até 2009)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ESTAB_ID	ID do estabelecimento (ano 2010 e seguintes)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
reg_reforma	Regime de reforma aplicado													x	x
ntrab	Número do Trabalhador	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nacio	Nacionalidade	x	x	x	x	x	x	x	x	x	x	x	x	x	x
sexo	Sexo	x	x	x	x	x	x	x	x	x	x	x	x	x	x
idade_Cod	Idade (idade_Cod <= 17 ou idade_Cod >=68)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
dt_adm	Data de admissão (mês ano)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
antig	Antiguidade na empresa	x	x	x	x	x	x	x	x	x	x	x	x	x	x
dt_prom	Data da última promoção (mês ano)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
IRCT_FICTICIO	Instrumento de Regulamentação Colectiva do Trabalho Fictício	x	x	x	x	x	x	x	x	x	x	x	x	x	x
IRCT_LABEL	IRCT Label	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ctpro	Categoria profissional	x	x	x	x	x	x	x	x	x	x	x	x	x	x
aplic_irct_ru	Aplicabilidade do IRCT													x	x
sitpro	Situação na profissão	x	x	x	x	x	x	x	x	x	x	x	x	x	x
tipo_contr	Tipo de contrato	x	x	x	x	x	x	x	x	x	x	x	x	x	x
tipo_contr1	Tipo de contrato (nível 1)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
reg_dur	Regime de duração do trabalho	x	x	x	x	x	x	x	x	x	x	x	x	x	x
habil1	Habilitações literárias (1 dígito)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
habil2	Habilitações literárias (2 dígitos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
habil	Habilitações literárias	x	x	x	x	x	x	x	x	x	x	x	x	x	x
nqual1	Nível de qualificação	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prof_4d	Classificação Portuguesa de Profissões (CPP 2010 - 4 dígitos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prof_3d	Classificação Portuguesa de Profissões (CPP 2010 - 3 dígitos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prof_2d	Classificação Portuguesa de Profissões (CPP 2010 - 2 dígitos)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prof_1d	Classificação Portuguesa de Profissões (CPP 2010 - 1 dígito)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
ctrem	Controle de remuneração	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rbase	Remuneração base paga (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rganho	Remuneração ganho (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
rextra	Remuneração suplementar (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prest_reg	Prestações regulares (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
prest_irreg	Prestações irregulares (euros)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
esc_rem_base	Escalão de remuneração mensal base	x	x	x	x	x	x	x	x	x	x	x	x	x	x
esc_rem_ganho	Escalão de remuneração mensal ganho	x	x	x	x	x	x	x	x	x	x	x	x	x	x
reminf_mot1	Remuneração base<devida - motivo 1													x	x
reminf_mot2	Remuneração base<devida - motivo 2													x	x
reminf_mot3	Remuneração base<devida - motivo 3													x	x
pnt	Período normal de trabalho semanal (PNT) (horas)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
hnormais	Horas mensais remuneradas - normais (horas)	x	x	x	x	x	x	x	x	x	x	x	x	x	x
hextra	Horas mensais remuneradas - suplementares (horas)	x	x	x	x	x	x	x	x	x	x	x	x	x	x

Em ambos os universos, a chave *EMP_ID* identifica de forma única a empresa, possibilitando a junção entre características da organização e o perfil do respetivo trabalhador. Essa relação foi crucial para a construção da base de dados analisado neste Projeto. Para o caso dos dados nos ficheiros dos estabelecimentos, optou-se por não utilizar, dado que grande parte da informação aí presente se encontra já reportada nos ficheiros das empresas, resultando em variáveis redundantes para os objetivos deste estudo. Adicionalmente, várias das variáveis específicas dos estabelecimentos não eram relevantes para a análise realizada ou apresentavam muitos valores omissos (*missing values)*. A inclusão destes ficheiros teria como consequência um aumento significativo do volume de dados a processar, tornando os cálculos mais pesados sem acrescentar valor analítico. Por estas razões, a informação mais detalhada utilizada restringe-se aos 28 ficheiros considerados, assegurando simultaneamente abrangência e eficiência no tratamento da base de dados.

3.4.Organização pré-importação

3.4.1. Renomeação dos ficheiros SPSS

Os ficheiros recebidos pelo INE apresentavam nomes extensos e o padrão variava entre os anos. Assim, e com o objetivo de simplificar, os nomes foram alterados manualmente para o padrão *Empresas_YYYY.sav* e *Trabalhadores_YYYY.sav*, (em que *YYYY* corresponde ao ano) facilitando a automação da leitura em lote.

3.4.2. Verificação de tipos

Antes da importação, a consistência tipológica das variáveis nos diferentes ficheiros foi avaliada, confirmando que cada campo mantinha o tipo (numérico, carácter ou fator). Sempre que surgiram discrepâncias, foi efetuado o registo para harmonização posterior. Destacam-se quatro casos:

- A variável do ano de constituição passou de *ancon* (formato AAAAMM) para *ANO_CONSTITUICAO* (formato AAAA) após 2017 (ver Tabela 3).
- Foi acrescentado *trabsind*, que contabiliza os trabalhadores sindicalizados por empresa (em 2022).
- Em 2023 todos os nomes de variáveis surgiram em maiúsculas; sendo normalizados diretamente no SPSS antes da importação pela linguagem R.
- Em 2023 introduziu-se o novo identificador territorial *NUT1_EMPRESA*, exigindo mapeamento para o esquema geográfico utilizado nos anos anteriores.

Este mapeamento precoce de anomalias evitou erros de coerção de tipos e simplificou as etapas subsequentes de limpeza e integração dos dados.

3.5. Importação e conversão

Cada ano segue o mesmo "triplo ciclo" importar → converter → limpar. Os ficheiros foram importados para o R Studio, do formato SPSS, individualmente por ano e por categoria (empresa e trabalhador). Após a importação, a união (*merge*) entre ficheiros do mesmo ano (empresa – trabalhador) foi feita a partir da variável *EMP_ID* (comum a ambos os ficheiros). À medida que eram efetuados os *merges*, os ficheiros ficavam mais pesados, sendo depois aplicados os comandos *rm()* e *gc()* para libertar memória RAM.

Na Figura 1 está representado um exemplo desse código.

```
Empresas_YYYY <- read_sav(".../Empresas_YYYY.sav")
Trabalhadores_YYYY <- read_sav(".../Trabalhadores_YYYY.sav")
Emp_Trab_YYYY <- merge(Empresas_YYYY, Trabalhadores_YYYY, by = "EMP_ID")
rm(Empresas_YYYY, Trabalhadores_YYYY); gc()
```

Figura 1 - Código de importação (exemplo)

As colunas *haven_labelled* foram codificadas em *factor*. Esta medida preserva a semântica das categorias, evitando perdas de informação aquando da gravação no formato RData (Figura 2).

```
Emp_Trab_YYYY_Convert <- as.data.frame(lapply(Emp_Trab_YYYY,
    function(col) {if (inherits(col, "haven_labelled")) as_factor(col) }
    else { col}}))
```

Figura 2 - Código de conversão SPPS para R (exemplo)

3.6. Limpeza, filtragem e normalização dos dados

Este subcapítulo descreve o processo de preparação da base de dados para os estudos subsequentes. Este processo incluiu a limpeza, correção, eliminação de duplicados, reorganização e normalização das variáveis.

3.6.1. Limpeza e filtragem dos dados

Após a conversão e união dos ficheiros por ano e por categoria (empresas e trabalhadores), foi efetuado um processo sistemático de limpeza, que incluiu a remoção de dados não conformes, eliminação de duplicados e padronização de variáveis. Importa salientar que uma base de dados com a dimensão dos Quadros de Pessoal tende a apresentar algumas incorreções ou inconsistências, pelo que estas etapas de tratamento foram essenciais para garantir a fiabilidade e a coerência da informação utilizada na análise. A regra geral de limpeza foi feita pelos seguintes passos (na Figura 3 está representado um exemplo do código limpeza *loop*.):

1. **Normalização do ano de constituição:** Entre 2010 e 2016, a variável *ancon* apresentava o ano e mês no formato AAAAMM. Esta foi renomeada para *ANO_CONSTITUICAO* (designação utilizada após 2016), extraindo-se apenas o ano (via divisão por 100 e arredondamento por defeito a um número inteiro).

2. **Eliminação de variáveis redundantes:** Foram eliminadas colunas duplicadas como *ANO.y*, *NPC_FIC.y* e *NUEMP.y*, uma vez que estas informações já estavam presentes noutras colunas. Nos anos de 2022 e 2023, campos adicionais como *nut1_emp2022*, *trabsind*, *reg_reforma* ou *aplic_irct_ru* foram descartados por não estarem presentes nos anos anteriores.
3. **Remoção de códigos inválidos:** Foram removidos registos com valores de *ntrab* inferiores a 10 000, considerados inválidos por não respeitarem a estrutura esperada.
4. **Eliminação de duplicados completos:** Usou-se a função *!duplicated()* para excluir observações exatamente iguais.
5. **Eliminação de duplicados parciais:** Para casos com *ntrab* repetido, mas com atributos divergentes, optou-se pelas seguintes soluções:
    I. Se os registos tinham o mesmo sexo e idade, foi mantido o registo com maior antiguidade e maior número de horas semanais (*pnt*).
    II. Se existiam discrepâncias nestes atributos essenciais, todos os registos desse trabalhador foram eliminados, por não ser possível garantir fiabilidade.

```
Emp_Trab_YYYY_Clean <- rename(Emp_Trab_YYYY_Clean,
    ANO_CONSTITUICAO = ancon)
Emp_Trab_YYYY_Clean$ANO_CONSTITUICAO <-
    floor(Emp_Trab_YYYY_Clean$ANO_CONSTITUICAO/100)
Emp_Trab_YYYY_Clean <- Emp_Trab_YYYY_Clean[,
    !(names(Emp_Trab_YYYY_Convert) %in%
    c("ANO.y", "NPC_FIC.y", "NUEMP.y"))]
Emp_Trab_YYYY_Clean <- Emp_Trab_YYYY_Clean[, !(names(Emp_Trab_YYYY_Clean)
    %in% c("nut1_emp2022", "nut2_emp2022", "trabsind", "reg_reforma",
    "aplic_irct_ru"))]
Emp_Trab_YYYY_Clean <- Emp_Trab_YYYY_Convert[Emp_Trab_YYYY_Clean$ntrab
    >= 10000, ]
Emp_Trab_YYYY_Clean <-
    Emp_Trab_YYYY_Clean[!duplicated(Emp_Trab_YYYY_Clean), ]
Emp_Trab_2018_Clean <- Emp_Trab_2018_Clean %>%
    group_by(ntrab) %>% filter(all(sexo == first(sexo) & idade_Cod ==
    first(idade_Cod)) | n() == 1) %>% arrange(ntrab, desc(pnt),
    desc(antig)) %>% distinct(ntrab, sexo, idade_Cod, .keep_all = TRUE)
    %>% ungroup()
```

Figura 3 - Código do *loop* limpeza (exemplo)

3.6.2. Consolidação inter-anual

Após a limpeza e junção por ano, procedeu-se à fusão inter-anual dos dados. Os conjuntos limpos foram unidos sequencialmente (via *rbind*) por ordem cronológica, começando por unir os anos 2010 e 2011, juntando depois à base 2010-2011 o ano 2012 e assim sucessivamente até 2023. Após cada etapa de fusão, foram guardadas versões intermédias (.RData) como *backups* de segurança.

De notar que, à medida que o volume de dados aumentava, a performance do RStudio era afetada, sendo necessário proceder frequentemente à limpeza do ambiente de trabalho para evitar interrupções ou falhas.

Na Figura 4 está representado um exemplo desta junção dos diferentes anos.

```
Emp_Trab_Clean_2010_ate_2023 <- rbind(Emp_Trab_Clean_2010_ate_2022,
    Emp_Trab_2023_Clean)
rm(Emp_Trab_2023_Clean)
rm(Emp_Trab_Clean_2010_ate_2022)
gc()
save(Emp_Trab_Clean_2010_ate_2023, file =
    ".../Emp_Trab_Clean_2010_ate_2023.RData")
```

Figura 4 - Código da junção dos dados por ano (exemplo)

3.6.3. Normalização de listas de categorias e títulos

As variáveis constantes na base de dados apresentavam inconsistências ao longo dos anos, incluindo diferenças em espaçamento, acentuação, prefixos e outros. Essas variações dificultavam, por exemplo, a correta agregação dos dados e a construção de tabelas de frequência coerentes.

Para garantir uniformidade, todas as variáveis foram revistas manualmente, seguindo um processo sistemático de normalização (Tabela 5). Este passo foi essencial para evitar duplicação de categorias e garantir que todas as variáveis fossem interpretadas corretamente ao longo do tempo.

Tabela 5 - Ações de revisão para normalização

Inconsistência	Ação
Espaçamentos	Remoção de espaços no início e fim dos rótulos
Prefixos	Eliminação de números ou letras no início do texto
Valores em branco	Substituição por "NA" padronizado para representar dados em falta (*missing values*)
Ordem dos níveis	Reorganização lógica dos fatores (por exemplo, ordenar categorias de regiões ou setores segundo uma hierarquia coerente, em vez da ordem alfabética automática)
Títulos incompletos ou com erros	Correção manual para garantir consistência

Na Figura 5 está representado um exemplo do código de normalização.

```
levels(Emp_Trab_Clean_2010_ate_2023$prof_3d) <-
    str_trim(str_replace(levels(Emp_Trab_Clean_2010_ate_2023$prof_3d),
    "^[0-9]+\\s*", ""))
levels(Emp_Trab_Clean_2010_ate_2023$prof_3d)
    [levels(Emp_Trab_Clean_2010_ate_2023$prof_3d) == ""] <- "NA"
Emp_Trab_Clean_2010_ate_2023$esc_antig_emp <-
    ordered(Emp_Trab_Clean_2010_ate_2023$esc_antig_emp,
    levels=c("Menos de 1 ano", "1 a 4 anos", "5 a 9 anos",
    "10 a 19 anos","20 a 49 anos", "50 e mais anos","Ignorado"))
levels(Emp_Trab_Clean_2010_ate_2023$vndesc2)
    [levels(Emp_Trab_Clean_2010_ate_2023$vndesc2) == "50000 -499999
    milhares de euros"] <- "50000 - 499999 milhares de euros"
```

Figura 5 - Código de "normalização"

3.6.4. Base de dados consolidada

Concluídas as etapas de limpeza, filtragem e normalização, obteve-se a primeira versão consolidada da base de dados, composta por 68 variáveis e com 41.433.063 observações. Este ficheiro foi guardado com o nome *Emp_Trab_Clean_2010_ate_2023*.RData, com uma dimensão aproximada de 2,7 GB (2.716.869 KB).

Esta versão representa um marco fundamental na construção da base analítica, permitindo avançar para as etapas seguintes do estudo com dados fiáveis, consistentes e organizados de forma adequada à aplicação de modelos econométricos em dados em painel.

3.7. Adição de novas variáveis para estudo

Com o objetivo de enriquecer a base de dados e facilitar a análise empírica, foram criadas variáveis que permitem observar fenómenos relevantes de forma mais direta e estruturada. Estas variáveis adicionais foram criadas para simplificar a modelação, melhorar a interpretação dos resultados e alinhar os dados com os objetivos específicos do estudo.

Neste subcapítulo, descrevem-se em detalhe todas as variáveis criadas, bem como os critérios e motivações que justificaram a sua inclusão.

3.7.1. Remuneração real ajustada à inflação (IPC)

Uma das primeiras variáveis criadas foi a remuneração real (*rganho_real*), obtida através da correção nominal (*rgranho*) pela inflação, com base no Índice de Preços no Consumidor (IPC). O IPC é um indicador económico através do qual é possível medir a variação média dos preços de um conjunto de bens e serviços consumidos pelas famílias, sendo utilizado como medida da evolução do custo de vida e da inflação. A utilização do IPC permite ajustar os valores monetários para eliminar o efeito da inflação ao longo do tempo. Os dados do IPC utilizados neste trabalho foram obtidos no site oficial do Instituto Nacional de Estatística (INE), adotando o ano de 2012 como ano base (IPC = 1 em 2012). Esta transformação foi igualmente aplicada a todas as outras variáveis financeiras, nomeadamente o volume de negócios total (*vn*) e ao volume de negócios anual (*vn_ano*), dando origem às variáveis ajustadas *vn_real* e *vn_ano_real*. Na Figura 6 é representado um exemplo do código na criação variável IPC (*ipc_data*) e das variáveis económicas ajustadas à inflação (*rganho_real*, *vn_real* e *vn_ano_real*).

A motivação principal para a criação destas variáveis foi garantir que as comparações salariais e económicas ao longo do período 2010–2023 fossem apenas relativas a variações reais, assegurando uma análise mais fidedigna e comparável ao longo do tempo.

```
ipc_data <- data.frame(
    ANO = c(2010:2023),
    IPC_100 = c(0.93872, 0.97302, 1, 1.00274, 0.99996, 1.00483, 1.01094,
    1.02477, 1.03496, 1.03846, 1.03833, 1.05147, 1.13383, 1.18271))
ETC_2010_2023_NovasColunas <- Emp_Trab_Clean_2010_ate_2023 %>%
    left_join(ipc_data, by = "ANO") %>%
    mutate(rganho_real = rganho / IPC_100,
         vn_real = vn / IPC_100, vn_ano_real = vn_ano / IPC_100)
```

Figura 6 - Código da criação das variáveis reais

3.7.2. Anos de escolaridade e cálculo das variáveis *overeducation*, *undereducation e match*

Com o objetivo de operacionalizar o conceito central deste estudo: o desalinhamento (*mismatch*) entre qualificação académica e a associada à profissão exercida, foi, em primeiro lugar, necessário transformar a variável nível de habilitações (*habil2*), correspondente aos diferentes níveis de escolaridade, numa variável quantitativa (*anos_escolaridade*), que refletisse o número aproximado de anos de estudo associados a cada grau de ensino.

A correspondência entre os níveis de habilitação e os anos de escolaridade foi estabelecida com base nos percursos académicos típicos do sistema de ensino português, conforme a Tabela 6.

Tabela 6 - Lista do nível de habilitações e anos de escolaridade associados

Nível de Habilitação	Anos de escolaridade
Inferior ao 1.º ciclo do ensino básico	0
1.º ciclo do ensino básico	4
2.º ciclo do ensino básico	6
3.º ciclo do ensino básico	9
Ensino secundário	12
Ensino pós-secundário não superior nível IV	13
Curso técnico superior profissional (TeSP)	14
Bacharelato	15
Licenciatura	15
Mestrado	17
Doutoramento	20

Na Figura 7 é representado o código de criação da variável que representa o número de anos de escolaridade por habilitação (*anos_escolaridade*).

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(anos_escolaridade = case_when(
    habil2 == "Inferior ao 1.° ciclo do ensino básico" ~ 0,
    habil2 == "1.° ciclo do ensino básico" ~ 4,
    habil2 == "2.° ciclo do ensino básico" ~ 6,
    habil2 == "3.° ciclo do ensino básico" ~ 9,
    habil2 == "Ensino secundário" ~ 12,
    habil2 == "Ensino pós secundário não superior nível IV" ~ 13,
    habil2 == "Curso técnico superior profissional" ~ 14,
    habil2 == "Bacharelato" ~ 15,
    habil2 == "Licenciatura" ~ 15,
    habil2 == "Mestrado" ~ 17,
    habil2 == "Doutoramento" ~ 20,
    TRUE ~ NA_real_ ))
```

Figura 7 - Código da criação variável *anos_escolaridade*

A partir da variável *anos_escolaridade*, calcularam-se duas métricas fundamentais por grupo profissional (*prof_4d*):

1. *media_ano_esc_prof*: correspondente à média dos anos de escolaridade dos trabalhadores dentro de cada profissão.
2. *desvio_media_ano_esc_prof*: correspondente à diferença entre a escolaridade de cada trabalhador e a escolaridade média da sua profissão.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  group_by(prof_4d) %>%
  mutate(media_ano_esc_prof = mean(anos_escolaridade, na.rm = TRUE)) %>%
  ungroup()
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  group_by(prof_4d) %>%
  mutate(desvio_media_ano_esc_prof = anos_escolaridade -
  mean(anos_escolaridade, na.rm = TRUE)) %>%
  ungroup()
```

Figura 8 - Código da criação variável *media_ano_esc_prof* e *desvio_media_ano_esc_prof*

Na Figura 8 está representado o código para a criação das variáveis média dos anos de escolaridade por profissão (*media_ano_esc_prof*) e diferença entre anos de escolaridade do trabalhador e a média dos anos de escolaridade por profissão (*desvio_media_ano_esc_prof*):

A determinação deste desvio é bastante relevante, pois permite identificar se um trabalhador se encontra numa situação de *overeducation* (nível de escolaridade acima da média da sua profissão) ou *undereducation* (abaixo da média da sua profissão). Para operacionalizar esta lógica, foi calculado o sumário estatístico da variável *desvio_media_ano_esc_prof*, tendo-se obtido os seguintes valores (Tabela 7):

Tabela 7 - Resultados estatísticos da variável *desvio_media_ano_esc_prof*

Mínimo	1º Quartil	Mediana	Media	3º Quartil	Máximo	NA's
-15,41	-1,80	0,11	0,00	1,96	13,95	110.344

A mediana próxima de zero e a simetria entre os quartis (-1.80 *versus* 1.96) indicam que a maioria dos trabalhadores apresenta uma escolaridade próxima da média do seu grupo profissional. No entanto, os desvios registados são suficientemente expressivos para justificar a distinção entre casos de *overeducation* e *undereducation*. Assim, este critério permite criar uma variável robusta, com potencial explicativo relevante para análises de desigualdade estrutural, segmentação do mercado de trabalho e ajustamento entre a oferta e a procura de qualificações.

Com base nesta análise, foram criadas três variáveis binárias (*dummies*) que permitem classificar os trabalhadores conforme as constantes na Tabela 8:

Tabela 8 - Designação dos níveis de qualificação e valor

Nível de qualificação	Designação	Assume o valor 1 se
Undereducated	*Under*	*desvio_media_ano_esc_prof* ≤ -1,80
Overeducated	*Over*	*desvio_media_ano_esc_prof* ≥ 1,96
Match	*Match*	-1.8 < desvio_media_ano_esc_prof < 1,96

A escolha dos patamares de classificação baseou-se nos valores empíricos do 1.º e do 3.º quartil da distribuição da variável *desvio_media_ano_esc_prof*. Esta decisão visa garantir uma segmentação estatisticamente equilibrada dos trabalhadores entre os níveis de

qualificação de *over* e *under*, refletindo desvios reais face à média da sua profissão. Deste modo, garante-se que a definição das categorias (*under*, *over* e *match*) resulta do comportamento efetivamente observado nos dados, refletindo as diferenças entre profissões e as variações ao longo do período em estudo [18, 19].

Optou-se pelos quartis em vez da alternativa mais comum da média acrescida ou subtraída de um desvio padrão, dado que os quartis são menos sensíveis a valores extremos e a distribuições assimétricas de escolaridade dentro dos grupos profissionais. Tal abordagem permite obter resultados mais estáveis e representativos, assegurando maior robustez na identificação de situações de *mismatch* educacional [18, 19]. Na Figura 9 está representado o código na criação variável *qualificacao*, com os respetivos níveis de qualificação.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(qualificacao = case_when(
    desvio_media_ano_esc_prof <= -1.80 ~ "Under",
    desvio_media_ano_esc_prof >= 1.96  ~ "Over",
    TRUE ~ "Match" ))
```

Figura 9 - Código da criação variável *qualificacao*

Esta variável é fundamental para os objetivos do estudo, pois permite quantificar e analisar a incidência de *overeducation* e *undereducation* ao longo do tempo e entre diferentes segmentos do mercado de trabalho, nomeadamente por idade, sexo, profissão, dimensão da empresa e localização geográfica.

3.7.3. Nacionalidade

A variável original nacionalidade (*nacio*), presente na base de dados, contém dezenas de categorias distintas, correspondentes às diversas nacionalidades existentes. No entanto, para os objetivos do presente estudo, essa granularidade não é necessária, uma vez que o foco da análise não reside nas diferenças entre países específicos, mas sim na distinção entre trabalhadores nacionais e estrangeiros. Assim, foi criada a nova variável *dummy Nacionalidade*, que distingue apenas dois grupos: portugueses e estrangeiros. Esta variável foi construída a partir da informação original *nacio*, atribuindo o valor 0 aos indivíduos com nacionalidade registada como "Portugal" e o valor 1 a todos os restantes, classificados como estrangeiros. Esta simplificação tem como principal objetivo permitir a análise de diferenças estruturais entre trabalhadores nacionais e estrangeiros.

Na Figura 10 está representado o código na criação variável Nacionalidade.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(Nacionalidade = if_else(nacio == "Portugal", "Português",
  "Estrangeiro"))
```

Figura 10 - Código da criação variável *Nacionalidade*

A variável *Nacionalidade* constitui uma dimensão adicional de análise para avaliar potenciais desigualdades no acesso ao mercado de trabalho ou no reconhecimento da qualificação formal, em particular entre trabalhadores imigrantes e Portugueses.

3.7.4. Idade

A variável *idade_Cod*, presente nos dados originais, identifica a idade dos trabalhadores em anos, sendo que no caso de extremos tem a identificação efetuada através de faixas etárias, utilizando categorias como "<=17" e ">=68". Esta estrutura categórica para os casos extremos limita a sua utilização em modelos como regressões lineares ou outros modelos lineares aplicáveis em dados em painel.

Para ultrapassar esta limitação, foi criada a variável *idade_numerica*, na qual:

- O valor "<=17" foi convertido para 17;
- O valor ">=68" foi convertido para 68;
- Os restantes valores foram mantidos como numéricos.

Na Figura 11 está representado o código na criação variável *idade_numerica*.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(idade_numerica = case_when(
    idade_Cod == "<=17" ~ 17,
    idade_Cod == ">=68" ~ 68,
    TRUE ~ as.numeric(as.character(idade_Cod))))
```

Figura 11 - Código da criação variável *idade_numerica*

Esta transformação permitiu tratar a idade como variável quantitativa, tornando-a compatível com métodos de análise económica e estatística mais robustos, nomeadamente modelos de regressão onde se pretende estimar, entre outros, o efeito marginal da idade sobre o rendimento.

3.7.5. Dimensão da empresa

Com base no número de trabalhadores (*empl*) e no volume de negócios anual (*vn*), foi criada uma variável categórica *dim_empresa*, que classifica cada empresa segundo os critérios definidos pela União Europeia para a tipologia de entidades empresariais. Esta variável agrupa as empresas em quatro categorias: Microempresa, Pequena empresa, Média empresa e Grande empresa, conforme os limiares apresentados na Tabela 9.

A classificação foi efetuada exigindo que a empresa cumpra simultaneamente os dois critérios (n.º de trabalhadores e volume de negócios), sendo atribuída à menor categoria em que satisfaça ambas as condições. Caso uma empresa não cumpra os dois requisitos em nenhuma das três primeiras classes (micro, pequena ou média), é automaticamente classificada como grande empresa.

Tabela 9 - Critérios de classificação da dimensão da empresa segundo a tipologia da União Europeia

Tipo de empresa	N.º de trabalhadores	Volume de Negócios Anual
Microempresa	< 10	≤ 2 milhões de euros
Pequena empresa	< 50	≤ 10 milhões de euros
Média empresa	< 250	≤ 50 milhões de euros
Grande empresa	---	Não cumpre pelo menos um dos critérios de "Média"

A inclusão desta variável permite analisar o impacto da dimensão empresarial sobre diversos indicadores, nomeadamente a remuneração, a produtividade e a distribuição de perfis de qualificação.

Na Figura 12 está representado o código na criação variável dimensão da empresa.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(dim_empresa = case_when(
    empl < 10  & vn <= 2000000  ~ "Microempresa",
    empl < 50  & vn <= 10000000 ~ "Pequena empresa",
    empl < 250 & vn <= 50000000 ~ "Média empresa",
    TRUE ~ "Grande empresa" ))
```

Figura 12 - Código da criação variável *dim_empresa*

Esta segmentação é particularmente útil na análise de padrões estruturais no mercado de trabalho, nomeadamente no que diz respeito ao ajustamento entre oferta e procura de qualificações em diferentes contextos empresariais.

3.7.6. Produtividade por trabalhador

Foi criada uma medida de produtividade média do trabalho, definida como o rácio entre o volume de negócios real (*vn_real*) e o número de trabalhadores da empresa (*pempl*). Esta variável constitui uma aproximação da receita média gerada por trabalhador, servindo como *proxy* da produtividade laboral ao nível da empresa. A unidade desta medida corresponde a euros por trabalhador (€/trabalhador), refletindo o volume de negócios médio anual associado a cada trabalhador da empresa.

Embora se trate de uma medida simplificada, esta métrica é particularmente relevante para efeitos comparativos e será utilizada, principalmente, para analisar a relação entre a produtividade das empresas e o rendimento ganho pelos trabalhadores [44, 45].

Neste estudo, a produtividade do trabalho foi aproximada através da seguinte métrica:

$$Produtividade = \frac{Volume\ de\ Negócios\ Real\ da\ Empresa\ (vn\_real)}{Nº\ de\ Trabalhadores\ (pempl)}.$$

Na Figura 13 está representado o código para a criação variável produtividade.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(produtividade = vn_real / pempl)
```

Figura 13 - Código da criação variável *produtividade*

A inclusão desta variável permite incorporar uma dimensão económica fundamental na análise empírica, aproximando o estudo das condições reais do mercado de trabalho.

3.7.7. Transformações logarítmicas

A aplicação de transformações logarítmicas a variáveis económicas, como o salário ou a produtividade, é uma prática comum em análises estatísticas e económicas, especialmente no contexto de modelos de regressão. Esta abordagem oferece várias vantagens teóricas e práticas, entre as quais se destacam:

- Redução da assimetria: em dados económicos é comum existirem valores muito altos em comparação com a maioria (assimetria positiva e existência de *outliers*);
- Conversão de variações absolutas em relativas, facilitando a interpretação dos coeficientes;
- Melhoria da linearidade entre variáveis explicativas e dependentes;
- Reduz problemas de variabilidade desigual nos dados, tornando as estimativas mais fiáveis.

Com o objetivo de aproximar as variáveis quantitativas às suposições de normalidade e homocedasticidade exigidas por muitos modelos econométricos, foram criadas as seguintes variáveis transformadas em logaritmo natural: *log_rganho_real*, *log_vn_real*, *log_vn_anual_real*, *log_produtividade*. Na Figura 14 está representado o código na criação destas novas variáveis logarítmicas.

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(log_rganho_real = ifelse(rganho_real > 0, log(rganho_real), NA))
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(log_vn_real = ifelse(vn_real > 0, log(vn_real), NA))
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(log_vn_anual_real = ifelse(vn_anual_real > 0,
  log(vn_anual_real), NA))
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(log_produtividade = ifelse(produtividade > 0,
  log(produtividade), NA))
```

Figura 14 - Código da criação das novas variáveis logarítmicas

Estas variáveis são especialmente úteis em modelos lineares, dado que permitem interpretar os coeficientes estimados como variações percentuais aproximadas. Adicionalmente, a aplicação do logaritmo ajuda a suavizar o impacto de valores extremos (*outliers*), tornando as distribuições mais simétricas e os modelos mais estáveis.

3.7.8. Variação intra e inter-grupo (efeitos por profissão)

No âmbito da preparação para a análise com dados em painel, foram criadas duas variáveis complementares com base na transformação logarítmica da remuneração real (*log_rganho_real*). Estas variáveis permitem decompor a variação salarial em dois

componentes analiticamente distintos: o efeito médio da profissão e o desvio individual face a esse grupo.

A primeira variável, *media_LogRganho_profissao*, corresponde ao valor médio do logaritmo do salário real (*log_rganho_real*) para cada código de profissão (*prof_4d*). Esta variável capta as diferenças estruturais entre grupos profissionais, funcionando como uma medida da remuneração típica de cada profissão.

A segunda variável, *desvio_LogRganho_individual*, corresponde à diferença entre o salário real (em termos do logaritmo) de cada trabalhador e a média do seu grupo profissional. Esta variável reflete a variação intra-grupo, permitindo analisar a diferença individual face ao padrão salarial da profissão.

Estas duas variáveis são particularmente relevantes para a aplicação de modelos com efeitos fixos e aleatórios, pois permitem distinguir entre:

- A variabilidade explicada por características do grupo profissional (efeito inter-grupo);
- A variabilidade explicada por atributos individuais do trabalhador (efeito intra-grupo), como idade, escolaridade, antiguidade ou tipo de contrato.

Na Figura 15 está representado o código na criação destas novas variáveis intra e inter-grupo:

```
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  group_by(prof_4d) %>%
  mutate(media_LogRganho_profissao = mean(log_rganho_real, na.rm = TRUE))
  %>% ungroup()
ETC_2010_2023_NovasColunas <- ETC_2010_2023_NovasColunas %>%
  mutate(desvio_LogRganho_individual = log_rganho_real -
  media_LogRganho_profissao)
```

Figura 15 - Código da criação variável intra e inter-grupo

Contudo, importa salientar que, apesar da utilidade analítica do *desvio_LogRganho_individual*, esta variável não foi incluída no modelo principal deste estudo, uma vez que deriva diretamente da variável dependente (*log_rganho_real*). A sua utilização no mesmo modelo poderia introduzir circularidade e problemas de multicolinearidade, comprometendo a validade estatística das estimativas.

Por outro lado, a variável *media_LogRganho_profissao* foi incluída no modelo de efeitos fixos como variável de controlo. Esta escolha permite capturar o contexto salarial médio da profissão, isolando melhor o impacto da variável *qualificacao* (*under*, *over* e *match*) sobre o salário real. Como o modelo controla efeitos fixos ao nível individual e não ao nível da profissão, a inclusão de *media_LogRganho_profissao* é metodologicamente válida e contribui para uma maior robustez dos resultados.

3.7.9. Base de dados consolidada – pós adição de novas variáveis

Após a adição das novas variáveis, ficamos com uma base de dados com 80 variáveis e 41.433.063 observações. Este ficheiro ficou gravado com o nome de *ETC_2010_2023_NovasColunas.RData* e tem uma dimensão de 3.600.128KB (aproximadamente 3,6GB).

Assim, com esta base de dados, estamos em condições de seguir os nossos objetivos de estudos sobre a qualificação *versus* profissão e o efeito sobre o salário.

4. Estatística descritiva

Este capítulo tem como objetivo apresentar a estrutura final da base de dados utilizada no estudo, bem como descrever as suas principais estatísticas. Para além da análise descritiva clássica, na primeira parte do capítulo são detalhadas as etapas de filtragem e construção de um painel equilibrado que servirá de base para os modelos desenvolvidos nos capítulos seguintes. As escolhas efetuadas foram orientadas por critérios de consistência metodológica e viabilidade computacional.

Na segunda parte é apresentada a estatística descritiva da amostra final, com o objetivo de identificar padrões gerais e verificar a coerência dos dados face às hipóteses de estudo.

4.1. Construção da amostra de painel equilibrado

Após a fase de tratamento, limpeza e normalização descrita no Capítulo 3, a base de dados consolidada passou a integrar 80 variáveis e um total de 41.433.063 observações, correspondentes a registos individuais de trabalhadores ao longo do período 2010–2023.

Contudo, para efeitos de modelação e análise empírica, foi necessário refinar adicionalmente a amostra, com o objetivo de garantir consistência nas variáveis centrais do estudo e simultaneamente otimizar o desempenho computacional.

4.1.1. Seleção de variáveis relevantes

Para simplificar a base de dados e garantir foco nas dimensões mais relevantes, foi extraído um subconjunto de variáveis essenciais à análise. A Tabela 10 apresenta as variáveis selecionadas.

Tabela 10 - Lista das variáveis selecionadas para estudo

Código da variável	Designação
log_rganho_real	Logaritmo do rendimento mensal do trabalhador (corrigido pelo IPC)
qualificacao	Grau de alinhamento entre escolaridade e profissão (*Under, Over, Match*)
anos_escolaridade	Anos de escolaridade do trabalhador (estimados com base nas habilitações)
antig	Antiguidade do trabalhador na empresa
idade_numerica	Idade do trabalhador (em formato contínuo)
sexo	Sexo do trabalhador
Nacionalidade	Nacionalidade (Português ou Estrangeiro)
tipo_contr1	Tipo de contrato de trabalho (sem termo; com termo certo; com termo incerto, outros)
media_LogRganho_profissao	Média do *log_rganho_real* por ocupação/profissão
log_produtividade	Logaritmo da produtividade média do trabalho na empresa, medida como o volume de negócios real por trabalhador (*vn_real* / *pempl*).
dim_empresa	Dimensão da empresa (Micro, Pequena, Média, Grande)
caem11	Setor de atividade económica (CAE – 1 letra, 19 categorias)
nut2_emp	Localização da empresa (NUTS II, 8 categorias)
ntrab	Identificador único do trabalhador
ANO	Ano de referência

Na Figura 16 apresenta-se o código da filtragem da base de dados com apenas estas variáveis.

```
Painel_2010_2023_Filtrado_5 <- ETC_2010_2023_NovasColunas
  %>%  select(ntrab, ANO, log_rganho_real, antig, anos_escolaridade,
  idade_numerica, qualificacao, log_produtividade,
  media_LogRganho_profissao, dim_empresa, sexo, Nacionalidade, nut2_emp,
  caem11, tipo_contr1)
```

Figura 16 - Código da filtragem da base de dados com as variáveis selecionadas

4.1.2. Remoção de observações com valores em falta (NA)

De seguida, foram removidos todos os registos com dados em falta para as principais variáveis. Na Figura 17 está representado o código utilizado na remoção das observações com valores em falta nas variáveis selecionadas:

```
colSums(is.na(Painel_2010_2023_Filtrado_5))
Painel_2010_2023_Filtrado_sem_NA_5 <- Painel_2010_2023_Filtrado_5 %>%
  filter(
    !is.na(rganho_real),
    !is.na(antig),
    !is.na(anos_escolaridade),
    !is.na(idade_numerica),
    !is.na(produtividade),
    !is.na(log_rganho_real),
    !is.na(log_produtividade),
    !is.na(media_LogRganho_profissao),
    !is.na(tipo_contr1))
```

Figura 17 - Código da remoção dos valores em falta (NA)

A Tabela 11 apresenta a dimensão da base antes e depois da filtragem:

Tabela 11 - Resultado da diferença das observações antes e após remoção dos NA

Etapa	Observações
Antes da remoção de NA	41.433.063
Após remoção de NA	36.624.042
Diferença (registos excluídos)	4.809.021

De forma a avaliar se a remoção dos registos com valores omissos poderia comprometer a representatividade da amostra, foi realizada uma comparação entre a distribuição das principais variáveis antes e depois da limpeza da base de dados. Conforme se apresenta em tabela em anexo (Tabela A.1), as proporções mantêm-se praticamente inalteradas, assegurando que a amostra final continua alinhada com a população inicial.

4.1.3. Restrição das observações com 13 ou mais anos de acompanhamento

Para garantir maior robustez estatística e estabilidade temporal na análise com dados em painel, foi definido efetuar as diferentes análises com foco num grupo de trabalhadores com registos em pelo menos 13 anos distintos (dos 14 possíveis, entre 2010 e 2023). A decisão de restringir a amostra a trabalhadores com pelo menos 13 anos de registo consecutivo baseou-se em critérios metodológicos e práticos. Do ponto de vista estatístico, esta abordagem permite garantir maior robustez nas análises com dados em painel, pois assegura que os efeitos individuais capturados ao longo do tempo se baseiam em observações suficientemente ricas e consistentes. Por outro lado, do ponto de vista computacional, esta filtragem permitiu reduzir significativamente o volume de dados a processar, tornando viável a execução de modelos estatísticos mais exigentes em termos de recursos.

Na Figura 18 está representado o código na remoção dos trabalhadores com menos de 13 anos de registo e a sua verificação.

```
Painel_2010_2023_Filtrado_13_sem_NA_5 <-
  Painel_2010_2023_Filtrado_sem_NA_5 %>%
    group_by(ntrab) %>%
    filter(n_distinct(ANO) >= 13)
track_painel <- Painel_2010_2023_Filtrado_13_sem_NA_5 %>%
  group_by(ntrab) %>%
  summarise(Anos_Acompanhados = n_distinct(ANO), .groups = "drop")
table(track_painel$Anos_Acompanhados)
```

Figura 18 - Código da seleção dos trabalhadores com 13 ou mais anos de registo

A distribuição final dos trabalhadores por anos de acompanhamento é apresentada na Tabela 12, correspondendo a 12.704.971 observações.

Tabela 12 - Resumo do número de trabalhadores com 13 ou mais anos de registo

Anos de acompanhamento	Nº de trabalhadores
13	344.919
14	587.216
Total	932.135

Importa ainda referir que esta restrição não compromete a representatividade nem a variabilidade da amostra, uma vez que os trabalhadores retidos representam uma fração substancial e diversificada da população ativa observada ao longo do período em análise (2010 a 2023).

Assim, a caracterização final da amostra de trabalho é a seguinte:

- Entre 13 a 14 observações por trabalhador;
- Cerca de 12,7 milhões de observações (≈ 932.135 × média de 13,63 anos);
- 15 variáveis finais, sem *missing values*.

4.2.Caracterização comparativa por nível de qualificação

A Tabela 13 resume o perfil médio dos trabalhadores de acordo com a posição relativa do seu nível de escolaridade face à média de escolaridade da sua profissão, classificando-os em três grupos: *overeducated* (*Over*), *undereducated* (*Under*) e próximo da média da sua profissão (*match*). A distinção resulta da variável *qualificacao*, construída com base no desvio entre os anos de escolaridade do indivíduo e a média do seu grupo profissional (conforme explicado na secção 3.7.2).

A análise apresentada na Tabela 13 detalha as principais diferenças entre os grupos:

Tabela 13 - Estatística das variáveis selecionadas repartido pelo nível de qualificação

Variáveis		Qualificação		
		Over	***Match***	***Under***
	Nº Observações	2.508.567	6.814.758	3.381.646
Sexo	Homem	59%	55%	60%
	Mulher	41%	45%	40%
	Idade (média)	40,4	41,3	46,7
Nacionalidade	Português	97%	98%	97%
	Estrangeiro	3%	2%	3%
	Anos Escolaridade (média)	13,2	10,9	5,94
	Antiguidade (média)	10	11,4	14,2
	Remuneração (média, em euros)	1.461	1.270	949
Tipo Contrato	Sem termo	69%	72%	76%
	Termo certo	24%	22%	18%
	Termo incerto	7%	6%	6%
Dimensão Empresa	Micro	15%	16%	19%
	Pequena	24%	25%	29%
	Media	25%	25%	26%
	Grande	36%	34%	26%
	Produtividade (média, em euros)	224.878	179.640	123.767
Mudança de Emprego	Média	6,6%	5%	3,7%

4.2.1. Sexo

De acordo com os dados da Tabela 13, a distribuição por sexo varia entre as categorias de qualificação, onde a categoria com maior proporção de homens é a *Under* (60%), enquanto a categoria com maior proporção de mulheres é a *Match* (45%).

4.2.2. Idade (média)

Os *undereducated* são significativamente mais velhos em média do que os que se encontram nas restantes categorias. Este resultado é coerente com a evolução do sistema educativo português nas últimas décadas. Trabalhadores mais velhos tendem a ter níveis de escolaridade inferiores, uma vez que cresceram num contexto com menor acesso à educação formal [46].

4.2.3. Nacionalidade

Em todos os grupos, pelo menos 97% dos trabalhadores são portugueses. A baixa representação de estrangeiros (2~3%) reflete as características da amostra com longos períodos de acompanhamento (dados de pelo menos 13 anos dos 14 observados), que exclui a maioria dos trabalhadores imigrantes com estadias temporárias ou instáveis no mercado formal. Adicionalmente, é plausível que parte dos trabalhadores inicialmente registados como estrangeiros tenha adquirido nacionalidade portuguesa durante o período em análise.

4.2.4. Antiguidade na empresa

A antiguidade média é maior nos *undereducated*. Este fenómeno pode estar associado a uma menor rotação do emprego ou à dificuldade em ascender a posições mais qualificadas. Essa estabilidade pode também estar relacionada com a idade e com a menor mobilidade interempresarial, uma vez que trabalhadores mais velhos tendem a apresentar maior antiguidade.

4.2.5. Anos de escolaridade (média)

Como seria de esperar, o grupo de trabalhadores *overeducated* apresenta, em média, níveis de escolaridade mais elevados, enquanto os *undereducated* têm, em média, menos de 6 anos de escolaridade formal, o que, em muitos casos, corresponde apenas ao 1.º ciclo.

4.2.6. Remuneração (média)

Os trabalhadores *overeducated* recebem, em média, cerca de 15% mais do que os trabalhadores que apresentam um nível de escolaridade compatível com a média da profissão, o que decorre também do facto de possuírem mais 2 ou 3 anos de escolaridade em média.

4.2.7. Produtividade (média)

A produtividade média é mais elevada na sub-amostra de *overeducated*, o que poderá refletir uma maior afetação destes trabalhadores a empresas com maior produtividade. Este efeito pode ser potenciado por um capital humano mais elevado e pela associação com contextos laborais mais estruturados, como grandes empresas ou contratos sem termo, conforme abordado nas variáveis seguintes.

4.2.8. Dimensão da empresa

Os *overeducated* predominam nas grandes empresas e os *undereducated* estão sobretudo afetos a micro e pequenas empresas. Estes resultados corroboram com estudos que associam empresas maiores a processos de seleção mais exigentes, maior complexidade funcional e maior retorno esperado da qualificação [47]. Empresas pequenas tendem a recrutar perfis menos qualificados e oferecem salários mais baixos.

4.2.9. Tipo de contrato

A maior percentagem de trabalhadores com vínculo definitivo ocorre para o caso dos *undereducated*. Este resultado não se explica apenas pelo efeito da amostra longitudinal, que exclui trabalhadores com vínculos precários e favorece perfis mais estáveis, mas também pelo facto de este grupo apresentar, em média, maior idade e antiguidade, características que aumentam a probabilidade de contratos sem termo.

Adicionalmente, a maior prevalência de contratos a termo entre os *overeducated* poderá refletir dois fenómenos complementares:

- A maior mobilidade interempresarial, associada a processos de transição entre empregos e períodos de experiência;
- A maior probabilidade de serem contratados em regime temporário para projetos específicos, dada a sua especialização técnica ou académica.

Este padrão sugere que, apesar das suas qualificações, os *overeducated* estão mais expostos à segmentação contratual e à rotatividade no mercado de trabalho, enquanto os *undereducated* tendem a manter vínculos mais estáveis, quer pela sua idade e antiguidade, quer pela menor disponibilidade de alternativas no mercado.

4.2.10. Mudança de emprego

A métrica de mudança de emprego serve para entender o comportamento do mercado de trabalho. Através dela, é possível comparar a mobilidade profissional entre diferentes qualificações e perceber, por exemplo, se trabalhadores com certas qualificações tendem a mudar mais de emprego do que outros.

Esta variável foi construída a partir do identificador único da empresa (*EMP_ID*) registado em cada ano. Para cada trabalhador (*ntrab*), verificou-se se o *EMP_ID* era diferente do ano imediatamente anterior. Sempre que tal ocorreu, registou-se uma mudança de emprego, atribuída ao ano de chegada à nova empresa. Assim, a variável assume valor 1 no ano em que se verifica a mudança e 0 nos restantes anos.

Neste caso, os trabalhadores *overeducated* mudam mais de empresa ao longo do tempo, o que pode indicar:

- Maior insatisfação com o posto atual;
- Maior ambição ou mobilidade de carreira;
- Maior empregabilidade por parte do mercado, que valoriza o seu capital humano.

Os *undereducated*, por contraste, apresentam menor mobilidade, o que pode refletir barreiras de progressão profissional ou segmentação do mercado de trabalho [48].

4.3. Perfil demográfico e educativo

Este subcapítulo analisa a evolução da escolaridade dos trabalhadores e a sua adequação às exigências médias das profissões entre 2010 e 2023. São exploradas duas dimensões complementares: o nível formal de instrução (anos de escolaridade) e a qualificação relativa (*Over*, *Match* ou *Under*), construída a partir do desvio face à média de cada grupo profissional.

4.3.1. Distribuição da escolaridade ao longo do tempo

A Figura 19 apresenta a distribuição da escolaridade dos trabalhadores incluídos na amostra, segmentada por níveis de ensino e observada ao longo do período 2010 a 2023. A variável analisada corresponde à variável original *habil2*, agrupada por níveis de escolaridade.

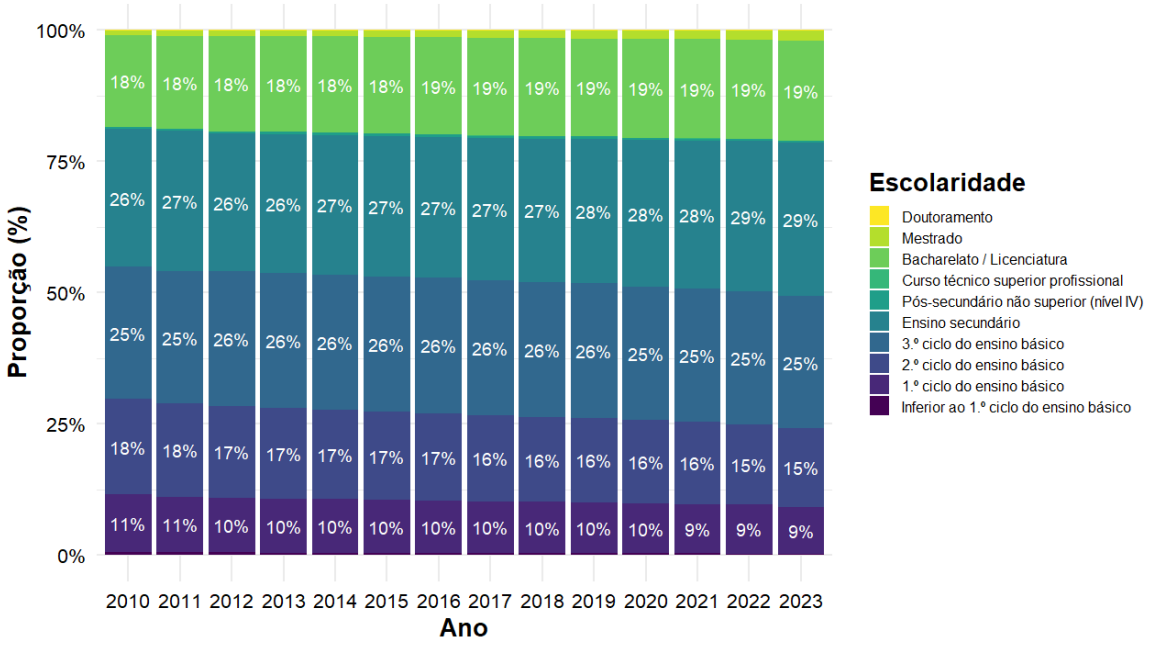

Figura 19 - Proporção de escolaridade ao longo dos anos

A análise da evolução da escolaridade da força de trabalho permite identificar uma tendência clara de elevação do nível de qualificação formal ao longo do período em análise. No entanto, importa salientar que, dado estarmos a considerar apenas trabalhadores com pelo menos 13 ou 14 observações consecutivas, a nossa amostra só inclui indivíduos que já se encontravam inseridos no mercado de trabalho em 2011. Assim, mais do que refletir a substituição geracional por novos trabalhadores mais qualificados, os resultados traduzem sobretudo a progressiva qualificação de indivíduos já ativos, através da conclusão de níveis adicionais de ensino.

Esta evolução manifesta-se em três padrões principais. Em primeiro lugar, observa-se um aumento da proporção de trabalhadores com ensino secundário (12 anos de escolaridade), que passa de 26 % em 2010 para 29 % em 2023, consolidando-se como o nível modal da amostra. Em segundo lugar, verifica-se uma diminuição progressiva dos níveis mais baixos de escolaridade. A proporção de trabalhadores com apenas o 1.º ciclo reduziu-se de 18 % para 15 %, enquanto a dos que possuem uma escolaridade inferior ao 1.º ciclo caiu de 11 % para 9 %, entre 2010 e 2023. Estes dados não resultam de substituição direta de trabalhadores menos qualificados por outros mais qualificados, mas sim do processo de reconversão e elevação das qualificações dentro da própria população observada.

Por fim, os níveis de qualificação intermédia e superior têm vindo a afirmar-se progressivamente. Destaca-se, em particular, o crescimento do ensino pós-secundário não superior (nível IV) e dos cursos técnicos superiores profissionais (cursos TeSP), que se têm consolidado como vias alternativas dentro do ensino superior, oferecendo formações especializadas com forte inserção no mercado de trabalho.

Este movimento reflete a transformação estrutural do perfil educativo da população ativa portuguesa, impulsionada por:

- A massificação do ensino secundário e superior nas últimas duas décadas, com forte impacto nas gerações mais jovens [49, 50];
- Políticas públicas de incentivo à conclusão da escolaridade obrigatória e reconversão profissional;
- O reforço da qualificação dos trabalhadores já inseridos no mercado de trabalho, através de programas de educação e formação ao longo da vida [50].

4.3.2. Proporção de qualificação (*Under*, *Match* ou *Over*) ao longo do tempo

A Figura 20 apresenta a evolução da distribuição dos trabalhadores segundo o seu grau de alinhamento entre escolaridade e exigência média da sua profissão, representado pela variável *qualificacao*, construída especificamente para esta investigação.

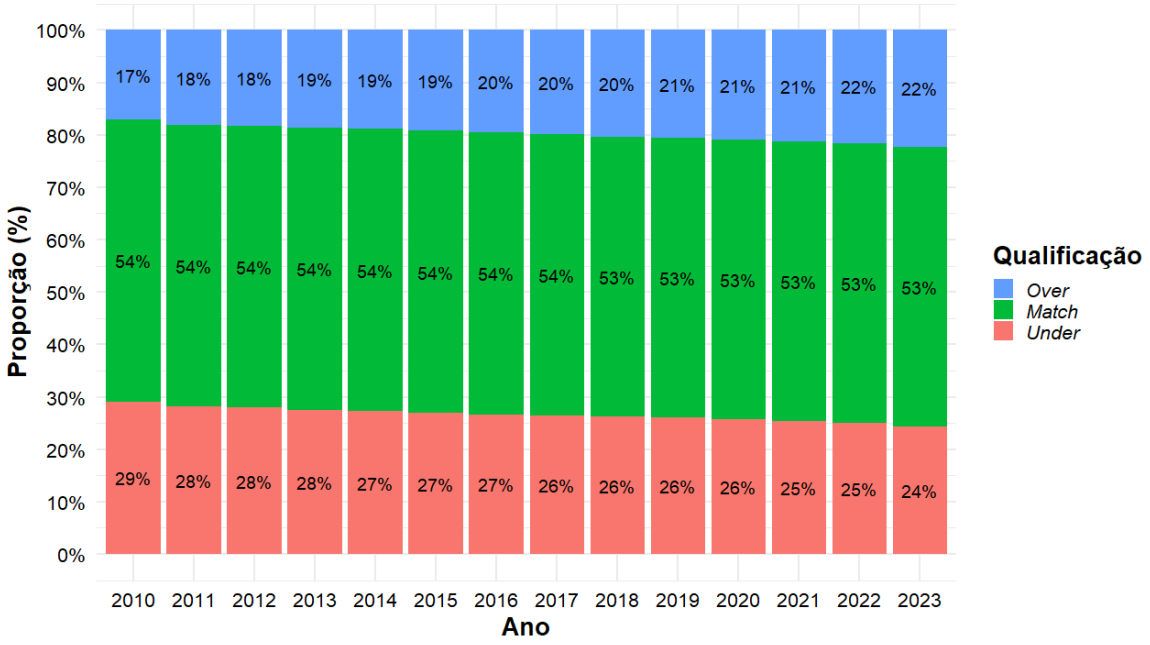

Figura 20 - Proporção de trabalhadores por tipo de qualificação ao longo dos anos

Ao longo do período de 2010 a 2023, observa-se uma estabilidade notável na proporção de trabalhadores classificados como *match*, mantendo-se entre 53% e 54%. Este grupo representa aqueles cuja escolaridade está próxima da média do seu grupo profissional, o que indica um ajustamento razoável entre qualificação e função desempenhada.

Por outro lado, a proporção de trabalhadores subqualificado mostra uma tendência decrescente, passando de 29% em 2010 para 24% em 2023. Esta evolução pode estar relacionada com:

- O aumento contínuo da escolaridade média entre os trabalhadores, o que reduz naturalmente o número de casos de *undereducation*. Este fenómeno é observado na literatura europeia e, em particular, na portuguesa, onde a diminuição da incidência de *undereducation* deve-se largamente ao crescimento da escolaridade da população ativa [26, 46, 51];
- A mobilidade profissional, dado que mudanças de emprego podem alterar o estatuto de qualificação dos trabalhadores, por exemplo, promoções ou reorientações de carreira podem levar a situações de *undereducation*, enquanto transições para funções menos exigentes ou em áreas distintas podem reduzir esse desajuste [11];
- O efeito de políticas públicas voltadas para a qualificação profissional e reconversão de trabalhadores, a partir de programas de formação contínua e reconversão parecem estar associados a uma melhoria no ajustamento entre escolaridade e profissão [52].

Em contraste, a percentagem de trabalhadores *overeducated* apresenta um crescimento progressivo, subindo de 17% em 2010 para 22% em 2023. Este fenómeno tem sido documentado em economias avançadas, refletindo um desalinhamento entre os perfis formados pelo sistema educativo e a estrutura da procura laboral [53].

4.4. Perfil profissional e contratual

Este subcapítulo analisa a evolução de aspetos estruturais do percurso profissional dos trabalhadores, incluindo o tipo de contrato, a dimensão da empresa, a localização geográfica, o setor económico e a ocupação. As tendências observadas refletem, não só, dinâmicas reais do mercado de trabalho português entre 2010 e 2023, mas também mudanças estruturais associadas a fatores económicos, tecnológicos e institucionais, influenciando a distribuição de oportunidades, a mobilidade laboral e a segmentação do mercado de trabalho.

4.4.1. Tipo de contrato ao longo do tempo

A Figura 21 apresenta a distribuição percentual dos trabalhadores por tipo de contrato entre 2010 e 2023, com base no variável *tipo_contr1*, onde distingue os principais regimes contratuais existentes: contrato sem termo, contrato com termo certo, contrato com termo incerto e outras.

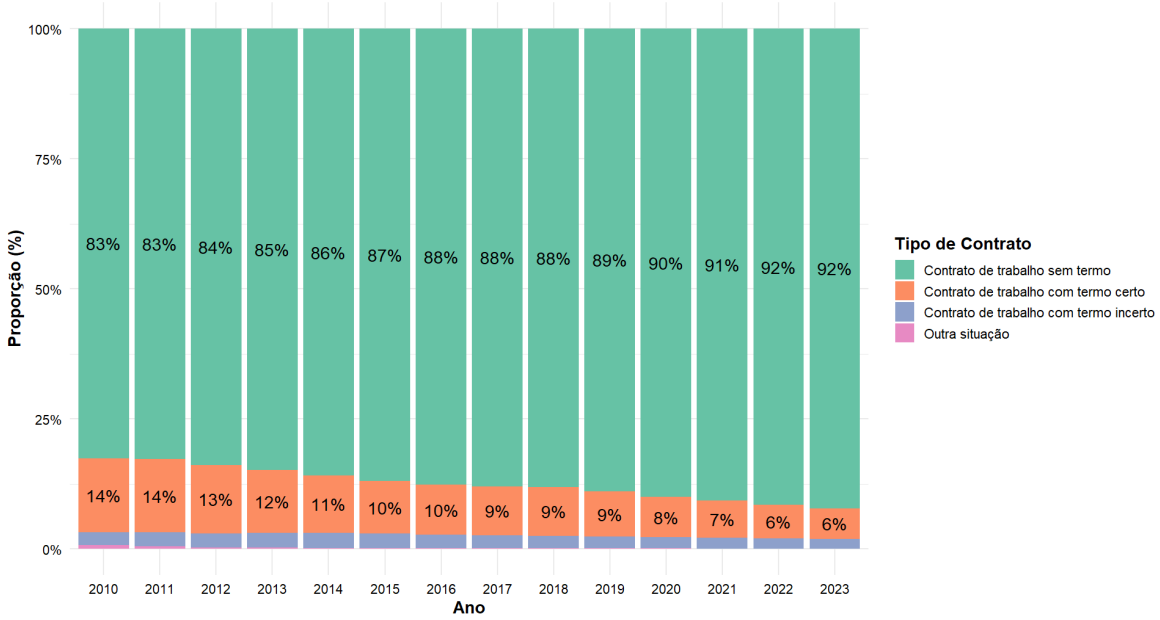

Figura 21 - Proporção por tipo de contrato ao longo dos anos

Os resultados mostram uma tendência clara e progressiva de aumento dos contratos sem termo, que passam de 83% em 2010 para 92% em 2023. Em contrapartida, os contratos com termo certo registam uma redução significativa (de 14% para 6%), enquanto os contratos com termo incerto mantêm-se estáveis, com valores marginais. No subcapítulo anterior foram apresentadas algumas sugestões que poderá justificar a tendência apresentada neste gráfico. Uma vez que todos os trabalhadores observados já estavam no mercado de trabalho em 2010 ou 2011, esta evolução descreve a melhoria do tipo de contrato destes trabalhadores, não refletindo necessariamente a tendência uma vez que não inclui o tipo de contrato dos novos trabalhadores que ingressaram no mercado de trabalho depois de 2012.

4.4.2. Distribuição por dimensão da empresa (*dim_empresa*)

A Figura 22 apresenta a proporção de trabalhadores por tipo de empresa ao longo do período em análise, segundo a variável *dim_empresa*, construída considerando quer o número de trabalhadores (*pempl*), quer o volume de negócios (*vn*).

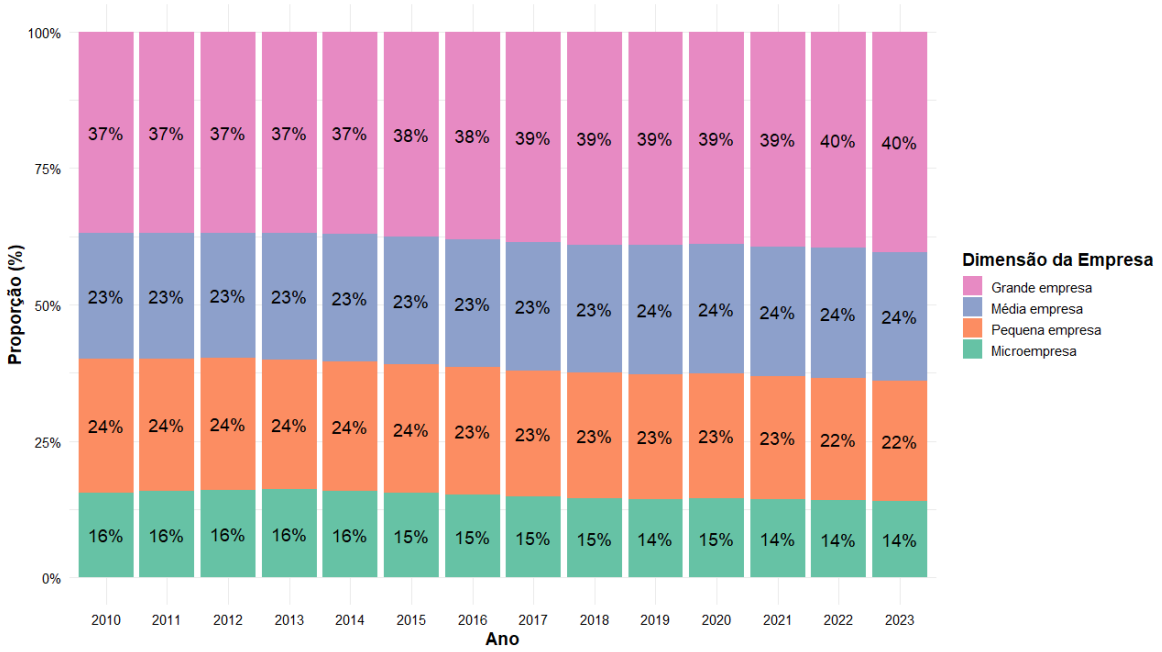

Figura 22 - Proporção por dimensão da empresa ao longo dos anos

Os dados revelam uma estrutura empresarial relativamente estável ao longo do tempo, com uma ligeira tendência de crescimento do peso das grandes empresas. Em 2010, cerca de 37% dos trabalhadores estavam afetos a este tipo de empresas, valor que aumentou para 40% em 2023. As médias e pequenas empresas mantiveram a sua expressão relativamente constante, representando entre 23% e 24% do total de trabalhadores. Por outro lado, os trabalhadores afetos a microempresas registaram uma ligeira diminuição no seu peso relativo, passando de 16% para 14% ao longo do período em análise.

Esta tendência pode refletir múltiplas dinâmicas:

- A maior resiliência e capacidade de retenção de pessoal das grandes empresas em períodos de crise, como a crise financeira e a pandemia de COVID-19;
- A concentração crescente do emprego em grupos empresariais com maior escala e organização formal;
- A redução relativa do emprego em microempresas, muitas das quais operando com vínculos informais ou com elevada rotatividade.

4.4.3. Distribuição geográfica por NUTS II

A Figura 23 apresenta a distribuição percentual dos trabalhadores por região NUTS II, com base na localização da empresa (*nut2_emp*), entre 2010 e 2023. Esta variável categórica

identifica a região onde o trabalhador exerce atividade, agrupando as empresas em sete regiões continentais e insulares e um grupo adicional "Estrangeiro", que representa situações residuais de empresas com sede fora de Portugal.

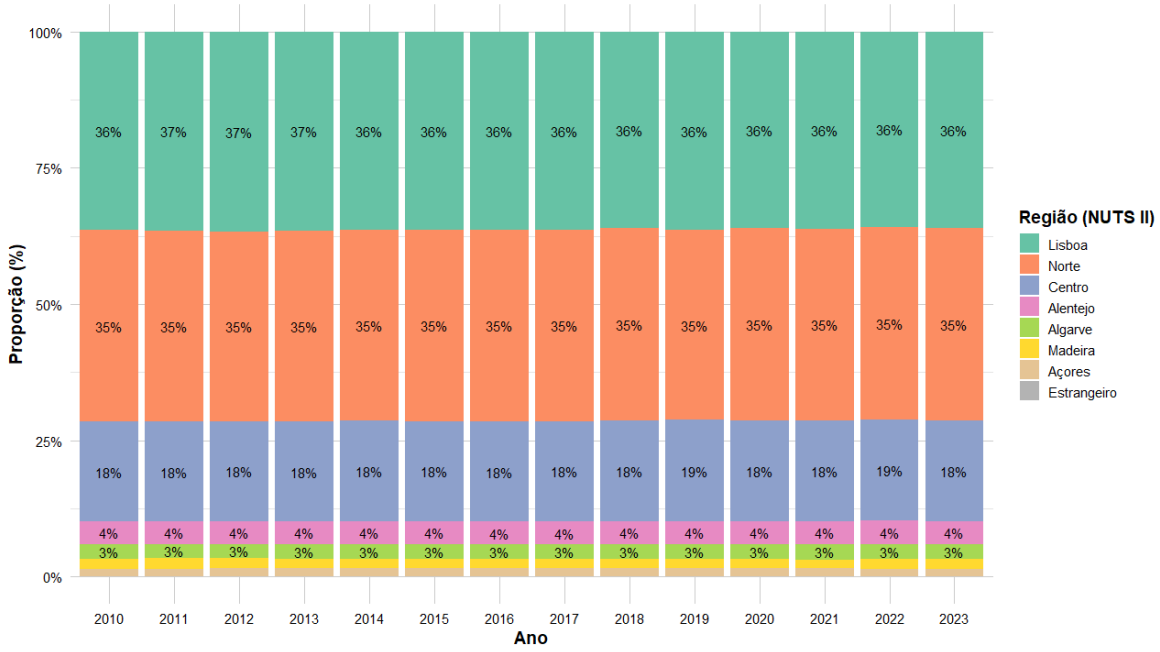

Figura 23 - Distribuição geográfica por região (NUTS II) ao longo dos anos

A estrutura regional da amostra apresenta uma elevada estabilidade ao longo do tempo, com três regiões a concentrarem mais de 85% do total de trabalhadores da base de dados. A região de Lisboa e Vale do Tejo assume o maior peso, representando entre 36% e 38% do total, seguida da região Norte, com uma proporção ligeiramente inferior em torno dos 35%. Já a região Centro emprega aproximadamente 18% dos trabalhadores da base, registando apenas pequenas variações entre anos. As restantes regiões apresentam um peso residual no total da amostra.

Este padrão é consistente com a distribuição geográfica da atividade empresarial em Portugal, refletindo a forte concentração económica e populacional nos eixos de Lisboa e Porto [54].

É importante recordar que a base de dados utilizada é composta apenas por trabalhadores com 13 ou mais anos de registos consecutivos, o que pode reforçar o peso relativo de regiões com maior concentração de emprego formal e estável, como é o caso de Lisboa, do Norte e do Centro do país. Regiões mais sazonais, como o Algarve e a Madeira, ou com maior

rotatividade laboral e desemprego de curta duração (ex.: setores turísticos ou agrícolas) apresentam, também por isso, valores menores.

4.4.4. Distribuição geográfica por setor (*caem1l*)

A Figura A.1, em anexo, apresenta a evolução da distribuição percentual dos trabalhadores por setor de atividade económica entre 2010 e 2023, de acordo com a classificação CAE a 1 letra (*caem1l*). Esta variável identifica o setor económico da empresa onde o trabalhador exerce funções, permitindo observar o enquadramento produtivo do universo analisado.

A estrutura setorial do emprego tem-se mantido relativamente estável ao longo dos anos, com destaque para dois setores principais: As indústrias transformadoras correspondem, ao longo de todo o período, ao maior empregador, com cerca de 27% a 28% dos trabalhadores; e o comércio por grosso e a retalho, com 20% a 21%.

Outros setores com peso relevante incluem as atividades de saúde humana e apoio social assim como os transportes e armazenagem que, desde 2017, empregam cada uma aproximadamente 7% a 8% da força de trabalho; a construção emprega cerca de 7%; as atividades administrativas e dos serviços de apoio apresentam entre 5% e 7%, diminuindo ao longo do tempo; o alojamento e a restauração bem como os serviços financeiros e de seguros apresentam participações próximas dos 5%. Setores como a agricultura, as atividades culturais, a eletricidade e gás e a indústria extrativa apresentam um peso residual, geralmente inferior a 2%.

Importa sublinhar que a classificação setorial se baseia na atividade da entidade empregadora e não na função efetivamente exercida pelo trabalhador. Assim, por exemplo, um profissional de informática que trabalhe numa empresa comercial será incluído no setor do comércio e não no das tecnologias de informação e comunicação.

4.4.5. Profissões mais representadas (Top 10)

A Figura A.2, em anexo, apresenta a evolução da distribuição percentual das 10 profissões com maior número de trabalhadores na amostra, ao longo do período 2010 a 2023, com base na variável *prof_4d*, correspondente à Classificação Portuguesa das Profissões de 2010 (CPP-2010) ao nível de 4 dígitos.

O gráfico revela uma forte dispersão ocupacional, com as 10 profissões mais comuns a representarem apenas 25% do total de trabalhadores ao longo de todo o período. Os restantes 75% pertencem a centenas de outras profissões com menor expressão individual, mas coletivamente dominantes, o que evidência de uma estrutura laboral diversificada e multifuncional.

Entre as profissões mais representadas destacam-se:

- Empregado de escritório em geral (5%);
- Outros trabalhadores relacionados com vendas (4% ~ 5%);
- Motoristas de pesados de mercadorias (2% ~ 3%)
- Trabalhador de limpeza em escritórios, hotéis e outros (2%);
- Empregados de aprovisionamento e armazém (2%);
- Vendedores em loja (2%);
- Supervisor de pessoal administrativo (2%);
- Operadores de máquinas de costura (2%);
- Restantes profissões (< 2%, cada, nos últimos anos observados).

4.5.Indicadores económicos

Este subcapítulo apresenta a evolução dos principais indicadores económicos da base de dados entre 2010 e 2023. Analisa-se o rendimento real médio ao longo do tempo e a sua desagregação por sexo, qualificação, tipo de contrato e dimensão da empresa, bem como a produtividade média por trabalhador. As tendências observadas são interpretadas à luz de transformações macroeconómicas, políticas salariais e dinâmicas estruturais do mercado de trabalho português. A análise baseia-se sempre em valores reais (corrigidos da inflação), garantindo comparabilidade temporal rigorosa e controlo de efeitos dos preços.

4.5.1. Evolução do rendimento médio real

A Figura 24 apresenta a evolução do ordenado médio real mensal dos trabalhadores no período de 2010 a 2023. Os valores foram ajustados pela inflação, com base no Índice de Preços no Consumidor (IPC), considerando o ano 2012 como base, permitindo assim uma análise mais rigorosa do poder de compra dos trabalhadores ao longo do tempo.

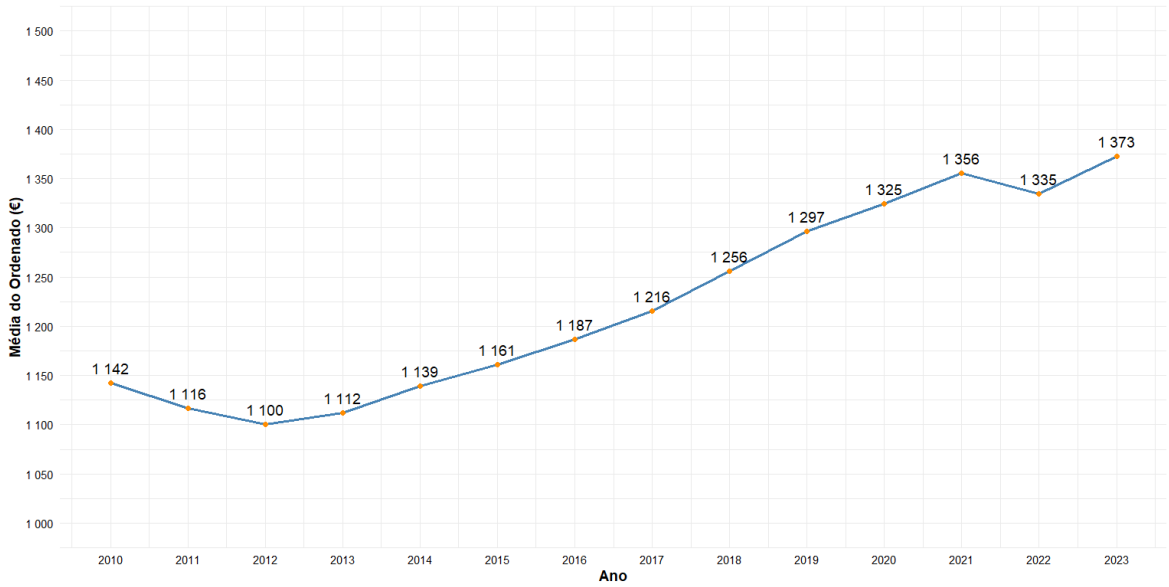

Figura 24 - Evolução do ordenado ganho ao longo dos anos

Verifica-se uma quebra do rendimento real entre 2010 e 2012, com uma descida de aproximadamente 40€ (de 1.142€ para cerca de 1.100€). Estes anos correspondem ao período final da crise da dívida soberana portuguesa e da intervenção externa da *Troika* (Fundo Monetário Internacional, Banco Central Europeu e Comissão Europeia), que implementou um conjunto de medidas de austeridade, incluindo o congelamento de salários no setor público e forte contenção salarial no privado [55].

A partir de 2013, observa-se uma trajetória de crescimento praticamente contínua dos rendimentos reais, até atingir cerca de 1.373€ em 2023. Note-se que este aumento é referente ao salário de indivíduos que entraram no mercado de trabalho pelo menos até 2011 (com pelo menos 13 observações entre 2010 e 2023). Assim sendo, não expressa a alteração observada no mercado de trabalho português, como o ingresso de novos indivíduos no mercado de trabalho, aos quais usualmente estão associados menores rendimentos. Neste caso, pode dever-se a um crescimento associado a uma maior antiguidade ou experiência bem como a outros fatores, tais como o aumento da escolaridade ou ajustamento da mesma face à ocupação.

O ano de 2022 regista uma ligeira quebra nos rendimentos reais, o que poderá refletir o impacto da inflação pós-pandemia e da crise energética mundial, motivada pela guerra na Ucrânia, que afetou o poder de compra de forma transversal. No entanto, esta quebra foi

compensada por uma recuperação no ano seguinte, refletindo ajustes salariais e políticas de mitigação do impacto da inflação.

4.5.2. Evolução do rendimento médio real por sexo

A Figura 25 apresenta a evolução do vencimento médio mensal real por sexo, ao longo do período 2010 a 2023.

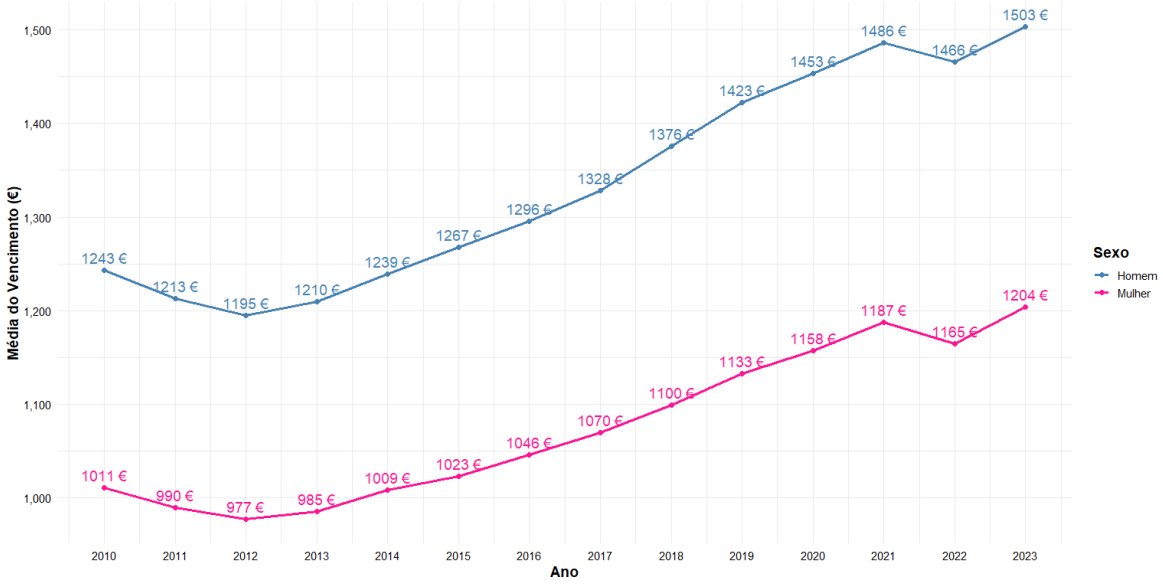

Figura 25 - Vencimento médio real por sexo ao longo dos anos

Entre 2010 e 2023, os vencimentos médios reais aumentaram em Portugal para ambos os sexos. No entanto, a diferença salarial entre homens e mulheres manteve-se persistente e até se agravou ligeiramente ao longo do período. Em 2010, os homens auferiam em média 1.243 €, enquanto as mulheres recebiam 1.011€ (homens com um rendimento médio cerca de 23% superior ao das mulheres). Em 2023, essa diferença agravou-se para cerca de 25%, sendo de 1.503 € o rendimento médio dos homens e 1.204 € o rendimento médio das mulheres.

Esta desigualdade salarial pode resultar de fatores estruturais como a segregação ocupacional e a sub-representação feminina em cargos de topo [56] ou dever-se a uma possível discriminação de género. De salientar que não estão a serem comparados trabalhadores da mesma profissão ou com níveis de experiência ou escolaridade semelhantes, pois é uma comparação em termos gerais, com a inclusão de todos os trabalhadores incluídos na base de dados de ambos os sexos. Todavia, a discrepância não deixa de evidenciar uma tendência

inequívoca entre os trabalhadores analisados. Deste modo, reiteramos que, nesta análise, não estão incluídos os trabalhadores que ingressaram no mercado de trabalho após 2011, podendo esta tendência ter sido mitigada com a aplicação das políticas de promoção de igualdade de género em Portugal.

4.5.3. Rendimento médio ao longo do tempo, por grupo de qualificação

A Figura 26 apresenta a evolução do rendimento médio real mensal dos trabalhadores ao longo dos anos, desagregado por grupo de qualificação: o*verqualified*, *underqualified* e *match*.

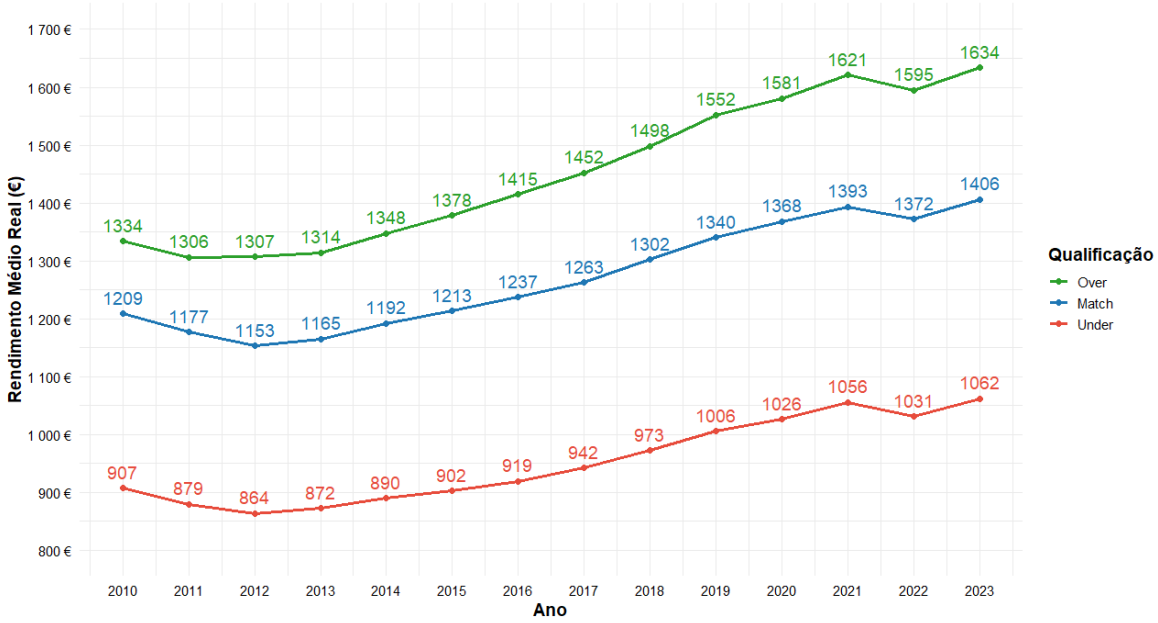

Figura 26 - Evolução do rendimento médio real por grupos de qualificação

A diferenciação salarial entre grupos mantém-se estável ao longo do tempo, onde podemos observar o seguinte:

- Trabalhadores *overeducated* (*Over*) apresentam sistematicamente os rendimentos médios mais elevados, atingindo 1.634 € em 2023, um crescimento de 22% face a 2010 (1.334 €).
- Trabalhadores com qualificação *match* (alinhados com a média da sua profissão) registam valores intermédios, subindo de 1.209 € para 1.406 € (crescimento de 16%).

- Trabalhadores *undereducated* (*Under*) apresentam os rendimentos mais baixos em todos os anos, embora também com tendência de crescimento, passando de 907 € para 1.062 € (crescimento de 17%).

4.5.4. Rendimento médio por tipo de contrato e dimensão da empresa

A Figura 27 apresenta a distribuição do rendimento médio real mensal dos trabalhadores em 2023, segmentado simultaneamente por tipo de contrato e dimensão da empresa. Esta análise permite observar de forma cruzada dois fatores estruturais que afetam a remuneração: a estabilidade contratual e a escala organizacional.

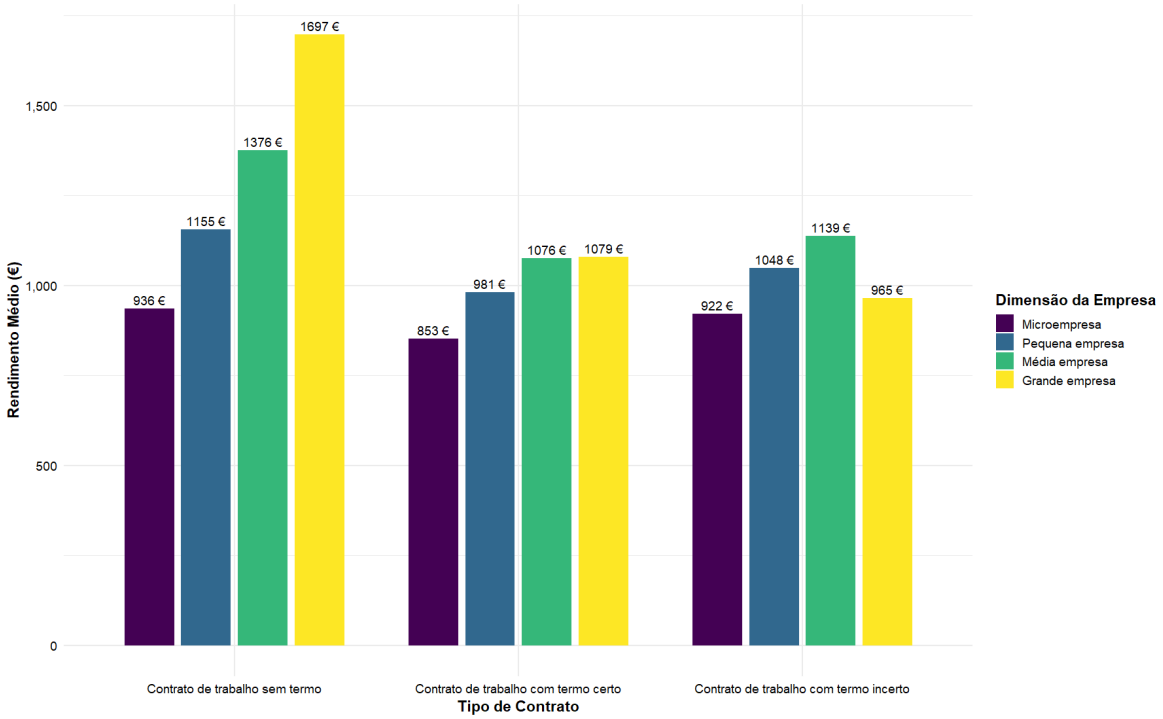

Figura 27 - Rendimento médio por tipo de contrato e dimensão da empresa (2023)

Os trabalhadores com contrato sem termo (ou efetivo) registam, em média, valores salariais mais elevados em todas as categorias de empresa, destacando-se os que trabalham em grandes empresas, com um rendimento médio de 1.697 €.

Por contraste, os trabalhadores com contrato a termo certo ou incerto apresentam rendimentos mais baixos em todos os tipos de empresa, com particular penalização nas microempresas, onde os salários se situam entre 853 € e 922 €.

As grandes empresas pagam, em média, os salários mais elevados independentemente do tipo de contrato, refletindo maior capacidade financeira, presença em setores de maior valor acrescentado e políticas de recursos humanos mais estruturadas.

Nas médias empresas, os valores são também relativamente elevados, sobretudo para contratos sem termo (1.376 €) e termo incerto (1.139 €), sugerindo um perfil intermédio entre estabilidade e remuneração.

Em contraste, as microempresas apresentam os níveis salariais mais baixos em todos os tipos de vínculo laboral. Esta tendência reflete a sua menor capacidade económica e níveis médios de produtividade mais reduzidos face a empresas de maior dimensão, o que limita a margem para remunerar melhor os trabalhadores [57].

4.5.5. Evolução da produtividade média do trabalho por ano

A Figura 28 apresenta a evolução da produtividade do trabalho (em média) ao longo do período de 2010 a 2023. Esta variável foi calculada como o rácio entre o volume de negócios real (corrigido pela inflação) e o número de trabalhadores (*vn_real* / *pempl*).

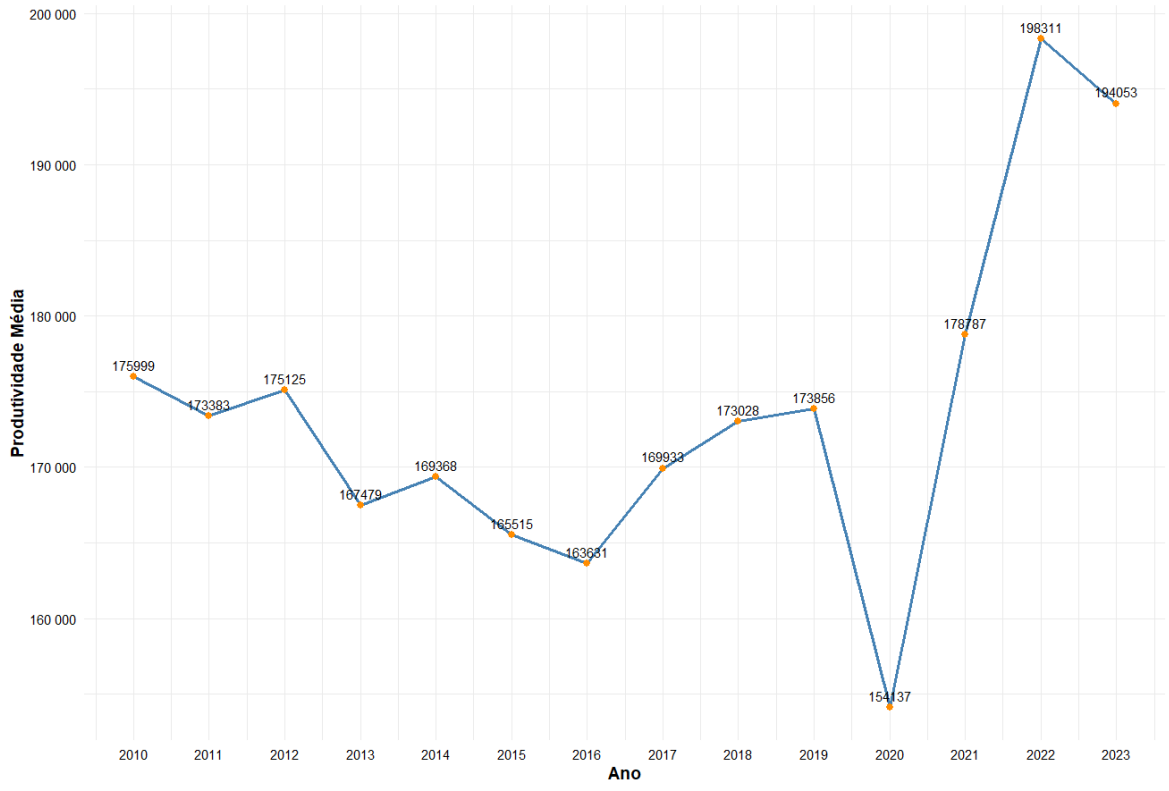

Figura 28 - Evolução da produtividade média por ano

Entre 2010 e 2016, a produtividade apresentou ligeira tendência decrescente (embora oscilante), com o valor a recuar de cerca de 176 mil euros para próximo de 164 mil euros. Esta trajetória reflete os impactos prolongados da crise económica e financeira internacional que atingiu Portugal entre 2011 e 2014, e que se seguiu da implementação de fortes medidas de austeridade no contexto da intervenção da *Troika*.

A partir de 2016, verifica-se uma recuperação moderada da produtividade, culminando em cerca de 173 mil euros em 2019, acompanhando o crescimento económico geral do país e a melhoria gradual dos indicadores de emprego.

O ano 2020 marca uma quebra abrupta na produtividade (que cai para próximo de 154 mil euros), refletindo diretamente os efeitos da pandemia de COVID-19, que causou fortes quebras na atividade económica, sobretudo em setores de elevada intensidade laboral. Após a pandemia, em 2021 e 2022, ocorre uma recuperação significativa, atingindo o pico de mais de 198 mil euros em 2022, seguido de uma ligeira diminuição em 2023, para próximo dos 194 mil euros.

4.6.Síntese final

A análise descritiva desenvolvida ao longo deste capítulo permitiu mostrar um retrato detalhado da estrutura e evolução dos trabalhadores presentes na amostra, abrangendo aspetos demográficos, educativos, profissionais, e económicos ao longo do período de 2010 a 2023.

A amostra utilizada corresponde a um painel quase equilibrado, composto exclusivamente por indivíduos com registo completo em pelo menos 13 anos dos 14 anos compreendidos entre 2010 e 2023. Esta decisão assegura ligação temporal e qualidade dos dados, mas implica também que os resultados obtidos reflitam um subconjunto particular da população ativa portuguesa, isto é, trabalhadores com carreiras mais longas, vínculos laborais mais estáveis e maior persistência no mercado formal. Consequentemente, setores e perfis com elevada rotatividade entre emprego e desemprego, sazonalidade ou informalidade tendem a estar sub-representados. Além disso, trabalhadores que ingressaram no mercado de trabalho depois de 2011 não estão incluídos na amostra.

Apesar desta restrição, os dados demonstraram elevada representatividade estrutural face à realidade nacional. A distribuição por sexo, regiões, setores de atividade e dimensão das empresas está alinhada com os padrões descritos nos relatórios oficiais do INE e outras entidades públicas.

Entre os principais resultados destacam-se:

- A progressiva elevação do nível médio de escolaridade da força de trabalho, acompanhada pelo crescimento dos rendimentos reais. Contudo, persistem diferenças salariais significativas entre grupos de qualificação, em particular entre trabalhadores com a escolaridade ajustada à média da profissão, *overeducated* e *undereducated*.
- O *overeducation* como fenómeno crescente, associado a rendimentos mais elevados, mas também a maior mobilidade contratual e eventual subaproveitamento de competências.
- A centralização do emprego nas grandes empresas e em regiões economicamente mais dinâmicas.
- A desigualdades salariais entre sexos, apesar do crescimento generalizado dos rendimentos.

Por fim, este capítulo forneceu o enquadramento necessário para os modelos dos capítulos seguintes, assegurando que as variáveis centrais foram devidamente definidas, analisadas e compreendidas. A próxima etapa consiste em explorar estas relações através de modelos de dados em painel, permitindo uma análise mais aprofundada das interações entre as variáveis e das dinâmicas entre trabalhadores e ao longo do tempo, consolidando assim a base para as conclusões finais do estudo.

5. Análise com dados em painel

Neste capítulo, aplicam-se técnicas de modelação econométrica com dados em painel para estimar os determinantes do rendimento real mensal dos trabalhadores em Portugal, com particular foco no papel do *mismatch* educacional. A escolha desta abordagem metodológica é possível dada a estrutura longitudinal da base de dados utilizada, composta exclusivamente por indivíduos observados durante 13 ou mais anos entre 2010 e 2023.

São exploradas duas especificações clássicas: o modelo de efeitos aleatórios (MEA) e o modelo de efeitos fixos (MEF), cuja escolha é fundamentada através do teste de Hausman. A análise inicia-se com a preparação dos dados no formato adequado ao painel, seguindo-se a verificação de pressupostos fundamentais como a ausência de multicolinearidade e a homocedasticidade. Após selecionado o modelo mais apropriado, os coeficientes finais são reestimados com erros padrão robustos, assegurando a fiabilidade estatística dos resultados.

A utilização desta abordagem permite controlar a heterogeneidade individual não observada constante ao longo do tempo (como é o caso das características inatas dos indivíduos) que, associada à inclusão de um vasto leque de variáveis de controlo, deverá possibilitar analisar adequadamente o impacto do *mismatch* educacional sobre os salários. Os resultados obtidos sugerem importantes considerações acerca da realidade do mercado de trabalho português ao longo da última década, oferecendo contributos relevantes para o debate académico e para a formulação de políticas públicas informadas.

5.1. Introdução teórica à análise com dados em painel

A análise com dados em painel constitui uma abordagem estatística robusta para estudar fenómenos socioeconómicos que evoluem ao longo do tempo, permitindo controlar simultaneamente a heterogeneidade individual e a variação temporal. No caso deste projeto, a estrutura longitudinal da base de dados, que acompanha os mesmos trabalhadores durante 13 ou 14 anos consecutivos, torna particularmente adequada a utilização de modelos em painel para estimar o impacto da qualificação sobre os rendimentos reais.

Os dados em painel são observações repetidas sobre as mesmas unidades estatísticas (indivíduos, empresas, etc) ao longo de vários períodos. Esta estrutura oferece vantagens consideráveis sobre modelos em corte transversal (*cross-section*), tais como [58, 59]:

- Controlo da heterogeneidade não observada: uma das maiores vantagens é a capacidade de controlar a heterogeneidade individual não observada, considerando que esta é constante ao longo do tempo. Fatores como o talento, a ambição e capacidades cognitivas podem influenciar tanto a escolaridade do individuo quanto o rendimento, mas são difíceis (ou mesmo impossíveis) de medir, podendo ser tratados adequadamente como dados em painel. Ignorar estes fatores pode resultar no enviesamento das estimações obtidas.
- Maior eficiência e variabilidade estatística: a combinação de várias unidades e múltiplos períodos de tempo gera uma amostra maior e mais rica, aumentando a variabilidade dos dados e proporcionando estimativas mais precisas.

Uma questão central na modelação com dados em painel é a escolha entre modelos de efeitos fixos e efeitos aleatórios. O modelo de efeitos fixos assume que os efeitos específicos de cada indivíduo estão correlacionados com as variáveis explicativas, o que permite controlar para toda a heterogeneidade não observada constante no tempo. Já o modelo de efeitos aleatórios pressupõe que os efeitos individuais são não correlacionados com os regressores, o que permite aproveitar também a variação entre indivíduos [58, 59]. Assim, podemos resumir que:

- Efeitos Fixos: A heterogeneidade é um "problema" que deve ser removido. O modelo foi desenhado para eliminar esse viés.
- Efeitos Aleatórios: A heterogeneidade é uma "característica" dos dados que deve ser modelada como aleatória. O modelo assume que ela não causa viés, mas apenas adiciona ruído.

A decisão entre os dois modelos será fundamentada no teste de Hausman, o qual avalia se existe correlação entre os efeitos não observados e as variáveis explicativas. Se essa correlação for significativa, o modelo de efeitos fixos é preferível, caso contrário o modelo de efeitos aleatórios pode ser mais eficiente [58, 59].

Além da escolha do modelo, a robustez dos resultados é uma preocupação metodológica central. Um dos pressupostos clássicos da regressão é a homocedasticidade, ou seja, a suposição de que a variância dos erros do modelo é constante para todas as observações. Quando essa suposição é violada, o que se denomina heterocedasticidade, as estimativas dos coeficientes permanecem válidas, mas os seus erros padrão tornam-se enviesados, podendo

comprometer os testes de significância. Para corrigir este problema, a análise subsequente aplica o cálculo de erros padrão robustos, ajustando a inferência estatística para garantir que as conclusões sobre o impacto das variáveis explicativas nos rendimentos sejam fiáveis, mesmo na presença de heterocedasticidade [58, 59].

A aplicação desta metodologia é particularmente apropriada para o nosso estudo sobre a relação entre o *mismatch* educacional e o rendimento. Ao utilizar dados em painel, podemos isolar o efeito causal da qualificação sobre o rendimento, separando-o de fatores não observados que podem estar enviesando a relação.

5.2. Especificação do modelo de regressão linear e definição das variáveis

O modelo estimado pode ser representado pela seguinte equação, considerando que $i = 1, \ldots, n$ (número de indivíduos na amostra, i.e. n=932.135) *e* $t = 1, \ldots, 14$ (correspondendo aos 14 anos observados: 2010 a 2023):

$$\begin{aligned} log\_rganho\_real_{it} = {} & \beta_0 + \beta_1\, antig_{it} + \beta_2\, anos\_escolaridade_{it} + \beta_3\, idade\_numerica_{it} \\ & + \beta_4\, qualificacaoUnder_{it} + \beta_5\, qualificacaoOver_{it} + \beta_6\, log\_produtividade_{it} \\ & + \beta_7\, media\_LogRganho\_profissao_{it} + \beta_8\, dim\_empresaPequena_{it} + \beta_9\, dim\_empresaMédia_{it} \\ & + \beta_{10}\, dim\_empresaGrande_{it} + \beta_{11}\, sexoMulher_{it} + \beta_{12}\, nacionalidadeEstrangeiro_{it} \\ & + \beta_{13}\, nut2\_empAlgarve_{it} + \ldots + \beta_{19}\, nut2\_empEstrangeiro_{it} + \beta_{20}\, caem1lB_{it} + \ldots \\ & + \beta_{37}\, caem1lS_{it} + \beta_{38}\, tipo\_contrCerto_{it} + \beta_{39}\, tipo\_contrIncerto_{it} + \beta_{40}\, tipo\_contrOS_{it} + \varepsilon_{it} \end{aligned}$$

onde o erro ε_{it} pode ser decomposto em $\varepsilon_{it} = \mu_i + \xi_{it}$, sendo ξ_{it} uma variável aleatória com distribuição normal com média zero e desvio padrão σ.

- Com dados em painel utilizando um modelo de efeitos fixos (MEF): μ_i é uma constante que varia de indivíduo para indivíduo.
- Com dados em painel utilizando um modelo de efeitos aleatórios (MEA): μ_i tem distribuição normal com média zero e desvio padrão σ_i (que depende de indivíduo para indivíduo) sendo independente de ξ_{it}.

Variável Dependente:

- log_rganho_real: o logaritmo do rendimento do trabalhador, que é a variável que o modelo pretende explicar.

Variáveis quantitativas:

- antig: antiguidade do trabalhador na empresa (anos).
- anos_escolaridade: número de anos de escolaridade formal.
- idade_numerica: idade em anos.
- log_produtividade: logaritmo da produtividade da empresa.
- media_LogRganho_profissao: média do log rendimento da profissão.

Variáveis dummy e categorias de referência:

- qualificacao (*mismatch* educacional): conjunto de duas variáveis *dummy* que indicam se o trabalhador está *overeducated* (*qualificacaoOver*) ou *undereducated* (*qualificacaoUnder*) relativamente ao posto de trabalho.
    - Referência: *match* (correspondência no nível de escolaridade).
- dim_empresa: três variáveis *dummy* que indicam se a empresa é de pequena dimensão (*dim_empresaPequena*), média dimensão (*dim_empresaMédia)* ou grande dimensão (*dim_empresaGrande*).
    - Referência: microempresa.
- sexo: variável *dummy* que indica se o trabalhador é mulher (*sexoMulher*)
    - Referência: homem.
- Nacionalidade: variável *dummy* que indica se o trabalhador é estrangeiro (*nacionalidadeEstrangeiro*)
    - Referência: português.
- nut2_emp (região da empresa): sete variáveis *dummy* que indicam a localização da empresa, nomeadamente se a localização é Centro (*nut2_empCentro*), Lisboa (*nut2_empLisboa*), Alentejo (*nut2_empAlentejo*), Algarve (*nut2_empAlgarve*), Açores (*nut2_empAçores*), Madeira (*nut2_empMadeira*) ou Estrangeiro (*nut2_empEstrangeiro*)
    - Referência: região Norte.
- caem1l: dezoito variáveis *dummy* (*caem1lB, caem1lC, …, caem1lS*) que indicam o setor de atividade da empresa, respetivamente B - Indústrias extractivas; C - Indústrias transformadoras; D - Electricidade, gás, vapor, água quente e fria e ar frio; E - Captação, tratamento e distribuição de água; saneamento, gestão de resíduos e despoluição; F - Construção; G - Comércio por grosso e a retalho; reparação de

veículos automóveis e motociclos; H - Transportes e armazenagem; I - Alojamento, restauração e similares; J - Actividades de informação e de comunicação; K - Actividades financeiras e de seguros; L - Actividades imobiliárias; M - Actividades de consultoria, científicas, técnicas e similares; N - Actividades administrativas e dos serviços de apoio; O - Administração pública e defesa; segurança social obrigatória; P - Educação; Q - Actividades de saúde humana e apoio social; R - Actividades artísticas, de espectáculos, desportivas e recreativas; S - Outras actividades de serviços.

    - Referência: setor A - Agricultura, produção animal, caça, floresta e pesca.

- tipo_contr1: três variáveis *dummy* que indicam o tipo de contrato: contrato de trabalho com termo certo (*tipo_contrCerto*); Contrato de trabalho com termo incerto (*tipo_contrIncerto*) e outra situação (*tipo_contrOS*).
    - Referência: contrato de trabalho sem termo.

5.3.Criação da estrutura em painel

Antes de proceder à estimação de modelos econométricos com dados em painel, é necessário transformar a base de dados num formato que permita ao software estatístico reconhecer a estrutura longitudinal das observações. No caso da linguagem R, esta preparação é feita através da função *pdata.frame*, do pacote *plm*.

A função *pdata.frame* cria um objeto de classe específica que identifica as dimensões do painel, nomeadamente o identificador individual (*ntrab*, representado no modelo pelo índice *i*) e o identificador temporal (*ANO*, representado no modelo pelo índice *t*). Esta definição é fundamental para que os modelos estimem corretamente os efeitos fixos ou aleatórios ao longo do tempo e entre indivíduos. A criação dos dados em painel foi realizada recorrendo ao código apresentado na Figura 29.

```
dados_painel_2010_2023 <-
  pdata.frame(Painel_2010_2023_Filtrado_13_sem_NA_5_Neutro, index =
  c("ntrab", "ANO"))
summary(dados_painel_2010_2023)
```

Figura 29 - Código preparação para modelo dados em painel

Ainda foi realizada uma verificação do objeto usando o comando *summary()*, o que confirmou que os dados em painel são equilibrados, com 13 a 14 observações por

trabalhador, e que todas as variáveis relevantes se encontram completas (sem valores omissos) e coerentes.

5.4. Verificação da multicolinearidade das variáveis

Outro aspeto fundamental é a avaliação da presença de multicolinearidade entre as variáveis independentes. A multicolinearidade ocorre quando duas ou mais variáveis explicativas estão fortemente correlacionadas, o que pode comprometer a precisão das estimativas dos coeficientes e inflacionar os erros padrão, dificultando a interpretação estatística dos resultados.

A verificação da multicolinearidade, entre as variáveis explicativas, foi feita utilizando ao fator de inflação da variância (VIF - *Variance Inflation Factor*). Para isso, foi estimada uma regressão linear auxiliar. Como o modelo inclui variáveis categóricas, foi usado o VIF Generalizado (GVIF) do pacote *car*, que ajusta o cálculo para variáveis com mais de uma dimensão, garantindo uma avaliação mais precisa da multicolinearidade. Assim, a verificação da multicolinearidade foi realizada com o código apresentado na Figura 30.

```
modelo_lm_vif <- lm(log_rganho_real ~ antig + anos_escolaridade +
  idade_numerica + qualificacao + log_produtividade +
  media_LogRganho_profissao + dim_empresa + sexo + Nacionalidade +
  nut2_emp + caem11 + tipo_contr1, data = dados_painel_2010_2023)
vif(modelo_lm_vif)
```

Figura 30 - Código cálculo da regressão linear e verificação da multicolinearidade

A interpretação habitual do $GVIF^{(1/(2\times GL))}$, que corresponde ao GVIF ajustado pelo número de graus de liberdade (GL, que nas variáveis categóricas corresponde ao número de categoria menos 1, i.e., ao número de variáveis *dummy* associadas), segue os seguintes critérios:

- $GVIF^{1/(2\times GL)} < 5$: ausência de preocupações relevantes;
- $GVIF^{1/(2\times GL)}$ entre 5 e 10: multicolinearidade moderada (que pode afetar a estimação);
- $GVIF^{1/(2\times GL)} > 10$: multicolinearidade severa (potencialmente problemática).

Com base nos valores de $GVIF^{1/(2\times GL)}$ obtidos (Tabela 14), conclui-se que não há indícios de multicolinearidade entre as variáveis explicativas incluídas no modelo, como todos os valores significativamente inferiores ao limiar 5. Deste modo, todas as variáveis

independentes podem ser mantidas na estimação subsequente com segurança, contribuindo para uma análise econométrica completa e informada.

Tabela 14 - Resultados da análise da multicolinearidade do modelo

Variável	GVIF	GL	$GVIF^{1/(2\times GL)}$
anos_escolaridade	5,10	1	2,26
caem11	4,31	18	1,04
qualificacao	3,12	2	1,33
media_LogRganho_profissao	2,66	1	1,63
log_produtividade	1,75	1	1,32
antig	1,60	1	1,27
nut2_emp	1,57	7	1,03
dim_empresa	1,58	3	1,08
idade_numerica	1,48	1	1,22
tipo_contr1	1,22	3	1,03
sexo	1,24	1	1,11
Nacionalidade	1,02	1	1,01

5.5.Modelo de efeitos aleatórios (MEA)

Nesta secção é estimado o modelo de efeitos aleatórios, assumindo que as características individuais não observadas dos trabalhadores estão correlacionadas aleatoriamente com as variáveis explicativas. O modelo foi estimado com o pacote *plm* em R, usando a especificação "*random*" sobre os dados em painel construídos previamente. Deste modo, a especificação do modelo é apresentada na Figura 31.

```
modelo_mea_log_9 <- plm(log_rganho_real ~ antig + anos_escolaridade +
  idade_numerica + qualificacao + log_produtividade +
  media_LogRganho_profissao + dim_empresa + sexo + Nacionalidade +
  nut2_emp + caem11 + tipo_contr1, data = dados_painel_2010_2023,
  model = "random")
summary(modelo_mea_log_9)
```

Figura 31 - Código criação do modelo de efeitos aleatório (MEA)

O modelo foi aplicado a um conjunto de dados em painel com cerca de 12,7 milhões de observações, cobrindo 932.135 indivíduos com 13 a 14 anos de registos.

Na Tabela 15 a variância dos efeitos individuais mostra que a variação total é dividida quase igualmente entre o efeito idiossincrático (48,8%) e o efeito individual (51,2%), o que sugere que a heterogeneidade entre trabalhadores desempenha papel relevante.

Tabela 15 - Resultados estatísticos do modelo MEA

Efeitos	Variância	Desvio Padrão	Partilha
Idiossincráticos	0,0807	0,2840	48,8%
Individual	0,0847	0,2910	51,2%

Na Tabela 16 o valor médio de θ (0,7445) confirma uma estrutura mista de efeitos intra e entre indivíduos.

Tabela 16 - Resultados dos valores θ do modelo MEA

Valores θ					
Mínimo	**1º Quartil**	**Mediana**	**Media**	**3º Quartil**	**Máximo**
0,7387	0,7387	0,7476	0,7445	0,7476	0,7476

Na Tabela 17, cerca de 17,4% da variação no rendimento é explicada pelo conjunto de variáveis incluídas no modelo. O R^2 para modelos com dados em painel, especialmente com um grande número de observações, tende a ser menor do que em modelos de dados seccionais, logo podemos assumir que este é um valor razoável para o nosso modelo. Significância global do modelo com *p-value* < $2{,}2e^{-16}$, indica que o modelo como um todo é estatisticamente significativo.

Tabela 17 - Resultados do R^2 e teste do qui-quadrado do modelo MEA

	Variância
R^2	0,17448
R^2 Ajustado	0,17448
Qui-quadrado (*p-value*)	< $2{,}22e^{-16}$

Quase todos os coeficientes são estatisticamente significativos, o que mostra uma estrutura estável e bem especificada.

5.6.Modelo de efeitos fixos (MEF)

Nesta secção é estimado o Modelo de Efeitos Fixos, que parte do pressuposto de que características não observáveis dos indivíduos, como motivação ou competências interpessoais, estão correlacionadas com as variáveis explicativas do modelo. O modelo foi estimado com o pacote *plm* em R, usando a especificação "*within*" sobre os dados em painel construídos previamente. Deste modo, a especificação do modelo é apresentada na Figura 32.

```
modelo_mef_log_9 <- plm(log_rganho_real ~ antig + anos_escolaridade +
    idade_numerica + qualificacao + log_produtividade +
    media_LogRganho_profissao + dim_empresa + sexo + Nacionalidade +
    nut2_emp + caem11 + tipo_contr1, data = dados_painel_2010_2023,
    model = "within")
summary(modelo_mef_log_9)
```

Figura 32 - Código criação do modelo de efeitos fixos (MEF)

O modelo foi aplicado a um conjunto de dados em painel com cerca de 12,7 milhões de observações, cobrindo 932.135 indivíduos com 13 a 14 anos de registos.

Tabela 18 - Resultados do R^2 e teste F do modelo MEF

	Variância
R^2	0,093147
R^2 Ajustado	0,021342
F-statistic (*p-value*)	$< 2{,}22e^{-16}$

Os resultados da Tabela 18 indicam:

- que cerca de 9.3% (R^2=0,093) da variação no rendimento, após remover os efeitos fixos individuais, é explicada pelas variáveis do modelo. O R^2 Ajustado de 0,021 é muito baixo, o que sugere que, embora o modelo seja estatisticamente significativo, o poder explicativo das variáveis é relativamente limitado.
- Significância global do modelo com *p-value* < $2{,}2e^{-16}$, o que indica que o modelo como um todo é estatisticamente significativo e contribui para explicar a variação na variável dependente.

5.7. Estimação e seleção do modelo final

A escolha entre o modelo de efeitos fixos (MEF) e o modelo de efeitos aleatórios (MEA) é uma etapa necessária na análise de dados em painel. Embora ambos os modelos controlem a heterogeneidade não observada entre indivíduos, a sua diferença essencial está na correlação entre os efeitos não observáveis e as variáveis explicativas, onde o MEF assume que existe correlação, enquanto o MEA assume que não existe.

Para decidir qual dos dois modelos é mais adequado, recorreu-se ao teste de Hausman, cuja hipótese nula assume que os efeitos aleatórios são consistentes e eficientes. A rejeição da hipótese nula favorece o uso do modelo de efeitos fixos uma vez que a estimação do modelo de efeitos aleatórios não é consistente, inviabilizando a sua aplicação. Deste modo, a aplicação do teste de Hausman é efetuada conforme apresentado na Figura 33.

```
test_Haus <- phtest(modelo_mef_log_9, modelo_mea_log_9)
print(test_Haus)
```

Figura 33 - Código aplicação do teste de Hausman

Dado que o *p-value* do teste de Hausman (Tabela 19) é inferior a um nível de significância de 5% (0,05), a hipótese nula é claramente rejeitada. Assim, os efeitos não observados estão correlacionados com as variáveis explicativas, tornando a estimação do modelo de efeitos aleatórios inconsistente neste contexto. O modelo de efeitos fixos é então o mais adequado para a análise final. Na Secção 5.8 será avaliada a presença de heterocedasticidade e proceder-se-á à estimação final com erros padrão robustos, garantindo maior fiabilidade nas conclusões estatísticas.

Tabela 19 - Resultados do teste de *Hausman*

Teste de *Hausman*		
Chisq	**Graus de liberdade (df)**	***p-value***
448.835	40	$< 2{,}2e^{-16}$

5.8. Verificação de heterocedasticidade e estimação com erros padrão robustos

Após a seleção do modelo MEF, como justificado na secção anterior, é fundamental verificar se os resíduos do modelo apresentam heterocedasticidade, isto é, variância não constante, o que compromete a validade das conclusões estatísticas baseadas nos erros padrão clássicos. A heterocedasticidade foi testada através do teste de *Breusch-Pagan*, aplicado ao modelo de efeitos fixos, conforme Figura 34.

```
bptest(modelo_mef_log_9, studentize = TRUE)
```

Figura 34 - Código da aplicação do teste de *Breusch-Pagan*

O *p-value* do teste de Breusch-Pagan (Tabela 20) é inferior a um nível de significância (0,05), o que indica que é estatisticamente significativa a heterocedasticidade nos resíduos. Este resultado invalida a suposição de homocedasticidade e exige a correção dos erros padrão.

Tabela 20 - Resultados do teste de Breusch-Pagan

Teste de Breusch-Pagan		
BP	**Graus de liberdade (df)**	***p-value***
70.890	40	$< 2{,}2e^{-16}$

Para corrigir este facto, os coeficientes do modelo foram reestimados com erros padrão robustos à heterocedasticidade, utilizando o estimador de Arellano (HC1), conforme Figura 35.

```
modelo_mef_log_9_rob <- vcovHC(modelo_mef_log_9, method = "arellano",
  type = "HC1", cluster = "group")
coeftest(modelo_mef_log_9, modelo_mef_log_9_rob)
```

Figura 35 - Código do cálculo da matriz de covariância robusta

Essa correção garante que as nossas conclusões sobre a significância estatística das variáveis sejam fiáveis, tornando o modelo mais robusto.

5.9. Estimação final e interpretação dos coeficientes

Após a verificação da heterocedasticidade e a subsequente correção com erros padrão robustos, a Tabela 21 apresenta as estimativas dos principais coeficientes do modelo de efeitos fixos, ajustados para permitir inferência estatística consistente. A tabela completa é apresentada em anexo na Tabela A.2.

Tabela 21 - Estimativas dos principais coeficientes do modelo

Variável	Estimativa	*p-value*	Significância
Undereducated	0,0039	0,0001042	***
Overeducated	0,0237	$< 2{,}2e^{-16}$	***
Rendimento Medio por Profissão (log)	0,1760	$< 2{,}2e^{-16}$	***
Anos de Escolaridade	0,0046	$< 2{,}2e^{-16}$	***
Antiguidade	-0,0011	$< 2{,}2e^{-16}$	***
Idade	0,0181	$< 2{,}2e^{-16}$	***
Sexo - Mulher	-0,0036	0,1523047	
Nacionalidade - Estrangeiro	-0.0131	$2.304e^{-07}$	***
Contrato de Trabalho com Termo Certo	-0.0660	$< 2{,}2e^{-16}$	***
Contrato de Trabalho com Termo Incerto	-0.0738	$< 2{,}2e^{-16}$	***
Contrato de Trabalho - Outra situação	-0.0731	$< 2{,}2e^{-16}$	***
Produtividade (log)	0.0410	$< 2{,}2e^{-16}$	***
Pequena empresa	0.0643	$< 2{,}2e^{-16}$	***
Média empresa	0.0982	$< 2{,}2e^{-16}$	***
Grande empresa	0.1043	$< 2{,}2e^{-16}$	***

Legenda dos níveis de significância estatística: **** → muito significativo; ** → significativo a 1%; * → significativo a 5%; . → marginalmente significativo a 10%; espaço em branco → não significativo.*

A variável dependente é o logaritmo do rendimento real mensal (*log_rganho_real*), o que implica que os coeficientes devem ser interpretados como variações percentuais aproximadas no rendimento real[1], associadas a uma unidade de variação na variável explicativa, mantendo as restantes constantes (*ceteris paribus*).

[1] Em rigor, a interpretação deveria ser efetuada considerando a exponencial das estimativas dos coeficientes, pois a variação percentual do rendimento real (supondo as restantes variáveis constantes) será dada por $\left(e^{\beta_i} - 1\right) \times 100\%$. Todavia, como para valores de x próximos de zero verifica-se $e^{x} - 1 \cong x$, então a variação percentual do rendimento real será aproximadamente dada por $\beta_i \times 100\%$. Deste modo, para valores $|\beta_i| \leq 0{,}1$ a aproximação é muito boa, sendo a aproximação razoável para valores de β_i no intervalo $0{,}1 < |\beta_i| < 0{,}2$.

Os principais resultados apresentam-se de seguida:

- Os trabalhadores *overeducated* recebem em média +2.4% face ao grupo de referência (*match*), e este diferencial mantém-se mesmo após o controlo pelos anos de escolaridade. Ou seja, comparando trabalhadores com igual nível de escolaridade, aqueles que estão empregados em ocupações que requerem menos estudos continuam a auferir salários superiores. De notar que este resultado decorre na inclusão da variável do rendimento médio da profissão, a qual permite interpretar os salários dos trabalhadores como um desvio relativamente à média da sua ocupação. Assim, a análise foca-se nas diferenças dentro das profissões, que mostra que no "interior" de uma ocupação, os *overeducated* auferem um prémio salarial. No entanto, quando essa variável é removida do modelo, passando a comparação a refletir diferenças entre profissões, o coeficiente associado à *overeducation* torna-se negativo (–1,3%), em linha com a literatura internacional, que tende a identificar prémios salariais mais baixos para os *overeducated* após o controlo da escolaridade [18, 60].
- No caso dos *undereducated*, observa-se igualmente um ganho marginal (+0,4%). Este resultado significa que, controlando pelo número de anos de escolaridade, estes trabalhadores beneficiam por estarem afetos a ocupações mais exigentes e mais bem remuneradas. Assim, não se trata de "ganharem mais por terem menos escola", mas de auferirem rendimentos relativamente superiores quando comparados com trabalhadores com o mesmo nível de escolaridade em profissões menos exigentes [20, 30, 61].
- A inclusão da variável do rendimento médio da profissão permite que a variável dependente seja interpretada como o desvio do salário do trabalhador face à média da sua profissão. Assim, tanto os diferenciais dos *overeducated* como os dos *undereducated* devem ser interpretados da seguinte forma: trata-se de diferenças relativas dentro da profissão, e não de efeitos absolutos entre grupos profissionais distintos.
- Os anos de escolaridade revelam um coeficiente positivo e estatisticamente significativo, o que confirma o efeito esperado do capital humano formal. Cada ano adicional de escolaridade está associado a ganhos salariais superiores, em linha com a teoria de Mincer [3].

- A variável antiguidade apresenta um coeficiente negativo e estatisticamente significativo, que é contraintuitivo. Este resultado pode justificar-se, em primeiro lugar, ao facto de após o controlo pela idade, a antiguidade tender a perder o seu efeito esperado, sugerindo que a progressão salarial está mais associada a mudanças de emprego do que à permanência na mesma empresa. Em segundo lugar, a própria variável apresenta limitações na base de dados, onde se verificou que cerca de 18% dos trabalhadores que mudam de empresa surgem registados com valores de antiguidade que variam entre 2 e 58 anos, quando seria de esperar uma reinicialização para zero. Este problema metodológico pode ajudar a explicar o sinal negativo observado, pelo que a variável será mantida no modelo (por ser significativa), não sendo interpretada nos resultados seguintes.
- A variável idade apresenta um coeficiente positivo e estatisticamente significativo, refletindo o impacto da experiência acumulada no mercado de trabalho. Este resultado está em linha com a literatura, que associa maior experiência a rendimentos salariais mais elevados [30, 60].
- As mulheres, mesmo após controlo por múltiplos fatores, apresentam um coeficiente negativo (≈ -0,36%), não estatisticamente significativo ao nível de 5%. Este resultado pode refletir, em parte, o facto de a disparidade salarial de género ser explicada pelas restantes variáveis de controlo. Contudo, importa notar que, no modelo de efeitos fixos, variáveis invariantes no tempo, como o sexo, tendem a perder significância estatística, uma vez que o efeito fixo individual absorve grande parte da sua variabilidade [62, 63].
- A nacionalidade estrangeira influência negativamente o valor do rendimento, o que poderá refletir segmentações ou desvantagens acumuladas no mercado [63, 64].
- Os contratos sem termo continuam associados a salários mais altos. Os vínculos a termo certo e incerto estão associados a penalizações salariais médias de -6,6% e -7,4%, respetivamente, refletindo a menor estabilidade e o poder negocial reduzido dos trabalhadores com vínculos precários [65].
- A variável produtividade da empresa apresenta um efeito positivo relevante (≈ 4,1%), sugerindo que trabalhadores em empresas onde a produtividade do trabalho é mais elevada tendem a receber salários superiores, mesmo controlando para outras variáveis [16].

- A dimensão da empresa continua a ter um papel determinante, sendo que quanto maior esta dimensão maior será tendencialmente o valor dos salários pagos aos trabalhadores. O destaque vai para as grandes empresas, com um prémio salarial de +10,4% quando comparado com as microempresas. Este resultado é confirmado pela literatura que associa a dimensão empresarial a maior capacidade de remuneração [66, 67].
- Quando comparados com o setor "Agricultura, produção animal, caça, floresta e pesca", a maioria dos restantes revela efeitos estatisticamente significativos, com setores como "atividades financeiras", "administração pública" ou "indústrias extractivas" a apresentarem os maiores prémios salariais [68].
- Empresas localizadas nas NUTS II de Lisboa e Vale do Tejo, Madeira e Algarve apresentam geralmente salários mais elevados [69].

6. Análise por regressão linear (dados seccionados por ano)

Este capítulo consiste na aplicação dos modelos de regressão linear estimados para cada um dos anos entre 2010 e 2023. Com esta estimação pretende-se, em primeiro lugar, apresentar uma análise que foque apenas a comparação entre indivíduos, em segundo lugar, traçar a evolução temporal dos coeficientes das variáveis-chave, como a escolaridade e a qualificação, e finalmente permitir uma comparação entre os resultados obtidos através desta modelação e os obtidos por via da análise em dados em painel. Para garantir a fiabilidade das conclusões, começamos por realizar um diagnóstico rigoroso dos modelos, avaliando a multicolinearidade e a heterocedasticidade. Após a devida correção para a heterocedasticidade, prosseguimos com a interpretação dos resultados obtidos com base nos dados recolhidos.

6.1. O modelo de regressão linear

A regressão linear é uma ferramenta estatística fundamental para analisar relações entre uma variável dependente contínua e um conjunto de variáveis independentes. O seu objetivo é estimar o efeito marginal de cada variável explicativa sobre a variável dependente, assumindo uma relação linear entre elas [58, 70].

Neste capítulo, a análise será realizada por meio de modelos seccionais anuais, isto é, será estimada uma regressão linear distinta para cada ano entre 2010 e 2023. Esta abordagem permite explicar as diferenças nos salários dos trabalhadores em cada ano com base nas suas características demográficas, educacionais, profissionais e da empresa em que estão inseridos, possibilitando ainda observar a evolução temporal dos coeficientes estimados.

A estimação será feita através do método dos mínimos quadrados ordinários (OLS), de acordo com os pressupostos clássicos. Quando esses pressupostos forem violados (ex.: presença de heterocedasticidade), serão aplicadas correções robustas nos erros padrão, garantindo a validade da inferência estatística [70].

Embora esta abordagem não permita controlar diretamente a heterogeneidade individual não observada, como nos modelos com dados em painel, oferece a vantagem de avaliar o impacto

das variáveis explicativas num contexto estático e anual, complementando a análise efetuada no Capítulo 5.

6.2.Especificação formal do modelo de regressão (dados seccionados)

O modelo estimado pode ser representado pela seguinte equação, considerando que $i = 1, \ldots, n$ (número de indivíduos na amostra, i.e. n=932.135):

$$\begin{aligned} log\_rganho\_real_i = {} & \beta_0 + \beta_1\, antig_i + \beta_2\, anos\_escolaridade_i + \beta_3\, idade\_numerica_i \\ & + \beta_4\, qualificacaoUnder_i + \beta_5\, qualificacaoOver_i + \beta_6\, log\_produtividade_i \\ & + \beta_7\, media\_LogRganho\_profissao_i + \beta_8\, dim\_empresaPequena_i + \beta_9\, dim\_empresaMédia_i \\ & + \beta_{10}\, dim\_empresaGrande_i + \beta_{11}\, sexoMulher_i + \beta_{12}\, nacionalidadeEstrangeiro_i \\ & + \beta_{13}\, nut2\_empAlgarve_i + \ldots + \beta_{19}\, nut2\_empEstrangeiro_i + \beta_{20}\, caem1lB_i + \ldots \\ & + \beta_{37}\, caem1lS_i + \beta_{38}\, tipo\_contrCerto_i + \beta_{39}\, tipo\_contrIncerto_i + \beta_{40}\, tipo\_contrOS_i + \varepsilon_i \end{aligned}$$

onde o erro ε_i é uma variável aleatória com distribuição normal com média zero e desvio padrão σ. A definição das variáveis utilizadas nesta regressão mantém-se idêntica à apresentada para os modelos de efeitos fixos e aleatórios, descritos no Capítulo 5.2.

6.3.Modelação e variáveis incluídas

A presente secção descreve o processo de estimação dos modelos de regressão linear múltipla aplicados a cada ano entre 2010 e 2023. O objetivo desta abordagem transversal é captar a relação entre rendimento real e um conjunto de determinantes estruturais e individuais, observando a sua estabilidade ou evolução ao longo do tempo.

```
dados_YYYY <- subset(dados_painel_2010_2023_13_sem_NA_Neutro_2,
    ANO == YYYY)
modelo_YYYY <- lm(log_rganho_real ~ antig + anos_escolaridade +
    idade_numerica + qualificacao + log_produtividade +
    media_LogRganho_profissao + dim_empresa + sexo + Nacionalidade +
    nut2_emp + caem1l + tipo_contr1, data = dados_YYYY)
```

Figura 36 - Código exemplo do cálculo da regressão linear

Para cada ano, foi filtrado um subconjunto da base de dados longitudinal, contendo apenas as observações relativas a esse ano. A estrutura base da modelação manteve-se constante, utilizando o modelo descrito na Figura 36. Deste modo, os dados utilizados neste capítulo

são os mesmo que foram utilizados no Capítulo 5, de forma a ser possível comparar os resultados obtidos.

Esta fórmula será aplicada para todos os anos, permitindo uma comparação direta com os resultados obtidos utilizando dados em painel. A variável dependente é o logaritmo do rendimento real mensal, o que permite interpretar os coeficientes estimados como variações percentuais aproximadas no salário real. Além da estimação básica com o método dos mínimos quadrados ordinários, foram aplicados testes complementares em cada modelo para garantir a robustez dos resultados:

- Multicolinearidade: verificada com o cálculo do *Variance Inflation Factor* (VIF), usando o pacote *car* [40].
- Heterocedasticidade: avaliada através do teste de Breusch-Pagan, que testa a hipótese de variância constante dos resíduos.
- Normalidade dos resíduos: para verificar se os resíduos seguem uma distribuição normal, foi utilizado o teste de Anderson-Darling (*ad.test*) [42]. Essa suposição é fundamental para garantir a validade dos intervalos de confiança e dos testes estatísticos do modelo.
- Autocorrelação dos resíduos: para detetar a presença de autocorrelação nos resíduos, foi aplicado o teste de Durbin-Watson (*dwtest*) [39]. A ausência de correlação entre os resíduos é uma suposição importante para que as inferências estatísticas do modelo sejam confiáveis.
- Erros padrão robustos: sempre que detetada heterocedasticidade (*p-value* < 0,05, no teste de Breusch-Pagan), os coeficientes foram reestimados com erros padrão robustos à heterocedasticidade, utilizando o estimador HC1, e os resultados foram reportados com a função *coeftest()* [39].

Esta metodologia é exemplificada na Figura 37.

```
vif(modelo_YYYY)
bptest(modelo_YYYY)
ad.test(residuals(modelo_YYYY))
dwtest(modelo_YYYY)
coeftest(modelo_YYYY, vcov = vcovHC(modelo_YYYY, type = "HC1"))
```

Figura 37 - Código exemplo dos vários testes de verificação e aplicação dos erros padrão robusto

Este procedimento foi repetido sistematicamente para cada ano, garantindo coerência metodológica e permitindo analisar tendências na relação entre qualificações e rendimento ao longo do tempo. A utilização de modelos seccionais também oferece uma perspetiva complementar ao modelo de efeitos fixos estimado anteriormente, permitindo verificar se os efeitos identificados nos dados em painel se mantêm quando analisados isoladamente ano a ano.

6.4.Diagnóstico do modelo: multicolinearidade e heterocedasticidade

Antes de interpretar os coeficientes das regressões lineares para cada ano entre 2010 e 2023, foi realizada uma verificação dos pressupostos fundamentais: a ausência de multicolinearidade entre variáveis explicativas, a homocedasticidade dos resíduos, a normalidade dos resíduos e a ausência de autocorrelação.

6.4.1. Multicolinearidade: análise do teste VIF

Para avaliar a multicolinearidade, foi calculado o VIF generalizado ajustado pelo número de graus de liberdade $GVIF^{1/(2\times GL)}$ para todas as variáveis de cada modelo anual. Os resultados mostram que:

- Em todos os anos analisados, nenhuma variável excedeu o valor crítico de 5, critério geralmente usado como sinal de multicolinearidade;
- A variável *anos_escolaridade* apresenta sistematicamente os $GVIF^{1/(2\times GL)}$ mais elevados, com valores próximos de 2, o que pode refletir a sua correlação com variáveis como idade, antiguidade e *qualificacao*;
- Todas as restantes variáveis apresentam valores significativamente inferiores a 2.

Em suma, não há evidência de multicolinearidade preocupante nos modelos anuais. A estrutura das variáveis foi mantida de forma consistente em todos os anos, assegurando a robustez da comparação temporal.

6.4.2. Heterocedasticidade: teste de *Breusch-Pagan*

A heterocedasticidade foi verificada com o teste de Breusch-Pagan aplicado aos resíduos dos modelos anuais. Os resultados revelam que:

- Todos os anos apresentam um *p-value* muito inferior a 0,05 (sempre inferior a 0,001), indicando forte evidência de heterocedasticidade;
- Os valores da estatística do teste variam anualmente, mas permanecem elevados, confirmando que a variância dos resíduos não é constante.

Esta violação do pressuposto de homocedasticidade não invalida os coeficientes estimados, mas compromete a validade dos erros padrão clássicos e, por consequência, os testes de significância. Em resposta a esta evidência, foi aplicado o método de correção robusta aos erros padrão na estimação final dos coeficientes.

6.4.3. Normalidade dos resíduos: teste de Anderson-Darling

A normalidade dos resíduos foi avaliada através do teste de Anderson-Darling. Pelos resultados obtidos, podemos concluir que em todos os anos, os *p-value* foram extremamente baixos ($\approx < 2{,}2\times10^{-16}$), rejeitando claramente a hipótese nula de normalidade. Estes resultados indicam que a distribuição dos resíduos se afasta de uma distribuição normal, o que pode afetar a validade de alguns testes paramétricos.

Como a distribuição dos resíduos não é normal, usaremos erros padrão robustos para garantir que os resultados sejam confiáveis.

6.4.4. Autocorrelação dos resíduos: teste de Durbin-Watson

A autocorrelação dos resíduos foi testada com o teste de Durbin-Watson. Os valores da estatística DW situam-se entre 1,95 e 1,98 ao longo dos anos, muito próximos do valor de referência 2, sugerindo ausência relevante de autocorrelação de primeira ordem. Embora os *p-values* sejam muito baixos, tal resultado é explicado pelo grande tamanho da amostra, que leva a detetar como estatisticamente significativas autocorrelações de magnitude mínima. De facto, a estimativa de $\rho \approx 1-DW/2$ assume valores muito reduzidos, entre 0,009 e 0,021, sendo praticamente nula na prática.

As figuras em anexo (Figura A.3 e Figura A.4), referentes aos dois últimos anos (2022 e 2023), reforçam esta interpretação, evidenciando que, ao longo de todo o intervalo de valores preditos, os resíduos apresentam padrões alternados de sinais positivos e negativos, sem tendência clara ou associação sistemática. Os gráficos obtidos para os restantes anos

revelaram sempre o mesmo comportamento. Assim, apesar de o teste indicar autocorrelação, a magnitude observada é negligenciável e não compromete a validade das estimativas.

6.5.Evolução temporal dos coeficientes (2010–2023)

Após a deteção de heterocedasticidade nos modelos de regressão linear, procedeu-se à reestimação dos coeficientes com erros padrão robustos, recorrendo ao estimador de Huber-White. Esta abordagem garante que os testes de significância estatística dos coeficientes sejam válidos, mesmo na presença de variância não constante dos resíduos [58].

A análise incide sobre as mesmas variáveis consideradas na interpretação dos resultados com dados em painel na Secção 5.9, mantendo a coerência temática. As variáveis em foco são: qualificação, escolaridade, antiguidade, idade, sexo, nacionalidade, tipo de contrato, dimensão da empresa, região e setor de atividade económica.

6.5.1. Qualificação

Os coeficientes associados à variável qualificação apresentam-se estatisticamente significativos em todos os anos do período analisado. O sinal confirma que, em média, trabalhadores *overeducated* auferem salários superiores, enquanto os *undereducated* recebem salários inferiores relativamente aos adequadamente qualificados (o coeficiente associado à *undereducated* é positivo, mas como têm pelo menos dois anos de escolaridade a menos, tenderão a receber menos se juntarmos o efeito destas duas variáveis), após controlo pelas restantes variáveis.

A Figura 38 e Figura 39 mostra a evolução destes coeficientes, evidenciando que, sobretudo a partir de 2015, o impacto estimado da condição de *overeducation* e de *undereducation* se intensificou de forma consistente.

Para os *overeducated*, o coeficiente cresce de forma mais acentuada, quase duplicando entre 2013 e 2022, o que reforça a ideia de que a *overeducation* passou a ser mais valorizada no mercado, possivelmente pela escassez de competências especializadas e pela crescente complexidade de determinadas funções.

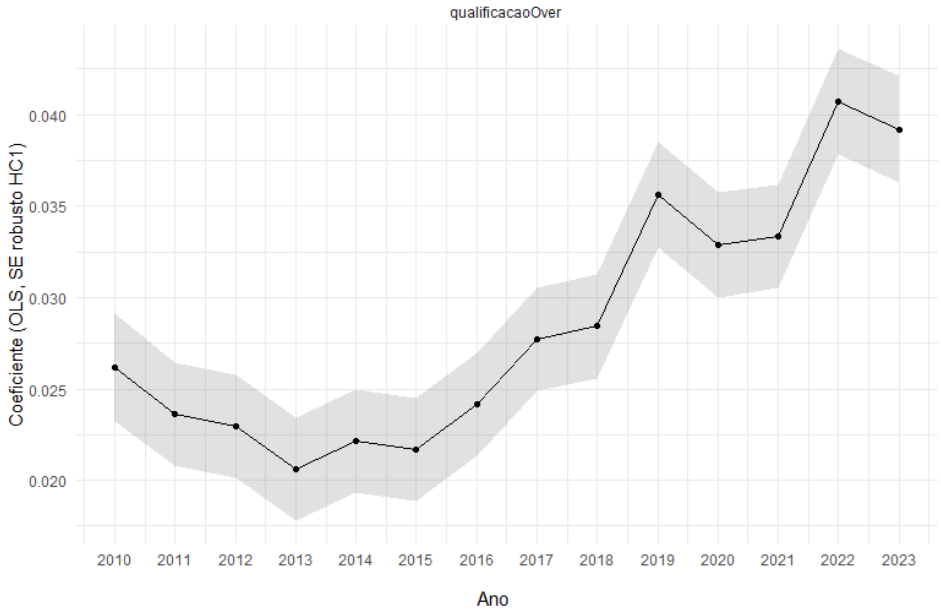

Figura 38 - Evolução do coeficiente ao longo dos anos da variável qualificação (*over*)

Nos *undereducated*, embora os valores absolutos sejam mais baixos, observa-se também uma tendência de valorização até 2021, o que pode indicar contextos em que a experiência prática e a antiguidade compensam parcialmente a menor escolaridade.

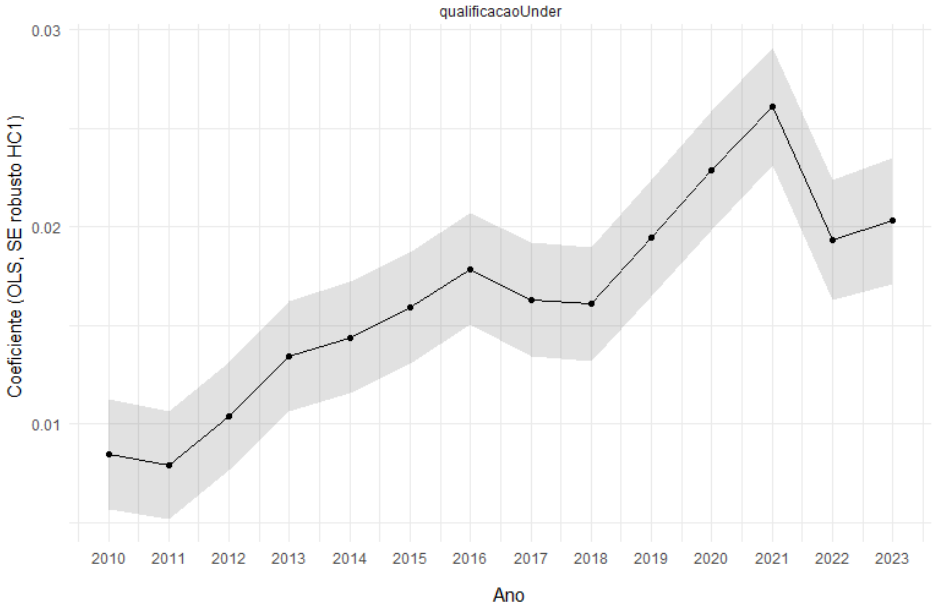

Figura 39 - Evolução do coeficiente ao longo dos anos da variável *qualificação (under)*

Contudo, é importante reiterar que o valor positivo desta variável deve ser interpretado em conjunto com a variável "anos de escolaridade". Assim, o resultado não implica que os *undereducated* recebam mais do que os trabalhadores *match*, mas que, para este grupo,

outros fatores, como experiência, contribuem para ganhos salariais adicionais, ainda que partindo de uma base mais baixa.

Em ambos os casos, há uma ligeira redução do coeficiente em 2022, mas os níveis mantêm-se superiores aos do início do período, sugerindo um prémio crescente associado à adequação da qualificação face ao posto de trabalho.

6.5.2. Escolaridade

A Figura 40 mostra a evolução do coeficiente associado aos anos de escolaridade que se apresenta positivo e estatisticamente significativo em todos os anos do período analisado.

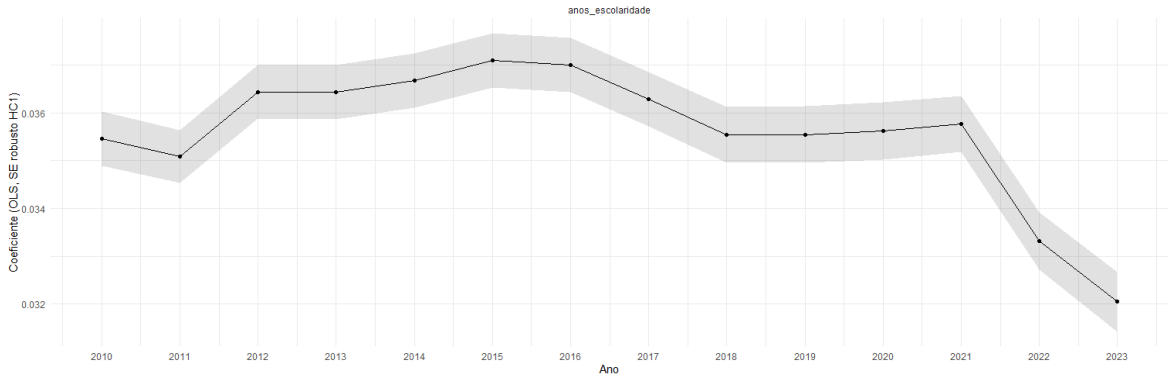

Figura 40 - Evolução do coeficiente ao longo dos anos da variável *escolaridade*

Isto confirma que trabalhadores com mais anos de estudo auferem, em média, salários superiores aos de colegas com menor escolaridade, mesmo após controlo pelas restantes variáveis do modelo. Entre 2010 e 2021, o coeficiente manteve-se relativamente estável, com valores a oscilar entre 0,035 e 0,037, indicando que cada ano adicional de escolaridade esteve associado a diferenças salariais médias na ordem dos 3,5% a 3,7%.

A partir de 2022, observa-se uma quebra, com o coeficiente a recuar para cerca de 0,032 em 2023. Uma possível explicação é que, como cada vez mais trabalhadores têm níveis de escolaridade elevados, as diferenças salariais médias entre graus de ensino ficam menores. Outra hipótese é que o aumento dos casos de *overeducation* reduza o prémio salarial associado a cada ano adicional de estudo.

6.5.3. Idade

O coeficiente associado à idade apresenta-se positivo e estatisticamente significativo em todos os anos do período analisado, indicando que trabalhadores mais velhos auferem, em média, salários superiores aos mais jovens, após controlo pelas restantes variáveis. Ainda assim, e conforme o que se observa na Figura 41, a evolução do coeficiente associado à variável idade mostra uma tendência clara de declínio ao longo do período 2010 a 2023.

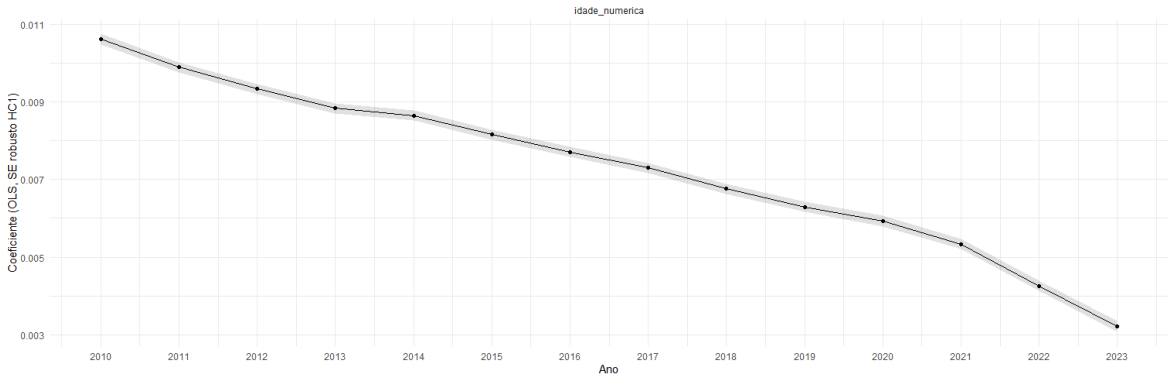

Figura 41 - Evolução do coeficiente ao longo dos anos da variável *idade*

Em 2010, o impacto marginal da idade sobre o rendimento real era de aproximadamente 0,011, sugerindo que cada ano adicional de idade estava associado a um salário mais elevado, após controlo pelas restantes variáveis. A partir desse ponto, verifica-se uma diminuição gradual e consistente, atingindo cerca de 0,0032 em 2023. Esta redução não significa uma desvalorização direta da experiência acumulada com a idade. O declínio do coeficiente resulta em grande parte da forma como a base de dados está definida. Ou seja, é de esperar que a idade, refletindo a valorização da experiência, impacte mais os salários nos primeiros anos do trabalhador no mercado de trabalho, sendo que para anos mais avançados, o efeito é positivo mas decrescente (o que vai ao encontro das evidências de Mincer considerando a experiência) [2, 6].

6.5.4. Sexo

Na Figura 42, o coeficiente associado à variável sexo (1 se mulher; 0 caso contrário), mantém-se consistentemente negativo e estatisticamente significativo ao longo de todo o período, confirmando a existência de uma diferença salarial desfavorável às mulheres, mesmo após controlar o efeito através de outras variáveis.

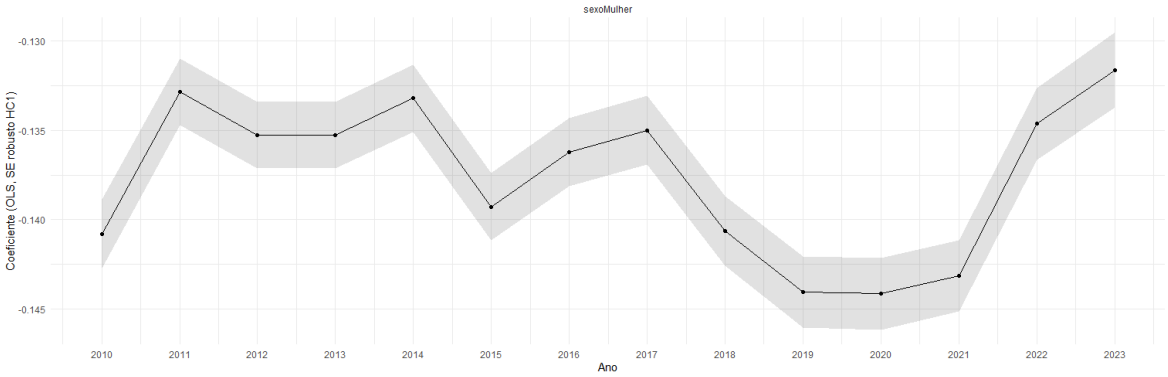

Figura 42 - Evolução do coeficiente ao longo dos anos da variável *sexo*

Entre 2010 e 2014, observa-se uma ligeira redução desta penalização, mas o diferencial volta a acentuar-se entre 2015 e 2019, atingindo o valor mais negativo nesse ano. A partir de 2020, verifica-se uma tendência de recuperação, com diminuição da penalização até 2023, embora sem eliminar a diferença salarial. Este padrão sugere que, apesar de algumas melhorias recentes, a desigualdade de género no rendimento persiste de forma estrutural, sendo pouco provável que se explique apenas por características observáveis no modelo [71, 72].

Importa salientar que, devido à definição da amostra, restrita a trabalhadores com 13 ou 14 anos de registo, foram exclusas as novas entradas no mercado de trabalho. A literatura internacional mostra que as novas gerações de trabalhadores tendem a apresentar diferenças salariais de género menores à entrada no mercado [73]. Em Portugal, os dados oficiais confirmam esta dinâmica, onde as diferenças salariais entre homens e mulheres são mais reduzidas no início da carreira, mas aumentam de forma significativa à medida que cresce a antiguidade no emprego [74].

6.5.5. Nacionalidade

O coeficiente associado à variável Nacionalidade (Estrangeiro), na Figura 43, apresenta-se negativo e estatisticamente significativo entre 2010 e 2013, evidenciando que, nesses anos, trabalhadores estrangeiros auferiam rendimentos inferiores aos nacionais, mesmo após controlo por fatores como sexo, escolaridade, antiguidade, setor e dimensão da empresa. Entre 2014 e 2019, embora o coeficiente se mantenha negativo, a sua magnitude diminui progressivamente, sugerindo uma redução gradual da penalização salarial.

A partir de 2020, o coeficiente aproxima-se de zero e perde significância estatística, deixando de haver evidência robusta de diferenças salariais entre nacionais e estrangeiros.

Em 2022 e 2023, os *p-values* atingem valores muito elevados (entre 0,6 e 0,8), e em 2023 o coeficiente situa-se praticamente em zero, indicando que a desigualdade remuneratória associada à nacionalidade desapareceu no período mais recente.

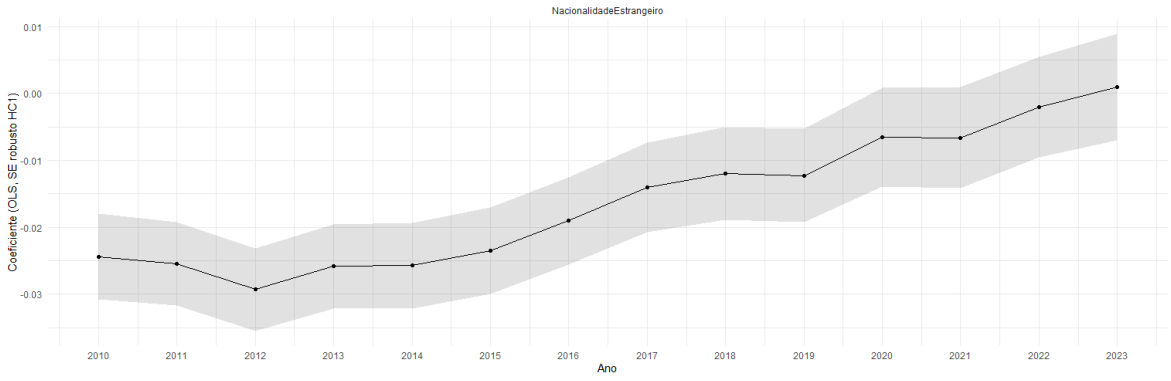

Figura 43 - Evolução do coeficiente ao longo dos anos da variável *nacionalidade*

Esta evolução sugere um processo de convergência salarial ao longo da última década, possivelmente associado a uma maior integração dos trabalhadores estrangeiros em setores e funções mais qualificados, bem como a mudanças no enquadramento legal e institucional. Importa salientar que, na base de dados utilizada, todos os trabalhadores estrangeiros considerados têm pelo menos 13 anos de registo, o que significa que representam indivíduos com presença estável e prolongada no mercado de trabalho português. Essa continuidade poderá contribuir para explicar a redução das disparidades, uma vez que, após vários anos no país, estes trabalhadores parecem ter acesso a oportunidades salariais semelhantes às dos nacionais.

6.5.6. Tipo de contrato

Os coeficientes associados ao tipo de contrato apresentam-se negativos e estatisticamente significativos em todos os anos do período analisado, confirmando a penalização salarial consistente dos vínculos não permanentes face ao contrato de trabalho sem termo (categoria de referência). A Figura A.5, em anexo, mostra a evolução destes coeficientes, revelando padrões diferenciados por categoria contratual:

- Contratos de trabalho com termo certo apresentam um prémio salarial negativo relativamente estável, oscilando entre -0,10 e -0,05 ao longo dos anos. A penalização

suavizou-se gradualmente até cerca de 2020, mas voltou a intensificar-se ligeiramente após esse período.

- Contratos de trabalho com termo incerto exibem a penalização mais acentuada entre 2010 e 2016 (cerca de -0,10 a -0,08), seguida de uma redução até 2021, atingindo o valor mais "suave" do período (-0,06), antes de voltar a agravar-se até 2023.
- Outras situações contratuais apresentam elevada volatilidade e penalizações muito variáveis, com episódios de quase neutralidade (2014 a 2016) mas também quedas expressivas (como em 2021 e 2023, abaixo de -0,15).

Embora a diferença negativa face aos contratos sem termo se tenha atenuado nalguns períodos, a precariedade contratual continua a traduzir-se em perdas salariais relevantes e persistentes, sendo mais estável nos contratos com termo certo, ligeiramente menos penalizada nos contratos com termo incerto nos últimos anos, e mais instável no grupo "outra situação".

6.5.7. Produtividade

O coeficiente da produtividade (Figura 44) apresenta-se positivo e estatisticamente significativo em todos os anos do período analisado, confirmando a sua relação direta com o rendimento real dos trabalhadores.

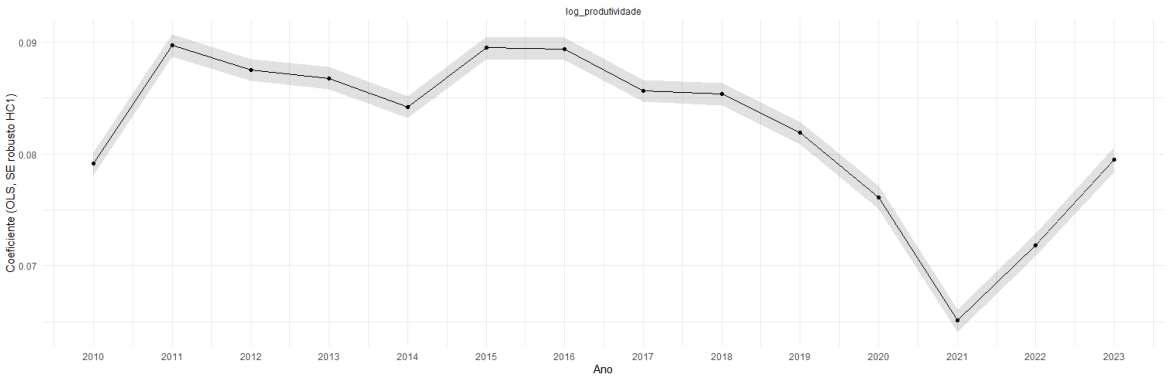

Figura 44 - Evolução do coeficiente ao longo dos anos da variável *produtividade*

Entre 2010 e 2016, observa-se uma ligeira valorização desta variável, seguida de um decréscimo gradual até 2021, quando atinge o valor mínimo da série (0,065). Nos dois anos seguintes regista-se recuperação, aproximando-se novamente dos níveis iniciais.

Apesar das variações anuais, a produtividade permanece entre os determinantes mais relevantes do rendimento, reforçando o papel estrutural da eficiência económica das empresas na definição salarial.

6.5.8. Dimensão da empresa

Os coeficientes associados à dimensão da empresa apresentam-se positivos e estatisticamente significativos em todos os anos do período analisado, confirmando que trabalhadores em empresas de maior dimensão auferem salários mais elevados do que os que trabalham em microempresas, após controlo pelas restantes variáveis.

A Figura A.6, em anexo, correspondente à evolução dos coeficientes associados à dimensão da empresa tendo como referência a categoria microempresa, mostra um padrão consistente: trabalhadores em empresas maiores tendem a receber um prémio salarial superior. Essa vantagem manteve-se ou até aumentou ao longo do período analisado.

- Nas grandes empresas, apesar de uma ligeira queda inicial até 2012, observa-se uma recuperação contínua e expressiva até 2020, estabilizando depois em valores próximos de 0,19, o que sugere um prémio salarial estável e elevado neste segmento face aos trabalhadores em microempresas.
- O coeficiente associado às médias empresas registou um crescimento acentuado entre 2012 e 2015, atingindo um pico em 2021 (0,21). Embora haja uma pequena descida até 2023, o efeito permanece forte e consistente.
- Nas pequenas empresas, embora com coeficientes inferiores, o padrão é semelhante, com um aumento até meados da década e posterior estabilização, apresentando alguma volatilidade nos últimos anos.

O conjunto destes resultados reforça a evidência de que o tamanho da empresa está fortemente associado a melhores condições salariais, possivelmente devido à maior capacidade financeira, oportunidades de progressão e acesso a setores de maior valor acrescentado.

6.5.9. Região (NUTS II)

A análise dos coeficientes associados às regiões NUTS II mostra que estes se apresentam estatisticamente significativos em todos os anos do período analisado, evidenciando a existência de diferenciais salariais relevantes consoante a localização da empresa.

A Figura A.7, em anexo, ilustra estes padrões, sempre comparados com a região de referência no modelo, o Norte. Abaixo destacam-se os principais *insights* regionais:

- Lisboa e Vale do Tejo apresenta consistentemente o coeficiente mais elevado, embora com tendência de redução ao longo do tempo (de cerca de 0,12 em 2010 para 0,07 em 2023), o que indica uma convergência parcial com a região de referência.
- Madeira também exibe prémios salariais positivos e relativamente altos, mas com decréscimo gradual desde 2010.
- Algarve e Açores mantêm coeficientes positivos, mas apresentam maior volatilidade, com o Algarve a mostrar uma quebra mais marcada entre 2010 e 2014 e recuperação parcial nos anos seguintes.
- Alentejo segue um padrão de declínio suave, mantendo-se positivo, mas menos expressivo em 2023.
- A região Centro regista coeficientes próximos de zero ou ligeiramente negativos na maior parte do período, sugerindo que, controlando pelos restantes fatores, não há um prémio salarial face à região de referência.
- O caso de Estrangeiro é atípico, com valores extremamente voláteis, especialmente em 2020, o que pode refletir o impacto do COVID-19.
- Comparando os efeitos da pandemia, observa-se que as regiões do Centro e dos Açores parecem ter sofrido menos impacto relativo do COVID-19 do que a região Norte, mantendo coeficientes mais estáveis no período 2020 a 2021.

Este padrão indica que, apesar da manutenção de alguns diferenciais salariais regionais, existe uma tendência geral de convergência, com a localização geográfica a perder importância relativa na determinação do rendimento ao longo do tempo.

6.5.10. CAE

A Figura A.8, em anexo, apresenta a evolução dos coeficientes associados a cada CAE ao longo dos anos, tendo como referência a CAE - Agricultura, produção animal, caça, floresta e pesca. Os resultados por setor de atividade mostram padrões consistentes de diferenciação salarial ao longo do período 2010–2023, revelando setores com prémios salariais persistentes e outros com penalizações estáveis ou crescentes:

- Setores com prémios salariais estáveis ou crescentes: "Indústrias extrativas" e "Eletricidade, gás, vapor, água quente e fria" destacam-se por manterem coeficientes positivos e predominantemente crescentes, reforçando a ideia de que atividades com elevada intensidade de capital e barreiras à entrada oferecem melhores remunerações. "Atividades de informação e comunicação" também registam tendência positiva recente.
- Setores com penalizações salariais consistentes: "Alojamento, restauração e similares", "Comércio por grosso e a retalho", "Atividades administrativas e dos serviços de apoio" e "Construção" apresentam coeficientes maioritariamente negativos, eventualmente refletindo a concentração de mão de obra menos qualificada e maior rotatividade.
- Setores com ligeira deterioração ao longo do tempo: "Educação" e "Atividades de saúde humana e apoio social" registam reduções graduais dos coeficientes, podendo estar associadas à contenção salarial nestes setores.

De forma geral, a análise confirma que a pertença a determinados setores funciona como um determinante estrutural do rendimento, reforçando desigualdades e refletindo diferenças na qualificação exigida, produtividade média e capacidade de negociação salarial de cada ramo.

6.5.11. Síntese dos resultados obtidos

Os resultados apontam para um deslocamento estrutural na valorização salarial em Portugal. A importância da idade na explicação das diferenças salariais diminuiu, enquanto a escolaridade manteve a sua influência e a variável *overeducated* ganhou relevância. Empresas de maior dimensão e com maior produtividade oferecem prémios salariais consistentes. Persistem desigualdades por género e setor de atividade, enquanto as diferenças por nacionalidade e região tendem a diminuir. Vínculos não permanentes continuam associados a penalizações salariais relevantes.

7. Comparação entre resultados obtidos com dados em painel e seccionados por ano

Este capítulo analisa e compara os resultados de modelos com dados seccionais (obtidos em cada ano) com os resultados obtidos com dados em painel (considerando efeitos fixos).

7.1. Introdução e objetivo

A comparação entre as estimativas obtidas pelos modelos de regressão em cada ano isoladamente e os modelos com dados em painel (efeitos fixos) permite avaliar o impacto de diferentes estratégias de controlo da heterogeneidade não observada [58, 59].

Nas regressões lineares anuais, cada ano é estimado de forma independente, captando apenas diferenças entre indivíduos em função das suas características observáveis naquele momento. Já o modelo de efeitos fixos utiliza a variação temporal dentro do mesmo indivíduo, controlando automaticamente para todas as características invariantes no tempo (como aptidões, motivação ou contexto familiar), mesmo quando estas não são observáveis [58].

Esta diferença metodológica explica porque é que os coeficientes estimados podem divergir. O modelo de efeitos fixos baseia-se também na evolução ao longo do tempo de cada trabalhador. Assim, e uma vez que muitas das variáveis têm pouca alteração ao longo do tempo, as estimativas em painel tendem a apresentar magnitudes menores refletindo tanto o efeito da heterogeneidade não observada como a distinta fonte de variação explorada [59].

7.2. Comparação de magnitude e sinal dos coeficientes

A Tabela A.2, em anexo, apresenta as médias dos coeficientes estimados nos modelos de regressão linear anuais (2010 a 2023) e as estimativas obtidas no modelo de efeitos fixos com dados em painel.

Ao analisar a tabela podemos tirar as seguintes observações principais:

7.2.1. Inversão de sinal

A inversão do sinal é uma situação de maior complexidade que ocorre no caso de algumas variáveis *dummy* que identificam a CAE, como é o caso das Indústrias transformadoras.

Neste setor, o coeficiente médio das regressões anuais é negativo (-0,065), sugerindo penalização salarial face ao setor de referência. Contudo, no modelo em painel o coeficiente torna-se positivo (+0,027), indicando que, controlando as características fixas dos trabalhadores, este setor oferece um prémio salarial. Situação semelhante verifica-se em setores como Captação, tratamento e distribuição de água (-0,087 / +0,024), Atividades imobiliárias (-0,019 / +0,016), Atividades de consultoria, científicas, técnicas e similares (-0,022 / +0,025), Administração pública, defesa e segurança social (-0,009 / +0,093), Educação (-0,016 / +0,046) e Atividades de saúde humana e apoio social (-0,052 / +0,057).

Em todos estes casos, penalizações aparentes nos modelos *cross-section* transformam-se em prémios positivos quando analisada apenas a variação intra-trabalhador ao longo do tempo. Tal sugere que parte do efeito captado nas estimativas anuais não decorre do setor em si, mas de características individuais não observadas, que levam determinados trabalhadores a concentrarem-se em setores mais dinâmicos ou mais estáveis.

7.2.2. Redução da magnitude:

Variáveis como a escolaridade, dimensão da empresa e média do rendimento da profissão apresentam coeficientes substancialmente mais elevados nas regressões lineares anuais do que no modelo de efeitos fixos. Esta diferença reflete não apenas a presença de características fixas não controladas nos modelos *cross-section*, mas também o facto de estas variáveis serem particularmente relevantes para distinguir salários entre trabalhadores diferentes, ao invés de explicar a evolução salarial do mesmo indivíduo ao longo do tempo.

Apesar da redução da magnitude dos coeficientes no modelo de efeitos fixos, todas estas variáveis mantêm-se estatisticamente significativas, confirmando a sua importância estrutural na determinação dos salários. A escolaridade, a dimensão da empresa e o rendimento médio da profissão revelam-se determinantes importantes para explicar diferenças salariais entre trabalhadores, ainda que o seu impacto seja menos expressivo quando a análise se concentra na evolução salarial do mesmo indivíduo ao longo do tempo [75, 76].

7.2.3. Coerência de sinal:

A maioria dos coeficientes mantém o sinal em ambos os métodos, reforçando a consistência dos resultados, embora com magnitudes distintas [58].

7.3.Implicações para variáveis-chave

Após a análise comparativa entre os resultados obtidos com regressões anuais em *cross-section* e os provenientes do modelo com dados em painel, este capítulo aprofunda as implicações das diferenças observadas nas variáveis-chave. O objetivo é compreender de que forma a utilização de dados longitudinais, combinada com o controlo de efeitos fixos característico do modelo, redefine a leitura da relação entre características demográficas, produtividade, estrutura empresarial e rendimento, permitindo uma interpretação mais rigorosa e dinâmica dos determinantes salariais.

7.3.1. Perfil demográfico

A variável idade apresenta um coeficiente positivo e estatisticamente significativo em ambos os métodos, confirmando que trabalhadores mais velhos auferem salários mais elevados do que os mais jovens. Contudo, a magnitude do efeito é maior no modelo de efeitos fixos (0,0181) do que nas regressões *cross-section* (0,0073). Esta diferença indica que, ao controlar pelas características fixas dos trabalhadores, a idade ganha maior relevância, refletindo o impacto da experiência acumulada ao longo do tempo sobre os rendimentos [76].

O coeficiente da variável que identifica as mulheres é negativo em ambos os métodos, mas a penalização é substancialmente menor nos dados em painel (-0,0036 versus -0,1376). Nos modelos *cross-section*, o efeito é estatisticamente significativo, confirmando a existência de uma disparidade salarial entre sexos. Já no modelo de efeitos fixos, o coeficiente deixa de ser estatisticamente significativo (*p-value* $\approx$ 0,15), o que pode sugerir que parte da disparidade salarial bruta captada nos modelos anuais é explicada por características fixas não observadas dos trabalhadores [77] ou pode dever-se ao facto desta variável não se alterar ao longo do tempo e por isso ter pouca importância em modelos longitudinais.

A nacionalidade (Estrangeiro) tem um impacto negativo consistente, com magnitudes próximas, o que indica um efeito robusto, pouco influenciado por efeitos fixos [78], apesar de não ser significativo nos últimos anos observados.

7.3.2. Escolaridade e qualificação

Os anos de escolaridade apresentam um coeficiente positivo e estatisticamente significativo em ambos os métodos, mas com magnitude muito superior na regressão linear anual. Este

resultado pode ser interpretado de duas formas complementares. Por um lado, a escolaridade é sobretudo importante para explicar diferenças salariais entre indivíduos, o que tende a ser captado com maior intensidade nos modelos *cross-section*. Por outro, parte do prémio salarial atribuído à escolaridade poderá estar correlacionado com características fixas não observadas, já que trabalhadores com maiores capacidades inatas, motivação ou recursos familiares tendem simultaneamente a atingir níveis mais elevados de escolaridade e a auferir salários mais altos, o que pode inflacionar a estimativa nos modelos anuais.

Quando se controlam estes fatores através do modelo de efeitos fixos, o coeficiente da escolaridade diminui, mas mantém-se claro e estatisticamente significativo. Isto indica que a escolaridade é um fator chave nos rendimentos, ainda que o seu efeito estimado seja mais modesto quando se utilizam dados em painel [76].

Já a variável qualificação (*overeducation* e *undereducation* face à profissão exercida) mantém um sinal positivo e significativo nos dois métodos, com menor diferença entre as estimativas quando comparado com outras variáveis referidas acima. Este resultado indica um efeito mais robusto e menos dependente da heterogeneidade não observada, reforçando a importância do *mismatch* educacional como fator explicativo das diferenças salariais [76].

7.3.3. Tipo de contrato

As penalizações associadas aos contratos a termo mantêm-se consistentes em ambos os métodos, com coeficientes negativos e estatisticamente significativos face à categoria de referência, o contrato sem termo. Este resultado sugere um efeito robusto da estabilidade contratual sobre o rendimento [77].

7.3.4. Produtividade

A variável produtividade mantém um efeito positivo e estatisticamente significativo em ambos os métodos, mas com magnitude mais baixa no modelo de efeitos fixos (0,0409 versus 0,0823). Esta diferença sugere que parte do impacto captado nos modelos anuais pode refletir características fixas dos próprios trabalhadores, como capacidades individuais ou preferências, que estão associadas simultaneamente à escolha de empresas mais produtivas e a salários mais elevados.

As empresas mais produtivas tendem a pagar salários mais altos, mas parte desta relação decorre também do facto de atraírem trabalhadores mais qualificados e motivados. Assim, a

produtividade reflete tanto fatores estruturais da empresa, como tecnologia ou gestão, como as características dos trabalhadores que nela se concentram. Quando estas características são controladas no modelo em painel, o coeficiente da produtividade reduz-se, embora permaneça relevante, confirmando a sua importância como determinante salarial [77].

7.3.5. Dimensão da empresa

Na variável de dimensão da empresa, as empresas maiores associam-se a salários mais elevados em ambos os métodos, sendo os coeficientes positivos e estatisticamente significativos. No entanto, o efeito estimado é cerca de metade no modelo de efeitos fixos, o que sugere que parte do prémio salarial das grandes empresas é explicado por características fixas dos trabalhadores que nelas se inserem [77].

7.3.6. Setores de atividade (CAE)

Observam-se várias inversões de sinal, revelando que, quando se controlam efeitos fixos, a posição relativa de certos setores no *ranking* salarial se altera significativamente [77].

7.3.7. Regiões (NUTS II)

Os prémios salariais regionais tendem a ser mais reduzidos nos dados em painel, com uma inversão (Alentejo), sugerindo que diferenças regionais captadas na regressão linear anual podem refletir características fixas dos trabalhadores [79].

7.4. Considerações finais

A comparação entre as estimativas obtidas pelos modelos *cross-section* anuais e pelos modelos de dados em painel com efeitos fixos demonstra que o controlo da heterogeneidade não observada conduz a estimativas mais prudentes e, potencialmente, mais fiéis à realidade [18, 20]. Este controlo permite eliminar relações que parecem existir, mas que são explicadas por fatores escondidos (como aptidões, motivação, etc.) e que não estão explicitamente incluídas nas variáveis observadas [67].

As principais diferenças entre métodos surgem em variáveis como a localização e setor da empresa, que, no modelo de efeitos fixos, apresentam magnitudes mais baixas ou sinais invertidos. Isto indica que, nos modelos *cross-section*, parte dos efeitos captados não resulta apenas da relação direta com o salário, mas também de fatores permanentes que não são

medidos, como práticas internas de remuneração, estabilidade das empresas ou características próprias dos trabalhadores [30, 60].

Este padrão confirma o que é apontado na literatura: a *Human Capital Theory* e a *Assignment Theory* reconhecem que o impacto da escolaridade, da experiência e da produtividade depende não só das características individuais, mas também do contexto em que o trabalhador está inserido [12, 20]. Ao eliminar o efeito dessas características invariantes, o modelo de efeitos fixos aproxima-se mais de uma relação de causa e efeito entre variáveis como escolaridade, qualificação, produtividade e estrutura empresarial sobre o rendimento [80].

Assim, a utilização de metodologias que considerem efeitos fixos revela-se fundamental para estudos que pretendam identificar relações causais mais robustas entre características dos trabalhadores, das empresas e o rendimento. No contexto do presente estudo, esta abordagem permitiu clarificar que, embora a qualificação e a produtividade mantenham aproximadamente a sua magnitude e um efeito positivo consistente, variáveis como antiguidade ou dimensão da empresa podem ter sido sobrestimadas quando analisadas apenas com métodos de *cross-section* [18, 60].

8. Conclusão

O presente estudo teve como objetivo analisar a relação entre a qualificação absoluta (medida em termos de escolaridade) e relativa (que corresponde à forma como o nível de escolaridade de um trabalhador se encontra acima, abaixo ou muito próximo da escolaridade média na profissão) dos trabalhadores e os seus rendimentos reais em Portugal. Para isso, utilizou-se uma base de dados longitudinal abrangendo o período de 2010 a 2023, que cobre todos os trabalhadores por conta de outrem do setor privado em Portugal. Duas metodologias complementares foram aplicadas: regressões *cross-section* anuais e regressão linear com dados em painel utilizando efeitos fixos. Nas regressões *cross-section*, a análise centra-se em explicar diferenças salariais entre indivíduos, enquanto os modelos de dados em painel com efeitos fixos (MEF) permitem não só captar variações ao longo do tempo, mas também controlar a heterogeneidade não observada.

Os resultados confirmam que os anos de escolaridade têm impacto positivo e significativo no rendimento, embora o seu efeito seja substancialmente reduzido quando controlamos por características fixas dos trabalhadores. A análise da qualificação relativa mostra que trabalhadores *overeducated* recebem em média, um prémio salarial face aos adequadamente qualificados, mesmo após o controlo do nível de escolaridade e quando se considera o rendimento médio por profissão no modelo. Ao incluir a variável rendimento médio por profissão, o estudo foca na análise das diferenças salariais dentro das profissões, sugerindo que em igual contexto profissional, os *overeducated* conseguem auferir ligeiros prémios. No entanto, quando essa variável é excluída, a comparação passa a refletir diferenças entre profissões, e o coeficiente da *overeducation* torna-se negativo, em linha com a literatura internacional. No caso dos trabalhadores *undereducated*, a análise revela que estes trabalhadores recebem, em média, menos do que os trabalhadores adequadamente qualificados, mas mais do que aqueles com o mesmo nível de educação que estejam em profissões onde esse nível é considerado adequado. Assim, um trabalhador com escolaridade inferior à média do seu grupo profissional pode auferir salários relativamente elevados, por se encontrar numa profissão mais exigente e mais bem remunerada (por exemplo, profissões técnicas ou especializadas em que a média de escolaridade é superior). Este resultado é consistente com a literatura, que aponta para situações em que os *undereducated* são

valorizados pelo contexto da profissão em que estão inseridos, e não unicamente pela sua escolaridade individual.

Variáveis como a produtividade, a dimensão da empresa e fatores demográficos exercem também influência significativa, assim como aspetos contratuais e regionais. A comparação entre métodos evidenciou que o controlo por efeitos fixos reduz a magnitude de vários coeficientes e, nalguns casos, inverte o seu sinal, eliminando relações influenciadas por fatores não observados presentes nas estimativas anuais.

Considera-se que os objetivos do estudo foram atingidos. Foi possível construir e explorar uma base de dados longitudinal de grande dimensão, abrangendo cerca de um milhão de trabalhadores do setor privado entre 2010 e 2023, com informação muito rica em termos de fatores individuais, organizacionais e contextuais, permitindo isolar os efeitos da qualificação absoluta e relativa sobre os salários. A aplicação complementar de regressões *cross-section* anuais e de modelos com dados em painel com efeitos fixos possibilitou comparar resultados e avaliar a robustez das estimativas, enquanto a análise temporal captou a evolução das relações entre variáveis ao longo do período. Desta forma, o estudo oferece uma visão abrangente e comparativa da influência da qualificação, escolaridade e perfil profissional sobre os rendimentos.

Apesar da robustez da análise, importa salientar que a base de dados inclui apenas trabalhadores com registos contínuos durante 13 ou mais anos, excluindo os que ingressaram no mercado de trabalho após 2011, o que limita a generalização dos resultados a perfis mais móveis ou recentes. Além disso, variáveis relevantes como competências específicas ou desempenho individual não estavam disponíveis, podendo a sua inclusão aprofundar as conclusões.

Estas limitações sugerem linhas futuras de investigação. Estudos posteriores poderão beneficiar da integração de informações adicionais sobre competências e desempenho, bem como da inclusão de trabalhadores com carreiras mais recentes ou descontínuas, de modo a captar fenómenos associados à transferência emprego-desemprego e às novas formas de emprego.

Este estudo contribui para um melhor entendimento da relação entre qualificação e rendimento, oferecendo evidência empírica útil para apoiar políticas de emprego e educação orientadas para a valorização do capital humano.

Bibliografia

1. Becker, G. S. (1964). *Human capital: a theoretical and empirical analysis, with special reference to education*. University of Chicago Press.

2. Mincer, J. (1974). *Schooling, Experience and Earnings*. *Clumbia Universi*. New York: National Bureau of Economic Research (NBER).

3. Lemieux, T. (2003). *The Mincer Equation Thirty Years after Schooling Experience, and Earnings*. Center for Labor Economics, University of British Columbia and UC Berkeley.

4. Sousa, S., Portela, M., & Sá, C. (2015). *Characterization of returns to education in Portugal: 1986-2009*. 4th Linked Employer-Employee data (LEED) Workshop, Instituto Superior Técnico, Lisbon.

5. Campos, M., & Reis, H. (2017). Uma reavaliação do retorno do investimento em educação na economia portuguesa. *Revista de Estudos Económicos do Banco de Portugal, vol. 3, n.º 2, pp. 1-30*.

6. Fernandes, A. (2012). *O que determina as desigualdades salariais em Portugal?* Dissertação de Mestrado, Universidade Técnica de Lisboa.

7. Cabral, S., & Duarte, C. (2012). O Diferencial de Salários dos Imigrantes no Mercado de Trabalho Português. *Boletim Económico do Banco de Portugal, pp. 85-103*.

8. Friedberg, R. M. (2000). You can't take it with you? Immigrant assimilation and the portability of human capital. *Journal of Labor Economics*, vol. 18, n. 2, pp. 221–251. https://doi.org/10.1086/209957

9. Eurofound, & Vacas-Soriano, C. (2015). *Recent developments in temporary employment: Employment growth, wages and transitions*. *Eurofound*. Publications Office, Luxembourg. https://doi.org/https://doi.org/10.2806/014550

10. Magda, I., Rycx, F., Tojerow, I., & Valsamis, D. (2011). Wage differentials across sectors in Europe. *Economics of Transition*, vol. 19, n. 4, pp. 749–769. https://doi.org/10.1111/j.1468-0351.2011.00417.x

11. Verhaest, D., & Omey, E. (2010). The determinants of overeducation: different measures, different outcomes? *International Journal of Manpower*, vol. 31, n. 6, pp. 608–625. https://doi.org/10.1108/01437721011073337

12. Singer, N. M. (1976). Generating Inequality Mechanisms of Distribution in the U.S. Economy by Lester C. Thurow . *Challenge*, vol. 19, n. 4, pp. 58–59. https://doi.org/10.1080/05775132.1976.11470249

13. Sattinger, M. *et al.,* (1993). Assignment Models of the Distribution of Earnings. *Journal of Economic Literature*, *31*(2), 831–880. Retrieved from https://www.jstor.org/stable/2728516

14. Sicherman, N., & Galor, O. (1990). A Theory of Career Mobility. *Journal of Political Economy*, vol. 98, n. 1, pp. 169–192. https://doi.org/10.1086/261674

15. Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, vol. 87, n. 3, pp. 355–374. https://doi.org/10.2307/1882010

16. Mahy, B., Rycx, F., & Vermeylen, G. (2015). Educational Mismatch and Firm Productivity: Do Skills, Technology and Uncertainty Matter? *De Economist (Netherlands)*, vol. 163, n. 2, pp. 233–262. https://doi.org/10.1007/s10645-015-9251-2

17. Instituto Nacional de Estatística. (2011). *Classificação Portuguesa das Profissões: 210*. Lisboa: INE. Retrieved from https://www.ine.pt/xurl/pub/107961853

18. Kiker, B. F., Santos, M. C., & de Oliveira, M. M. (1997). Overeducation and undereducation: Evidence for Portugal. *Economics of Education Review*, vol. 16, n. 2, pp. 111–125. https://doi.org/10.1016/S0272-7757(96)00040-4

19. Verdugo, R. R., & Verdugo, N. T. (1989). The Impact of Surplus Schooling on Earnings: Some Additional Findings. *The Journal of Human Resources*, vol. 24, n. 4, pp. 629–643. https://doi.org/10.2307/145998

20. Hartog, J. (2000). Over-education and earnings: where are we, where should we go? *Economics of Education Review*, vol. 19, n. 2, pp. 131–147. https://doi.org/10.1016/S0272-7757(99)00050-3

21. Leuven, E., & Oosterbeek, H. (2011). Overeducation and Mismatch in the Labor Market. *Handbook of the Economics of Education*, vol. 4, pp. 283–326. https://doi.org/10.1016/B978-0-444-53444-6.00003-1

22. Duncan, G. J., & Hoffman, S. D. (1981). The incidence and wage effects of overeducation. *Economics of Education Review*, vol. 1, n. 1, pp. 75–86. https://doi.org/10.1016/0272-7757(81)90028-5

23. Cohn, E., Johnson, E., & Ng, Y. C. (2000). The incidence of overschooling and underschooling and its effect on earnings in the United States and Hong Kong. *Research in Labor Economi*cs, vol. 19, pp. 29–61. https://doi.org/10.1016/s0147-9121(00)19003-x

24. Bauer, T. K. (2002). Educational mismatch and wages: a panel analysis. *Economics of Education Review*, vol. 21, n. 3, pp. 221–229. https://doi.org/10.1016/S0272-7757(01)00004-8

25. Kampelmann, S., & Rycx, F. (2012). The impact of educational mismatch on firm productivity: Evidence from linked panel data. *Economics of Education Review*, vol. 31, n. 6, pp. 918–931. https://doi.org/10.1016/j.econedurev.2012.07.003

26. Rocha, A. B., Figueiredo, H., Sá, C., & Portela, M. (2025). Mismatch matters: education and productivity in laggard and frontier firms. *Journal of Productivity Analysis*. https://doi.org/10.1007/s11123-025-00772-4

27. Krzywda-Starzyk, P. (2024). There's more to overeducation than a wage penalty: A systematic review of 2011–2021 literature on a vertical job-education mismatch in Europe. *Kultura-Społeczeństwo-Edukacja*, vol. 26, n. 2. https://doi.org/10.14746/kse.2024.26.2.14

28. Groeneveld, S., & Hartog, J. (2004). Overeducation, wages and promotions within the firm. *Labour Economics*, vol. 11, n. 6, pp. 701–714. https://doi.org/10.1016/j.labeco.2003.11.005

29. Büchel, F., & Mertens, A. (2004). Overeducation, undereducation, and the theory of career mobility. *Applied Economics*, vol. 36, n. 8, pp. 803–816. https://doi.org/10.1080/0003684042000229532

30. Quintini Glenda. (2011). Over-Qualified or Under-Skilled - A Review of Existing Literature, *n.º 121*. https://doi.org/10.1787/5kg58j9d7b6d-en

31. Pimenta, A. C., & Pereira, M. C. (2019). Aggregate educational mismatches in the Portuguese labour market. *Banco de Portugal, Economic Studies*, vol. 5, pp. 41–66. Retrieved from https://www.bportugal.pt/sites/default/files/anexos/papers/re201903_e.pdf

32. Banco de Portugal. (2019). *O Crescimento Economico Português: Uma Visão sobre Questões Estruturais Bloqueios e Reformas*. Departamento de Estudos Económicos, Lisboa.

33. McGuinness, S., Bergin, A., & Whelan, A. (2019). Youth Overeducation in Europe: Is there scope for a common policy approach? *Youth Labor in Transition: Inequalities, Mobility, and Policies in Europe*, 530–559. https://doi.org/10.1093/oso/9780190864798.003.0018

34. RStudio Team. (2023). RStudio: Integrated Development for R. Boston, MA: RStudio, PBC. Retrieved from https://posit.co/

35. R Core Team. (2024). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

36. Wickham, H. *et al.,* (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, vol. 4, n. 43, pp. 1686. https://doi.org/10.21105/joss.01686

37. Barrett, T. *et al.* , (2025). data.table: Extension of `data.frame`. *CRAN: Contributed Packages*. Journal of Statistical Software. https://doi.org/10.32614/CRAN.package.data.table

38. Schauberger, P., & Walker, A. (2025). openxlsx: Read, Write and Edit xlsx Files. *CRAN: Contributed Packages*. https://doi.org/10.32614/CRAN.package.openxlsx

39. Hothorn, T., Zeileis, A., Farebrother, R. W., & Cummins, C. (2002). Diagnostic Checking in Regression Relationships. *R News*, vol. 2, n. 3, pp. 7–10. https://doi.org/10.32614/CRAN.package.lmtest

40. John Fox e Sanford Weisberg. (2019). An R Companion to Applied Regression. Sage Publications.

41. Zeileis, A., Köll, S., & Graham, N. (2020). Various Versatile Variances: An Object-Oriented Implementation of Clustered Covariances in *R*. *Journal of Statistical Software*, vol. 95, n. 1. https://doi.org/10.18637/jss.v095.i01

42. Gross, J., & Ligges, U. (2015). nortest: Tests for Normality. https://doi.org/10.32614/CRAN.package.nortest

43. Croissant, Y., & Millo, G. (2008). Panel Data Econometrics in R: The plm Package. *Journal of Statistical Software*, vol. 27, n. 2. https://doi.org/10.18637/jss.v027.i02

44. Martins, P. S. (2009). Rent sharing before and after the wage bill. *Applied Economics*, vol. 41, n. 17, pp. 2133–2151. https://doi.org/10.1080/00036840701736164

45. Card, D. (2022). Who Set Your Wage? *American Economic Review*, vol. 112, n. 4, pp. 1075–1090. https://doi.org/10.1257/aer.112.4.1075

46. Abrantes, P. (2022). Education and social classes in Portugal: continuities and changes in the 21st century. *Sociologia, Problemas e Praticas*, n. 99, pp. 9–27. https://doi.org/10.7458/SPP20229924309

47. Troske, K. R. (1999). Evidence on the Employer Size-Wage Premium from Worker-Establishment Matched Data. *Review of Economics and Statistics*, vol. 81, n. 1, pp. 15–26. https://doi.org/10.1162/003465399557950

48. Cruz, I. *et al.,* (2019). Labour market segmentation: Piloting new empirical and policy analyses. *Eurofound*. https://doi.org/10.2806/30796

49. PORTUGAL.GOV.PT. (2023). OCDE: mais jovens a concluir o ensino secundário e superior, mais alunos no ensino profissional. *PORTUGAL.GOV.PT*. Acedido 26 de Julho, 2025. Disponível em:

https://www.portugal.gov.pt/pt/gc23/comunicacao/noticia?i=ocde-mais-jovens-a-concluir-o-ensino-secundario-e-superior-mais-alunos-no-ensino-profissional

50. Banco de Portugal. (2024). Como éramos e como mudámos - Educação em Portugal: esta é a madrugada que eu esperava. *Banco de Portugal*. Acedido 26 de Julho, 2025. Disponível em: https://www.bportugal.pt/page/como-eramos-e-como-mudamos-educacao-em-portugal-esta-e-madrugada-que-eu-esperava

51. Redmond, P., & Brosnan, L. (2025). Educational Mismatch in Europe: Incidence, Determinants and the Impact of Increased Remote Working. *Social Science Research Network (SSRN)*. https://doi.org/10.2139/ssrn.5133819

52. Redmond, P., & Brosnan, L. (2024). Skills Mismatch in Europe. *Enabling Data Analytics for Actions Tackling Skills Shortages & Mismatch (TRAILS)*. Acedido 6 de Setembro, 2025. Disponível em: https://www.trails-project.eu/skills-mismatch-in-europe

53. Fundação José Neves. (2023). Quase 1 em cada 4 jovens com ensino superior trabalha em profissões que não exigem o seu nível de escolaridade. *Fundação José Neves*. Acedido 26 de Julho, 2025. Disponível em: https://www.joseneves.org/artigo/quase-1-em-cada-4-jovens-com-ensino-superior-trabalha-em-profissoes-que-nao-exigem-o-seu-nivel-de-escolaridade

54. INE. (2024). Place of residence vs Population density. *Instituto Nacional de Estatística*. Acedido 14 de Agosto, 2025. Disponível em: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_indicadores&contecto=pi&indOcorrCod=0013189&selTab=tab0

55. Monastiriotis, V. *et al.,* (2013). Austerity measures in crisis countries — results and impact on mid-term development. *Intereconomics*, vol. 48, n. 1, pp. 4–32. https://doi.org/10.1007/s10272-013-0441-3

56. Passinhas, J., & Araújo, T. (2021). Gender-based occupational segregation: a bit string approach. *REM – Research in Economics and Mathematics, arXiv preprint*. Disponível em: https://arxiv.org/pdf/2108.10343v1

57. Banco de Portugal. (2024). A economia portuguesa em 2024–27. *Banco de Portugal*. Acedido 26 de Julho, 2025. Disponível em: https://www.bportugal.pt/publicacao/boletim-economico-dezembro-2024

58. Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*.

59. Baltagi, B. H. (2021). *Econometric Analysis of Panel Data* (6th ed.). Springer. https://doi.org/10.1007/978-3-030-53953-5

60. Groot, W., & Henriëtte, M. van den B. (2000). Overeducation in the labor market: a meta-analysis. *Economics of Education Review*, vol. 19, n. 2, pp. 149–158. https://doi.org/10.1016/S0272-7757(99)00057-6

61. Green, F., & McIntosh, S. (2007). Is there a genuine under-utilization of skills amongst the over-qualified? *Applied Economics*, vol. 39, n. 4, pp. 427–439. https://doi.org/10.1080/00036840500427700

62. Kunze, A. (2018). The Gender Wage Gap in Developed Countries. In S. L. Averett, L. M. Argys, & S. D. Hoffman (Eds.), *The Oxford Handbook of Women and the Economy*, Oxford University Press pp. 368–394. https://doi.org/10.1093/oxfordhb/9780190628963.013.11

63. Blau, F. D., & Kahn, L. M. (2017). The gender wage gap: Extent, trends, & explanations. *Journal of Economic Literature*, vol. 55, n. 3, pp. 789–865. https://doi.org/10.1257/jel.20160995

64. Longhi, S., Nijkamp, P., & Poot, J. (2005). A Meta-Analytic Assessment of the Effect of Immigration on Wages. *Journal of Economic Surveys*, vol. 19, n. 3, pp. 451–477. https://doi.org/10.1111/j.0950-0804.2005.00255.x

65. Diogo, F. (2012). *Precariedade no emprego em Portugal e desigualdades sociais: alguns contributos*. Observatório das Desigualdades. Disponível em: http://hdl.handle.net/10400.3/2722

66. Oi, W. Y., & Idson, T. L. (1999). Chapter 33 Firm size and wages. *Handbook of Labor Economics*, *3*, 2165–2214. https://doi.org/10.1016/S1573-4463(99)30019-5

67. McGuinness, S., Pouliakas, K., & Redmond, P. (2018). Skilss mismatch: concepts, measurement and policy approaches. *Journal of Economic Surveys*, vol. 32, n. 4, pp. 985–1015. https://doi.org/10.1111/joes.12254

68. Campos, M. M., & Centeno, M. (2011). *Diferenças salariais entre os setores público e privado no período que antecedeu a adoção do euro: uma aplicação baseada em dados longitudinais*. Economic Bulletin and Financial Stability Report Articles and Banco de Portugal Economic Studies. Disponível em: https://www.bportugal.pt/sites/default/files/anexos/papers/ab201115_p.pdf

69. Pereira, J., & Galego, A. (2011). Regional wage differentials in Portugal: Static and dynamic approaches. *Papers in Regional Science*, vol. 90, n. 3, pp. 529–548. https://doi.org/10.1111/j.1435-5957.2010.00328.x

70. Gujarati, D. N., & Porter, D. C. (2008). Causality in economics: The Granger causality test. In *Basic Econometrics* (5th ed.). McGraw-Hill Education.

71. Gabinete de Estratégia e Planeamento. (2023). *Barometro das diferenças remuneratórias entre mulheres e homens 2023*. Retrieved from http://www.gep.mtsss.gov.pt/trabalho

72. OECD. (2023). Reporting Gender Pay Gaps in OECD Countries: Guidance for Pay Transparency Implementation, Monitoring and Reform. In *Gender Equality at Work*. OECD Publishing, Paris. https://doi.org/10.1787/ea13aa68-en

73. Arellano-Bover, J., Bianchi, N., Lattanzio, S., & Paradisi, M. (2024). *One Cohort at a Time: A New Perspective on the Declining Gender Pay Gap*. National Bureau of Economic Research, Cambridge, MA. https://doi.org/10.3386/w32612

74. CIG. (2024). *Igualdade de género em Portugal indicadores chave 2024 juventude e modernização*. Comissão para a cidadania e a igualdade de género. Disponível em: https://www.cig.gov.pt/wp-content/uploads/2024/12/IC2024_Paginados-A4_v6.pdf

75. Lazear, E. P., & Oyer, P. (2004). Internal and external labor markets: a personnel economics approach. *Labour Economics*, vol. 11, n. 5, pp. 527–554. https://doi.org/10.1016/j.labeco.2004.01.001

76. Card, D. (1999). The Causal Effect of Education on Earnings. In *Handbook of Labor Economics*, Vol. 3, pp. 1801–1863. https://doi.org/10.1016/S1573-4463(99)03011-4

77. Abowd, J. M., Kramarz, F., & Margolis, D. N. (1999). High Wage Workers and High Wage Firms. *Econometrica*, vol. 67, n. 2, pp. 251–333. https://doi.org/10.1111/1468-0262.00020

78. Dustmann, C., & Görlach, J. S. (2016). The economics of temporary migrations. *Journal of Economic Literature*. American Economic Association. https://doi.org/10.1257/jel.54.1.98

79. Combes, P. P., Duranton, G., & Gobillon, L. (2008). Spatial wage disparities: Sorting matters! *Journal of Urban Economics*, vol. 63, n. 2, pp. 723–742. https://doi.org/10.1016/j.jue.2007.04.004

80. Pereira, P. T., & Martins, P. S. (2002). *Education and Earnings in Portugal*. Bank of Portugal Conference Proceedings.

Anexos

Tabela A.1 - Estatística antes e após remoção dos NA

Antes e após remoção dos NA			
Variáveis		**Antes**	**Após**
	Nº Observações	41.433.063	36.624.042
Sexo	Homem	53,56%	53,59%
	Mulher	46,44%	46,40%
	Idade (média)	40.84	40.28
Nacionalidade	Português	93,79%	93,82%
	Estrangeiro	6,21%	6,18%
	Anos Escolaridade (média)	10,1	10,1
	Antiguidade (média)	7,72	7,50
	Remuneração (média, em euros)	997,65	1033,25
	Produtividade (média, em euros)	125.035	132.711
Tipo Contrato	Sem termo	68,96%	69,18%
	Termo certo	23,06%	23,26%
	Termo incerto	6,85%	6,92%
Dimensão Empresa	Micro	23,38%	19,83%
	Pequena	24,74%	25,14%
	Media	21,37%	22,29%
	Grande	30,51%	32,73%

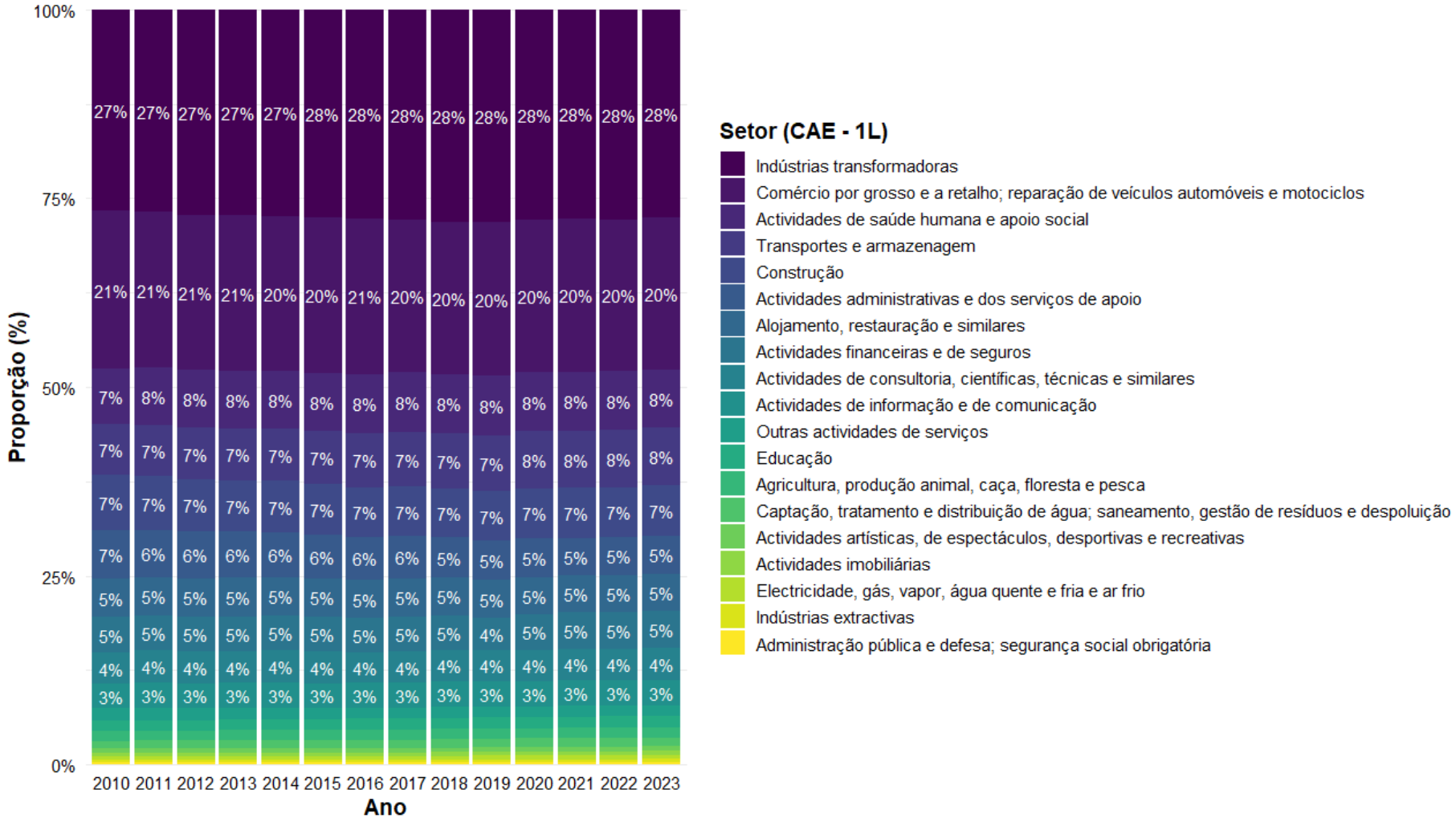

Figura A.1 - Setor CAE - 1L

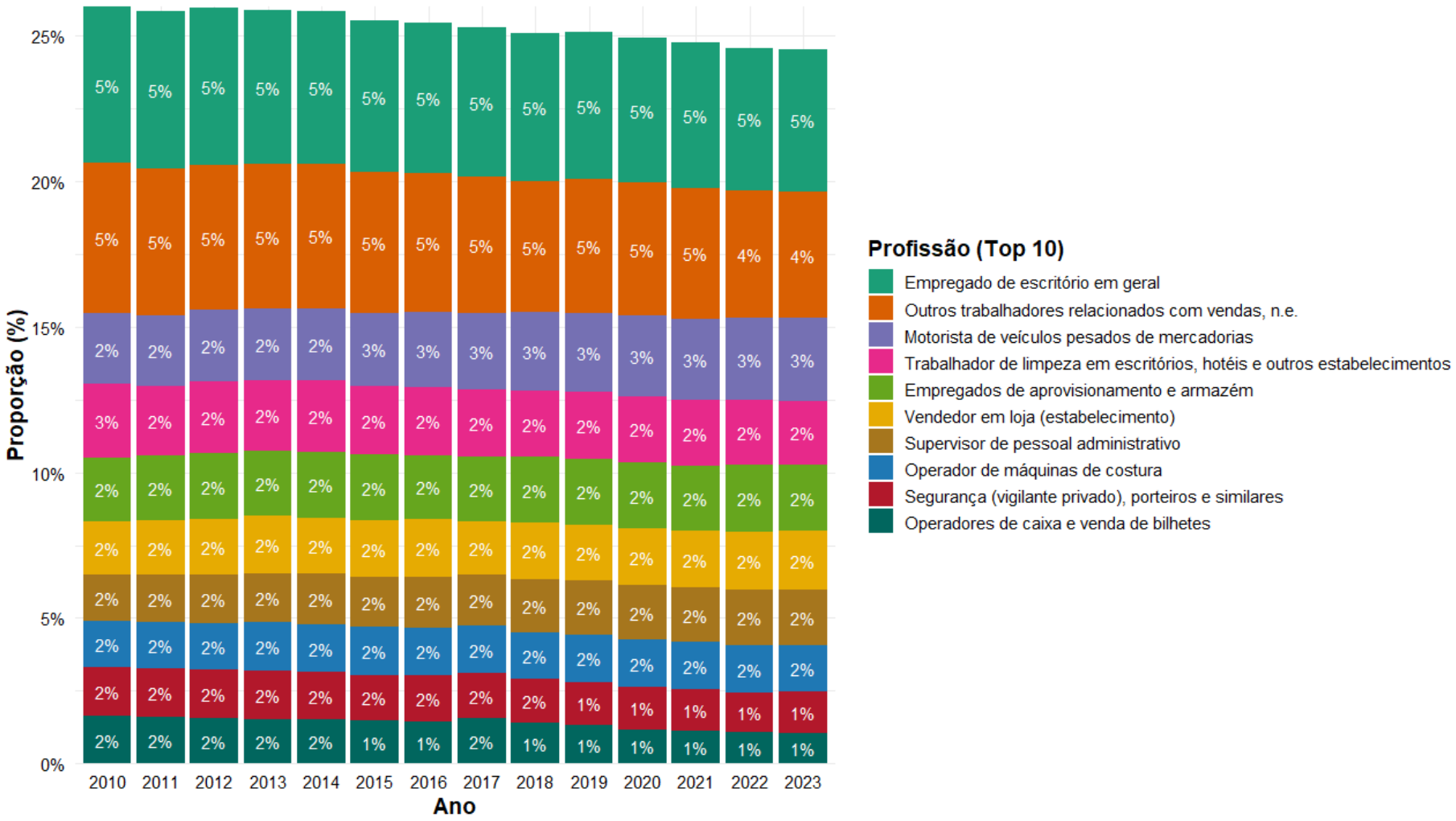

Figura A.2 - Profissões mais representadas (Top 10)

Tabela A.2 - Coeficientes estimados do MEF e cross-section (média)

Variável	*Cross-section*		Dados em Painel	
	Estimativas (média)	Significância	Estimativa	Significância
antig	0,0059	***	-0,0011	***
anos_escolaridade	0,0356	***	0,0046	***
idade_numerica	0,0073	***	0,0181	***
qualificacaoUnder	0,0163	***	0,0039	***
qualificacaoOver	0,0285	***	0,0237	***
log_produtividade	0,0823	***	0,0410	***
media_LogRganho_profissao	0,6280	***	0,1760	***
dim_empresaPequena empresa	0,1394	***	0,0643	***
dim_empresaMédia empresa	0,1931	***	0,0982	***
dim_empresaGrande empresa	0,1749	***	0,1043	***
sexoMulher	-0,1387	***	-0,0036	
NacionalidadeEstrangeiro	-0,0161		-0,0132	***
nut2_empAlgarve	0,0723	***	0,0190	***
nut2_empCentro	0,0039		0,0003	
nut2_empLisboa	0,0940	***	0,0100	***
nut2_empAlentejo	0,0489	***	-0,0047	
nut2_empAçores	0,0463	***	0,0121	.
nut2_empMadeira	0,0983	***	0,0649	***
nut2_empEstrangeiro	0,3162		0,0936	
caem1lIndústrias extractivas	0,0308	.	0,0867	***
caem1lIndústrias transformadoras	-0,0649	***	0,0268	***
caem1lElectricidade, gás, vapor, água quente e fria e ar frio	0,1545	***	0,0515	***
caem1lCaptação, tratamento e distribuição de água; saneamento, gestão de resíduos e despoluição	-0,0869	**	0,0237	***

caem1lConstrução	-0,0855	***	-0,0364	***
caem1lComércio por grosso e a retalho; reparação de veículos automóveis e motociclos	-0,1189	***	-0,0123	**
caem1lTransportes e armazenagem	-0,0199	.	0,0512	***
caem1lAlojamento, restauração e similares	-0,0145		-0,0253	***
caem1lActividades de informação e de comunicação	0,0001		0,0329	***
caem1lActividades financeiras e de seguros	-0,0293	.	0,1187	***
caem1lActividades imobiliárias	-0,0191	.	0,0162	*
caem1lActividades de consultoria, científicas, técnicas e similares	-0,0218		0,0248	***
caem1lActividades administrativas e dos serviços de apoio	-0,1282	***	-0,0646	***
caem1lAdministração pública e defesa; segurança social obrigatória	-0,0089		0,0927	***
caem1lEducação	-0,0163	*	0,0456	***
caem1lActividades de saúde humana e apoio social	-0,0525	*	0,0568	***
caem1lActividades artísticas, de espectáculos, desportivas e recreativas	0,0077		0,0493	***
caem1lOutras actividades de serviços	0,0182		0,0506	***
tipo_contr1Contrato de trabalho com termo certo	-0,0666	***	-0,0660	***
tipo_contr1Contrato de trabalho com termo incerto	-0,0837	***	-0,0738	***
tipo_contr1Outra situação	-0,0725	**	-0,0731	***

Legenda dos níveis de significância estatística: **** → muito significativo; ** → significativo a 1%; * → significativo a 5%; . → marginalmente significativo a 10%; espaço em branco → não significativo.*

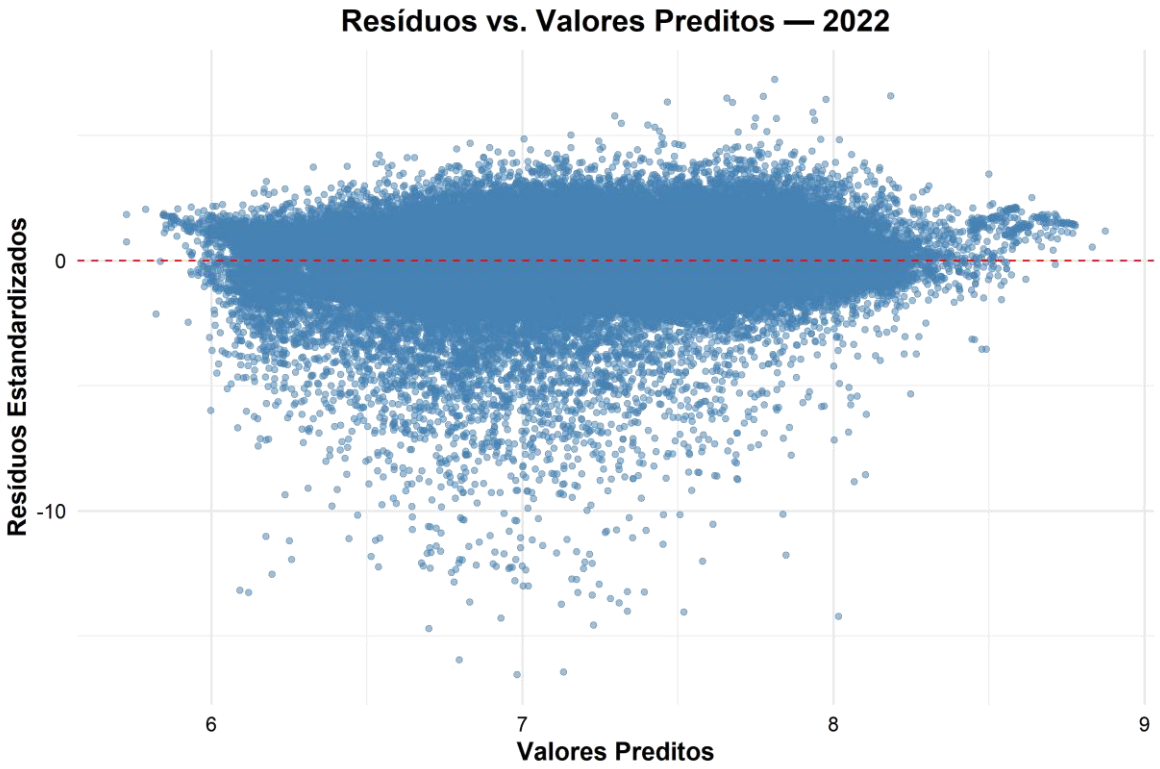

Figura A.3 - Análise dos resíduos da regressão linear do ano 2022

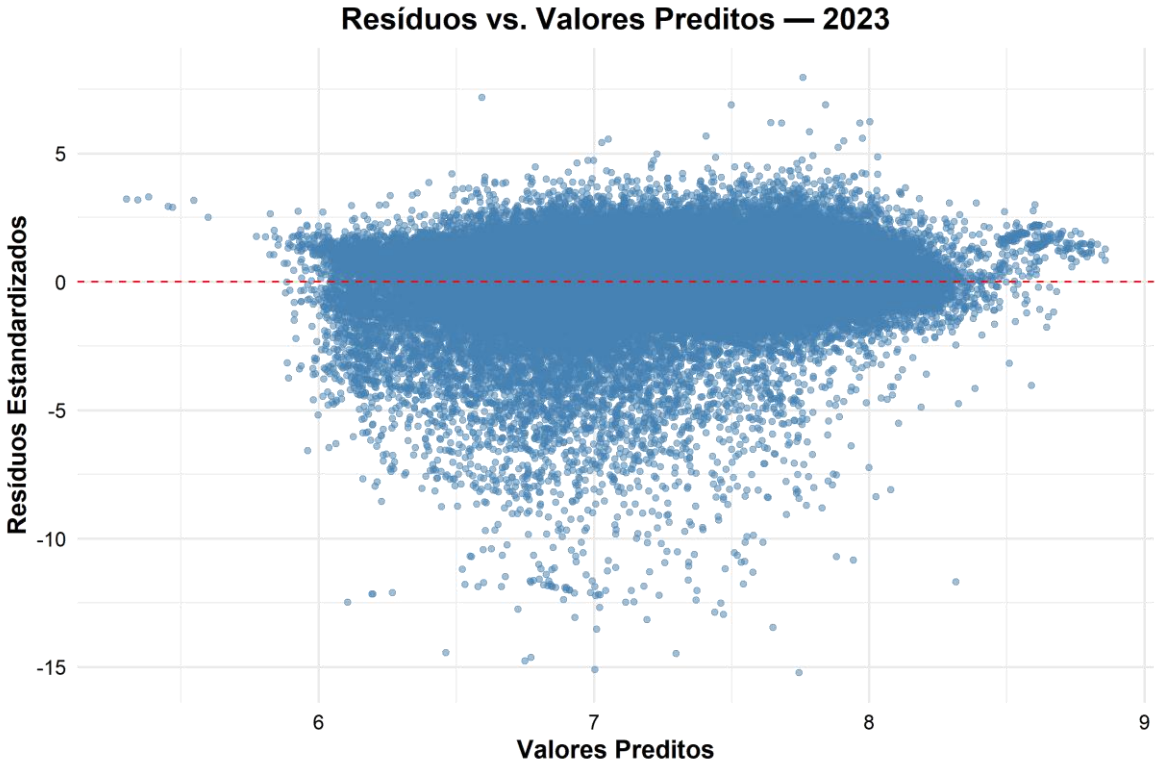

Figura A.4 - Análise dos resíduos da regressão linear do ano 2023

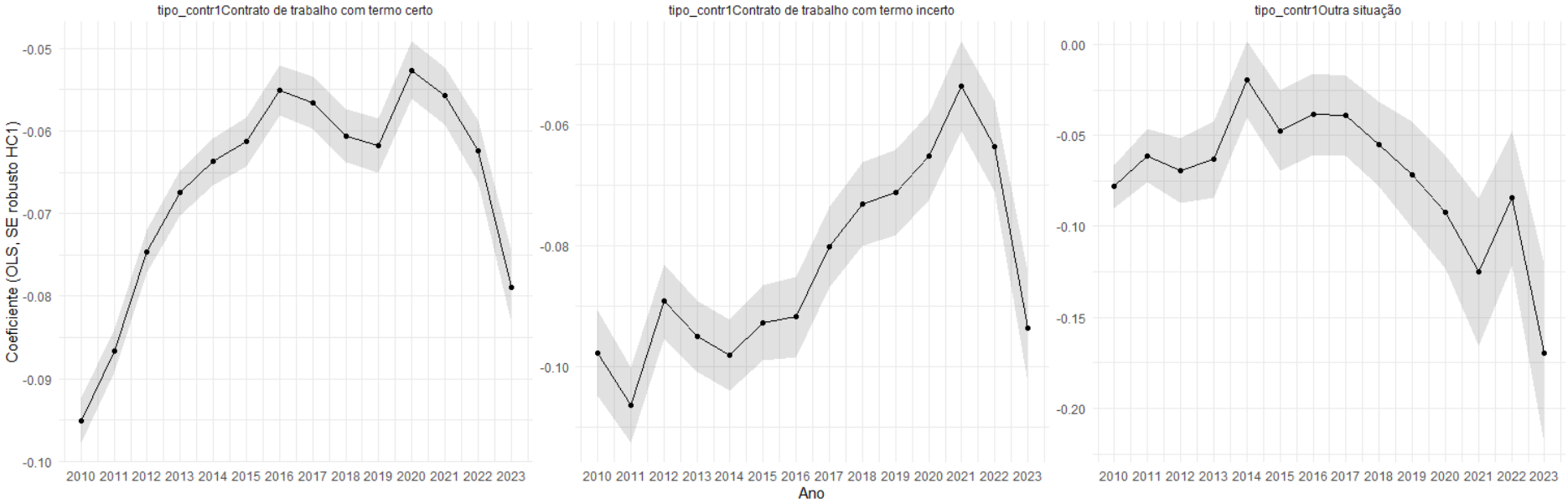

Figura A.5 - Evolução do coeficiente ao longo dos anos da variável *tipo_contr1*

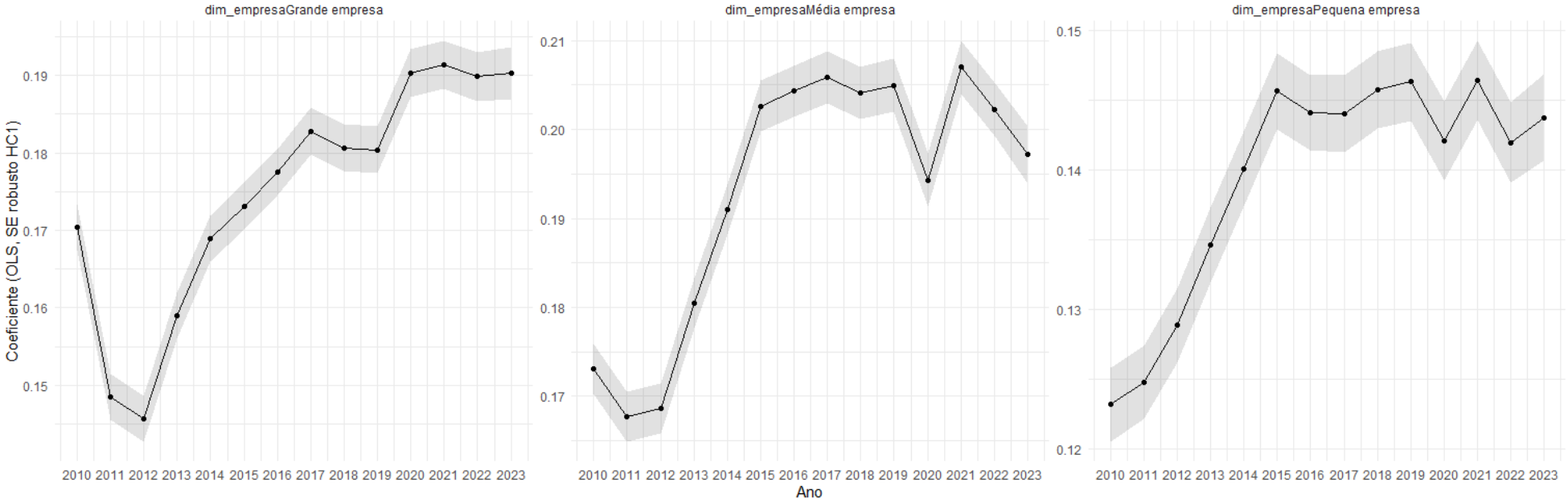

Figura A.6 - Evolução do coeficiente ao longo dos anos da variável *dim_empresa*

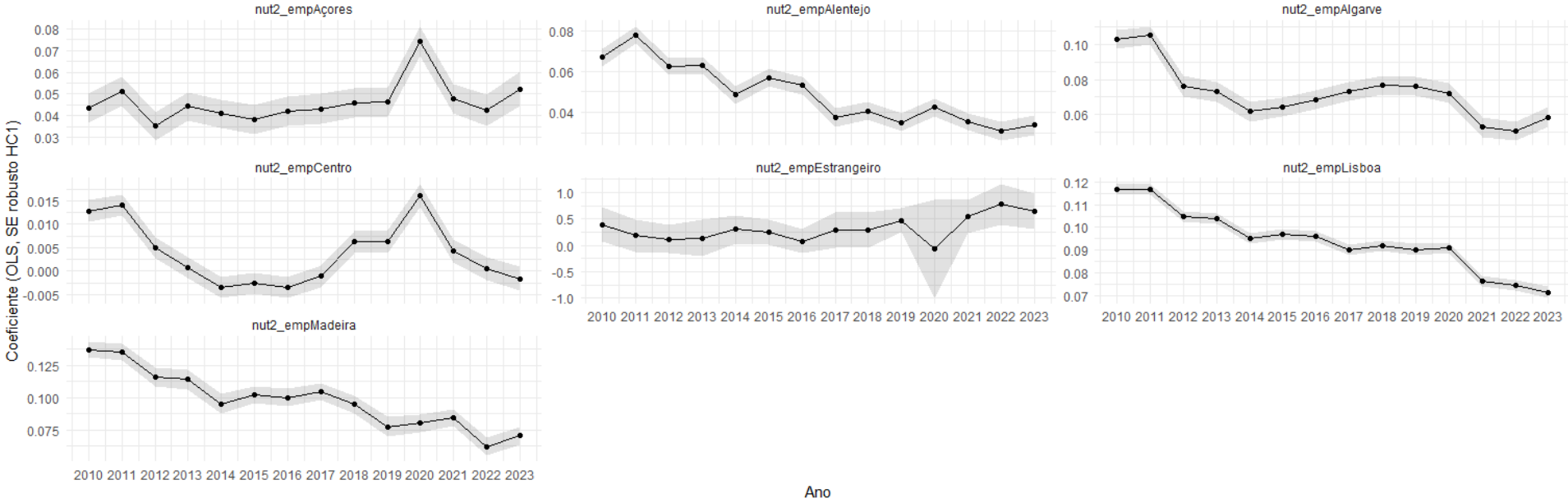

Figura A.7 - Evolução do coeficiente ao longo dos anos da variável *nut2_emp*

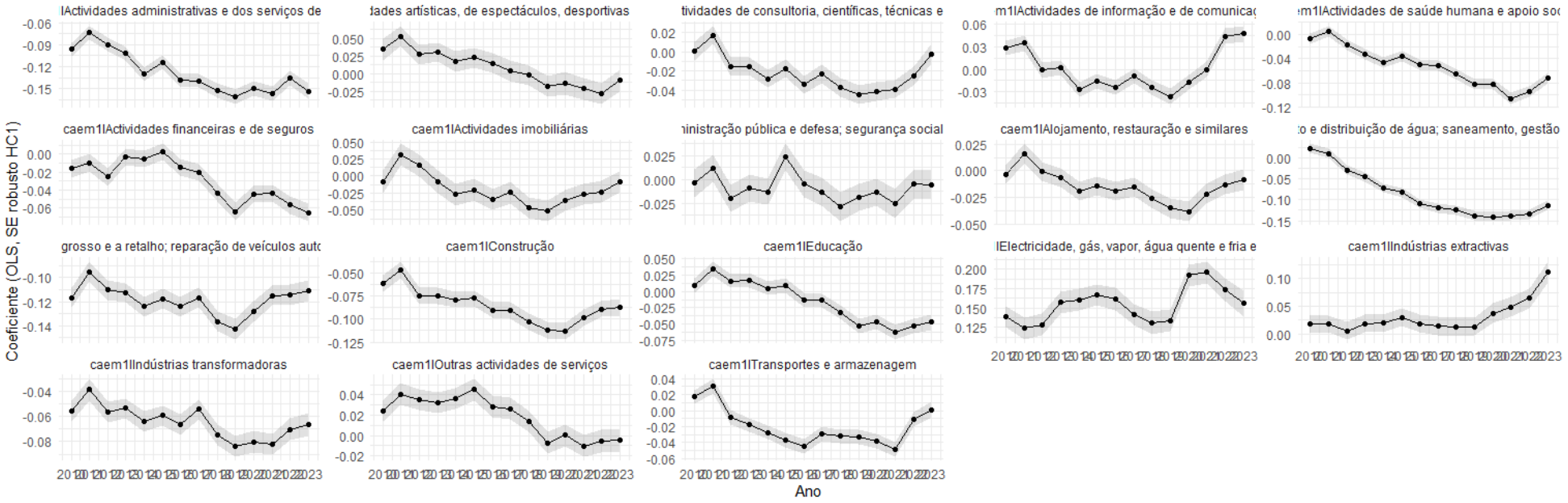

Figura A.8 - Evolução do coeficiente ao longo dos anos da variável *caem1l*