NOTA DE ANÁLISE N.º 11

PLANAPP
Centro de Planeamento e de Avaliação de Políticas Públicas

Evolução dos salários por nível de ensino em Portugal

Dinâmicas recentes

Apêndice Técnico

Novembro 2024

FICHA TÉCNICA

Título

Evolução dos salários por nível de ensino em Portugal: dinâmicas recentes – Apêndice Técnico

Autoria

PLANAPP – Centro de Planeamento e de Avaliação de Políticas Públicas

Data

Novembro 2024

PLANAPP – Centro de Planeamento e de Avaliação de Políticas Públicas
Campus XXI, Av. João XXI, n. 63
1000-300 Lisboa
planapp@planapp.gov.pt
www.planapp.gov.pt

Introdução

Este apêndice técnico detalha as abordagens analíticas utilizadas na Nota de Análise “Evolução dos salários por nível de ensino em Portugal: dinâmicas recentes”. A análise é dividida em duas partes principais: a estimação de uma equação Minceriana (Mincer, 1958) usando uma regressão linear e a exploração da heterogeneidade do efeito do tratamento utilizando um modelo de efeitos de tratamento médios condicionais, estimado com recurso a florestas causais (Athey *et al.*, 2019), uma abordagem que cruza aprendizagem automática com inferência causal.

As duas metodologias empregues complementam-se no sentido de, a par de se identificarem dinâmicas da tendência geral dos prémios salariais da educação (através do conceito estatístico de efeitos de tratamento médios), se explorar a natureza heterogénea destes prémios para períodos concretos, descrevendo-se a sua distribuição ao longo de diversas dimensões de análise.

1. Descrição dos dados

As tabelas de dados “Quadros de Pessoal” (GEP - MTSS) resultam de uma operação estatística de carácter censitário resultante de procedimentos administrativos. A obrigatoriedade de entrega dos “Quadros de Pessoal” aplica-se a todas as entidades empregadoras de trabalhadores, com exceção das administrações central, regional e local, dos institutos públicos (que não no caso dos trabalhadores que estejam em regime de contrato individual de trabalho) e das entidades empregadoras de trabalhadores do serviço doméstico. Os dados dos “Quadros de Pessoal” têm como referência o mês de outubro de cada ano e abrangem todo o território nacional (continente e regiões autónomas). Estas tabelas permitem a ligação entre informação dos empregadores e dos trabalhadores, tendo sido utilizadas, neste estudo, as tabelas referentes ao período de 1991 a 2021.

2. Preparação dos dados

Esta secção descreve, pela sua ordem de implementação, os passos tomados para assegurar a consistência e a precisão dos dados utilizados, enquanto tentando manter o máximo de observações possível.

2.1. Harmonização

Os dados das variáveis de escolaridade foram harmonizados de modo a fazer corresponder as categorias anteriores a 2006 com as categorias posteriores a 2006, para garantir a coerência entre os diferentes anos.

Embora existam quebras estruturais relevantes noutras variáveis, estas não têm qualquer influência no nosso contexto.

2.2. Tratamento dos dados originais em falta

Os dados em falta em certo ano e para certo indivíduo foram colmatados para as variáveis 'sexo', 'data de nascimento' e 'escolaridade', sempre que existiam dados consistentes e não em falta nos restantes registos. Por exemplo, se o dado sobre a categoria de sexo estava em falta em dado ano, mas outras entradas para o mesmo indivíduo reportavam consistentemente uma dada categoria (por exemplo, sexo masculino), a categoria mais comum foi utilizada para imputar os dados em falta.

2.3. Tratamento de inconsistências

Sexo: no caso de deteção de inconsistências nos dados relativos ao sexo e se as datas de nascimento eram consistentes, foi utilizada a categoria de sexo mais comum para correção.

Data de nascimento: no caso de deteção de inconsistências nas datas de nascimento e se os reportes de categoria de sexo eram consistentes, o valor mais comum da data de nascimento foi utilizado para correção.

Escolaridade: no caso de deteção de inconsistências, e apenas quando o sexo e data de nascimento já se encontravam consistentes, as correções foram feitas com base na consistência histórica da escolaridade. Foram verificadas condições específicas, como, por exemplo, se os dados de escolaridade diminuíam ou aumentavam apenas uma vez e depois regressam ao nível anterior. As situações de diminuição da escolaridade foram tratadas como anómalas, mas o aumento da escolaridade foi permitido. O nível de escolaridade anterior mais comum foi utilizado para imputações, sempre que adequado.

Após as correções dos dados em falta e das inconsistências, o conjunto de dados foi submetido a uma nova verificação da coerência das informações relativas ao sexo e à data de nascimento. Quaisquer inconsistências remanescentes foram assinaladas como valores em falta. As entradas com valores em falta foram então eliminadas, mantendo-se apenas as entradas completas para cada indivíduo.

2.4. Amostra considerada

Foram mantidos apenas os indivíduos com idade igual ou superior a 16 anos, bem como aqueles com idade igual ou inferior a 65 anos. Além disso, foram excluídas as profissões relacionadas com as forças armadas, a agricultura e a pesca. Foram consideradas somente as remunerações horárias superiores a zero.

2.5. Remoção de valores extremos

Apenas foram analisados os valores extremos presentes na distribuição, ano a ano, da nossa variável-alvo, uma vez que podiam ter um impacto significativo na análise de regressão e nos algoritmos de florestas causais (ver Secção 3, abaixo) devido à sua influência desproporcional face ao seu peso na amostra. Assim, aplicámos uma transformação logarítmica à distribuição do salário horário, limitando as suas caudas com base no método de identificação de valores extremos de Tukey, utilizando-se o parâmetro k=2. A escolha da transformação logarítmica é justificada pela distribuição aparentemente log-normal observada nos dados. É importante notar que a estratégia de identificação de valores extremos de Tukey é geralmente adequada para distribuições que se aproximam da normalidade (Tukey, 1977).

3. Metodologia

3.1. Análise de regressão linear: Prémios salariais médios de um nível de ensino adicional

3.1.1 Especificação do modelo

A primeira parte da nossa análise estima o efeito médio do tratamento da escolaridade nos salários através de um modelo de regressão linear. Especificamente, utilizamos uma equação salarial, ao estilo da equação Minceriana, da seguinte forma:

$$log\ (Remuneração\ horária) = \beta_0 + \sum_{j=1}^{k-1} \beta_j \cdot D_{ij} + \beta_k \cdot idade + \beta_{k+1} \cdot idade_i^2 + \beta_{k+2} \cdot sexo\ +\ \epsilon_i$$

onde D_{ij} são indicadores binários para $k-1$ diferentes níveis de educação.

Considerámos que a abordagem de variável instrumental (IV) era muito limitadora para os nossos objetivos. Os potenciais instrumentos identificados estavam relacionados apenas com os anos de escolaridade e não com níveis de escolaridade específicos, apresentando, além disso, pouca variabilidade anual, nomeadamente por sexo. Estas limitações não permitiam uma análise flexível da dinâmica dos prémios salariais associados aos diferentes níveis de ensino. Por conseguinte, optámos

por um modelo de regressão linear com indicadores binários para diferentes níveis de escolaridade (exceto um).

Na nossa análise, e no que diz respeito aos dados disponíveis, consideramos que a idade e o sexo são as únicas características observáveis dos indivíduos que não foram influenciadas pelo seu nível de educação no momento da observação. O condicionamento de quaisquer outras variáveis poderia introduzir um enviesamento adicional nas estimativas do efeito do tratamento. Para evitar este risco, optou-se por não condicionar a estimação da nossa quantidade de interesse a quaisquer outras variáveis.

Reconhecemos que o desenho metodológico empregue não satisfaz adequadamente as condições de identificabilidade necessárias para se poder considerar a quantidade de interesse aqui discutida como causal, uma vez que existem potencialmente características relevantes que confundem o tratamento e que não são observáveis nos nossos dados. Por conseguinte, consideramos as nossas estimativas como as melhores possíveis, tendo em conta os dados disponíveis.

3.1.2. Procedimento de estimação

Os parâmetros da equação Minceriana são estimados pelo método dos mínimos quadrados (OLS). Os coeficientes β_j captam o retorno médio do nível de educação correspondente, *j*, relativamente à categoria omitida (inferior ao 9º ano de escolaridade). Note-se que esta quantidade ainda não é a nossa quantidade de interesse.

Para obter o prémio salarial médio associado a um nível de ensino adicional, é necessário considerar as diferenças entre os regressores dos sucessivos níveis de escolaridade.

De forma a permitir inferência, estimamos o erro padrão da diferença entre os coeficientes (por exemplo, ensino superior e ensino secundário) utilizando o seguinte procedimento. Sejam $\hat{\beta}_{\text{Licenciatura}}$ e $\hat{\beta}_{\text{Secundário}}$ os coeficientes estimados para o ensino superior e para o ensino secundário, respetivamente. A diferença entre esses coeficientes é dada por:

$$\Delta\hat{\beta} = \hat{\beta}_{\text{Licenciatura}} - \hat{\beta}_{\text{Secundário}}$$

Definimos um vetor de contrastes a de dimensão $k-1$ que especifica os coeficientes a serem subtraídos,

$$\mathrm{a} = \begin{pmatrix} 0 \\ \ldots \\ -1 \\ \ldots \\ 1 \\ \ldots \\ 0 \end{pmatrix}$$

Seja $\mathrm{Var}(\hat{\beta})$ a matriz de covariância dos coeficientes estimados, obtida a partir de um estimador da matriz de covariância robusto a heterocedasticidade. A variância da diferença $\Delta\hat{\beta}$ pode, então, ser calculada utilizando a matriz de covariância e o vetor de contraste a:

$$\mathrm{Var}(\Delta\hat{\beta}) = a^{\top}\mathrm{Var}(\hat{\beta})a$$

O erro padrão da diferença é a raiz quadrada desta variância:

$$\mathrm{SE}(\Delta\hat{\beta}) = \sqrt{a^{\top}\mathrm{Var}(\hat{\beta})a}$$

Este procedimento fornece o erro padrão da diferença entre os coeficientes, permitindo-nos construir intervalos de confiança e efetuar testes de hipóteses sobre os retornos incrementais de um nível adicional de escolaridade.

3.2. Florestas Causais: modelação de efeitos de tratamento heterogéneos

Nesta secção, apresentamos as florestas causais, uma técnica avançada adequada a modelar efeitos de tratamento heterogéneos. Ao contrário dos modelos de regressão tradicionais, que estimam os efeitos médios do tratamento, as florestas causais permitem explorar a forma como o impacto de uma intervenção varia entre diferentes indivíduos.

3.2.1. Quadro conceptual

O conceito fundamental por trás das florestas causais envolve a construção de um conjunto de árvores de decisão, adaptadas para serem árvores causais (Athey & Imbens, 2016), em que cada árvore representa uma possível partição dos dados em subgrupos com base em covariáveis. Ao contrário das árvores de decisão padrão, que visam prever diretamente a variável de resultado, as árvores causais concentram-se na estimativa do efeito médio condicional do tratamento (CATE) para cada subgrupo. As florestas causais resultam da agregação de várias árvores causais, construídas tendo por base diferentes amostras de bootstrap dos dados iniciais, utilizando cada uma subconjuntos de variáveis escolhidos aleatoriamente.

A análise com recurso ao conceito de CATE tem por objetivo compreender a forma como o efeito de um tratamento varia entre diferentes subgrupos de uma população, em função das suas características observadas. Baseia-se no quadro dos resultados potenciais, que pressupõe que cada unidade de uma população tem um resultado potencial nas condições de tratamento e de controlo. A análise CATE

reconhece que os efeitos do tratamento podem variar entre diferentes indivíduos ou grupos devido à heterogeneidade de características como a idade, o sexo, o estatuto socioeconómico ou outras covariáveis relevantes.

3.2.2. Especificação do modelo

Formalmente, o modelo de floresta causal estima o efeito do tratamento τ_i para cada indivíduo i como:

$$\hat{\tau}_i = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i]$$

onde $Y_i(1)$ e $Y_i(0)$ representam os resultados potenciais do indivíduo i em condições de tratamento e de controlo, respetivamente, e X_i denota o vetor de covariáveis.

Utilizamos apenas dados de dois níveis de escolaridade consecutivos, em que o nível de escolaridade mais elevado representa o tratamento e o nível mais baixo serve de contrafactual. No contexto do nosso estudo, o nosso objetivo é captar o efeito da obtenção de um nível adicional de escolaridade. Consequentemente, o grupo de tratamento é constituído por indivíduos que prosseguiram um nível de escolaridade adicional, enquanto o contrafactual assume, *ceteris paribus*, o que teria acontecido aos indivíduos não tivessem estes prosseguido esse nível adicional de estudos. Assim, os indivíduos com maior escolaridade constituem a amostra tratada, enquanto os indivíduos com menor escolaridade constituem a amostra de controlo.

3.2.3. Procedimento de estimação

O procedimento de estimativa envolve os seguintes passos:

1. **Árvores de crescimento**: os dados são divididos em subconjuntos para a construção de árvores e a estimativa do efeito do tratamento. As árvores de decisão são construídas utilizando amostras de *bootstrap* dos dados. As divisões da árvore são escolhidas para maximizar a heterogeneidade dos efeitos do tratamento entre os nós filhos resultantes da partição. Os efeitos do tratamento em cada folha da árvore são estimados utilizando dados que não foram "vistos" durante a construção da árvore, evitando p-*hacking* e o sobreajuste do modelo.

2. **Construção do conjunto de árvores (floresta)**: os efeitos de tratamento individuais previstos por cada árvore são agregados para obter uma estimativa de conjunto do efeito do tratamento final, dado um conjunto de características observáveis.

3. **Previsão**: uma vez construída a floresta causal, esta pode ser utilizada para prever o efeito do tratamento para novos indivíduos com base nos seus valores de covariáveis.

As árvores construídas são complementadas por um estimador duplamente robusto (R-*learner*) que ajuda a colmatar potenciais enviesamentos decorrentes de características observáveis e da utilização de algoritmos de aprendizagem automática para os controlar (especificamente contornando os enviesamentos resultantes da regularização implícita feita pelas árvores).

Para a construção do R-*learner*, utilizámos a regressão linear para modelar o resultado e a regressão logística para modelar o tratamento, utilizando a mesma especificação que a regressão Minceriana apresentada nas secções anteriores. Optámos por estes métodos porque o conjunto limitado de fatores de confundibilidade e as possíveis não linearidades não justificavam a utilização de algoritmos mais complexos, como as florestas aleatórias ou *gradient boosting*.

3.2.4. Heterogeneidade induzida por mediadores no âmbito da estrutura da floresta causal

Uma caraterística fundamental das florestas causais, que justifica a sua utilização neste contexto, é a sua capacidade de captar a heterogeneidade. A versão original do algoritmo trata todas as características como fontes de heterogeneidade e potenciais fatores de confusão, simultaneamente. A construção das árvores, por si, controla rudemente os efeitos de confundibilidade existentes em qualquer das características consideradas. Isto deve-se ao facto de as unidades (tratadas e de controlo) que estão a ser comparadas durante a estimativa do efeito se tornarem semelhantes em termos de características devido às sucessivas divisões da árvore.

No nosso caso, permitimos que algumas características determinem diferentes efeitos de tratamento sem as tratarmos como fatores de confundibilidade. Estas características revelam-se, frequentemente, apenas após o tratamento e podem inclusive ser influenciadas pelo tratamento; são conhecidas na literatura por características mediadoras do efeito causal.

Embora não estejamos diretamente interessados numa análise de mediação causal tradicional, em que se tenta identificar e quantificar o efeito dos mediadores na cadeia causal, é importante não tratar os mediadores como fatores de confundibilidade. De outro modo, a sua inclusão poderia enviesar as estimativas do efeito do tratamento, tal como discutido anteriormente neste apêndice para o caso de inclusão de características adicionais na especificação do modelo de regressão linear.

A nossa abordagem distingue claramente entre fatores de confusão e mediadores. Adaptámos o algoritmo de construção de árvores fornecido pelos autores para considerar estes diferentes conjuntos de características separadamente e de forma distinta.

As alterações específicas efetuadas no algoritmo são as seguintes:

- Os fatores de confundibilidade são avaliados quanto à qualidade da divisão, considerando todas as unidades, tanto tratadas como de controlo.
- A melhor regra de divisão baseada em fatores de confundibilidade é utilizada para atribuir as unidades tratadas e de controlo a diferentes ramos da árvore.
- Os mediadores são avaliados quanto à qualidade da divisão que possibilitam, mas a regra de divisão a ser testada apenas afeta a divisão das unidades tratadas, sendo as unidades de controlo atribuídas aleatoriamente a qualquer ramo da árvore.
- A melhor regra de divisão baseada em mediadores é utilizada para atribuir apenas as unidades tratadas a diferentes ramos da árvore, enquanto as unidades de controlo são atribuídas aleatoriamente entre os ramos resultantes da divisão das unidades tratadas.
- Os critérios de paragem do algoritmo foram simplificados para considerar apenas a dimensão da amostra das unidades tratadas e de controlo, separadamente, presente nos nós terminais (folhas) resultantes das sucessivas divisões feitas pelas árvores.

Estas alterações foram validadas pelo facto de a inclusão de mediadores não ter alterado estatisticamente os efeitos médios do tratamento calculados pela nossa versão adaptada do algoritmo da floresta causal. Para além disso, estas estimativas são estatisticamente iguais às oriundas da versão original do algoritmo para os mesmos dados. São esperadas apenas pequenas diferenças devido à natureza aleatória de certas partes do procedimento. Além disso, diferentes execuções do algoritmo produzem conclusões coincidentes, desde que um grande número de árvores seja cultivado (combinámos 2000 árvores diferentes na nossa implementação para construir a floresta final).

Referências bibliográficas

Athey, Susan and Imbens, Guido. Recursive partitioning for heterogeneous causal effects, *Proceedings of the National Academy of Sciences,* Vol. 113, Issue 27, pp. 7353–7360, 2016.

Athey, Susan, Julie Tibshirani and Stefan Wager. Generalized random forests, *Annals of Statistics,* 47(2): 1148-1178, 2019.

Mincer, Jacob. Investment in human capital and personal income distribution, *Journal of Political Economy*, 66 (4): 281–302, 1958.

Tukey, John W. *Exploratory Data Analysis*. Addison-Wesley Pub, 1977.

www.planapp.gov.pt

PLANAPP

@planapp_

Newsletter

PLANAPP

Boas Políticas Públicas
Melhor Sociedade