MESTRADO

MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

TRABALHO FINAL DE MESTRADO

RELATÓRIO DE ESTÁGIO

DOCUMENTO ESPECIALMENTE ELABORADO PARA A OBTENÇÃO DE GRAU DE MESTRE

DOCUMENTO PROVISÓRIO

FATORES QUE INFLUENCIAM O COMPORTAMENTO DOS UTILIZADORES NUM QUESTIONÁRIO *WEB*

RAQUEL BARROS NUNES

ORIENTAÇÃO:

PROFESSOR DR. PAULO PARENTE
DR.ª TÂNIA CORREIA

2021

AGRADECIMENTOS

Ao Instituto Nacional de Estatística, agradeço a oportunidade de realizar este estágio. Em especial, agradeço à Dr.ª Tânia Correia por todo o apoio e esclarecimentos essenciais para a execução deste trabalho.

À coordenação do Mestrado de Métodos Quantitativos para a Decisão Económica e Empresarial agradeço por toda a ajuda, em particular ao Professor Paulo Parente, cuja orientação e disponibilidade permanente foram absolutamente essenciais para a realização deste trabalho.

Agradeço também ao Professor Doutor João C. M. Santos Silva por ter sugerido a utilização do método Poisson ML.

À minha família em geral, agradeço pelo apoio incondicional. Ao meu pai e à minha mãe pela paciência e ajuda, à Joana por manter a minha sanidade mental. A todos os meus avós por quererem sempre saber como o meu TFM estava, mesmo quando não percebiam muito bem do que se tratava; ao Avô Manel, em particular, por compreender melhor que ninguém a importância de uma educação superior. Ao Miguel, por ser o meu companheiro de todas as horas e a pessoa que mais acredita em mim.

Por fim, aos meus amigos de Licenciatura e de Mestrado, Bárbara, Sara, Alex, Margarida e Raimundo, que contribuíram através de muitas horas de apoio moral e revisões ao presente documento para que isto fosse possível.

RESUMO

A utilização de inquéritos *web* é cada vez mais frequente. Neste relatório de estágio, pretende-se relacionar paradados e dados auxiliares resultantes da resposta a um inquérito com o comportamento dos utilizadores no questionário *web*. Nesse sentido, utilizaram-se modelos de Poisson estimados por Pseudo-Máxima Verosimilhança para modelar o tempo total de resposta ao questionário e utilizaram-se modelos Probit para modelar a probabilidade de acesso ao inquérito na *internet* e a probabilidade de abandono desse mesmo questionário. Encontrou-se evidência estatística de que o tempo de resposta aumenta quando um mesmo respondente utiliza dispositivo móvel e computador em simultâneo, quando a sua idade aumenta e quando o número de insucessos no login aumenta. O tempo de resposta diminui quando o respondente utiliza apenas dispositivo móvel na resposta, face a quem usa apenas computador. Relativamente ao acesso ao inquérito *web*, as mulheres têm menos probabilidade de o fazer, bem como os indivíduos residentes em áreas rurais. Quanto maior o nível de educação completa, maior a probabilidade de tentar responder via *web*. A evidência estatística aponta para que as mulheres e os residentes em áreas medianamente urbanas tenham probabilidade acrescida de abandonar o questionário.

Palavras-chave: Questionário *web*; Paradados; Probit; PPML; Comportamento do Respondente; Características do Respondente

ABSTRACT

Web surveys are being used more and more everyday. In this internship report we want to study the relationship between paradata and auxiliar data and the respondent's behaviour in the web survey. We used Poisson Pseudo Maximum Likelihood estimators to model total response time. Probit models were used to study the respondents' probability of accessing and breaking-off from the web survey. We found evidence that the use of computer and mobile device by the same respondent, their age and the number of unsuccessful logins have a positive effect on response time. Time response is shorter when respondents use a mobile device to complete the survey, compared to respondents exclusively using the computer. We found out that women and those who live in rural areas are less likely to access the web survey. On the other hand, education has a positive effect on the probability of loging in to the survey. It was found that women and respondents living in averagely urban areas are more likely to break-off from the web survey.

Keywords : Web Survey; Paradata; Probit; PPML; Response Behaviour; Respondent Characteristics

Índice

Índice de Tabelas

Lista de Abreviaturas

AMU – Área Medianamente Urbana

APR – Área Predominantemente Rural

APU – Área Predominantemente Urbana

CAPI – *Computer Assisted Personal Interview*

CATI – *Computer Assisted Telephone Interview*

CAWI – *Computer Assisted Web Interview*

INE – Instituto Nacional de Estatística

IUTICF – Inquérito à Utilização de Tecnologias de Informação e da Comunicação pelas Famílias

PPML – *Poisson Pseudo-Maximum Likelihood*

TIPAU – Tipologia de Áreas Urbanas

1. Introdução

Apesar de não serem novos, os inquéritos via *web* trazem promessas de facilidade de administração e análise dos dados, tendo a frequência da sua utilização aumentado a cada ano (Daikeler et al., 2020). Não obstante, estudos revelam que as taxas de resposta a questionários *web* são consistentemente mais baixas do que em outros modos de recolha (Daikeler et al, 2020). O presente documento resulta do trabalho desenvolvido no estágio realizado no Instituto Nacional de Estatística (INE), no Departamento de Recolha e Gestão de Dados, e foi motivado, em primeira instância, pelo facto de, ainda que os inquéritos *online* estejam estabelecidos no INE, os paradados (dados que resultam do processo de resposta ao questionário (Callegaro, 2010)) neles recolhidos, e que têm um grande potencial de análise e informação, não têm sido alvo de estudo sistematizado.

O principal objetivo do trabalho desenvolvido é a análise das interações dos respondentes com o Inquérito à Utilização de Tecnologias de Informação e da Comunicação pelas Famílias (IUTICF), resultantes das entrevistas CAWI (*Computer Assisted Web Interview*ing), utilizando para isso paradados, dados de resposta e dados auxiliares do mesmo inquérito. A finalidade é a otimização do processo de recolha de resposta e os custos associados ao inquérito.

Pretende-se assim responder às seguintes questões:

Q1: Existe relação entre o tempo de resposta e as características do respondente?

Q2: Existe relação entre a tentativa de resposta em CAWI e as características do respondente?

Q3: Existe relação entre o abandono do questionário em CAWI e as características do respondente?

No Capítulo 2, apresenta-se uma revisão da literatura e, em particular, discutem-se alguns trabalhos anteriores que desenvolveram o tema dos inquéritos *web*, dos paradados e a influência das características sócio-demográficas dos respondentes no comportamento perante os inquéritos. No Capítulo 3, expõe-se a metodologia utilizada. No Capítulo 4, explicam-se os processos de preparação dos dados e apresentam-se estatísticas descritivas. No Capítulo 5, são apresentados e discutidos os principais resultados obtidos e o Capítulo 6 é dedicado à apresentação das principais conclusões do estudo desenvolvido.

2. Revisão de Literatura

Nas últimas décadas, a melhoria significativa dos *browsers* e do acesso à *internet* viabilizou a realização de inquéritos na *web*, cuja utilização é cada vez mais generalizada (Daikeler et al., 2020). Apesar de os questionários *online* serem frequentemente informais, muitos organismos oficiais começam a implementá-los, como é o caso do INE.

Em parte, a popularidade dos inquéritos *web* resulta das vantagens que trazem, quando comparados com os inquéritos com recolha presencial ou via telefone. Os inquéritos realizados via *internet* têm custos mais baixos, devido aos custos variáveis praticamente inexistentes – o processo de resposta custa aproximadamente o mesmo, contenha a amostra cinco ou cinco mil indivíduos. As respostas podem ser recolhidas mais rapidamente e o processo de resposta pode ser mais curto também para o respondente. A autoadministração não só previne a necessidade de contratar entrevistadores como facilita a resposta a questões de teor mais sensível. A flexibilidade geográfica e temporal são também vantagens que não devem ser ignoradas e que facilitam a vida do respondente. Os questionários *web* são mais fáceis de implementar e permitem a utilização de multimédia (Callegaro et al., 2015), embora atualmente a tecnologia já seja transversal a outros modos de recolha. Adicionalmente, facilitam a recolha de paradados (Matjašič et al., 2018)

Contudo, as baixas taxas de resposta nos inquéritos *online*, quando comparadas a entrevistas presenciais ou telefónicas, são uma preocupação bem presente. Em 2008, Manfreda et al. (2008) realizaram uma análise comparativa entre taxas de resposta de inquéritos *web* e outros modos (telefone, carta, fax, etc.). Concluíram que os inquéritos *online* tinham uma taxa de resposta 11% menor do que os outros modos. Em 2020, Daikeler et al. (2020) estenderam o número de inquéritos analisados e realizaram uma nova meta-análise, descobrindo que, ao contrário do que se esperava, a significativa melhoria no acesso à *internet* não tinha mudado as baixas taxas de resposta aos inquéritos *web*. Os autores apontaram duas possíveis causas. Por um lado, este tipo de inquéritos é encarado com falta de seriedade, porventura por ser mais fácil de criar, levando o respondente a pensar que o investigador não se esforçou tanto. Por outro, a frequência exagerada com que as pessoas recebem convites para participar em inquéritos *online* torna difícil distinguir os importantes dos menos importantes. A realização deste tipo de

questionários tem também outros problemas associados. Primeiro, é necessário que o respondente se sinta confortável com a utilização de um computador e da *internet*, caso contrário, poderá ser difícil completar o questionário. Mais, o questionário deixa de ser apenas o meio de recolha de informação e passa também a ser o meio de comunicação com o respondente. Decorre daqui que negligenciar um bom texto introdutório e instruções claras pode comprometer as taxas de resposta. Finalmente, a possibilidade de responder mais rapidamente ao questionário pode originar respostas de menor qualidade (Callegaro et al. , 2015).

Associada aos inquéritos *web*, está também a recolha de paradados. Apesar de outros modos de recolha gerarem paradados, os inquéritos *web* popularizaram-nos. A importância da recolha e estudo de paradados em questionários CASIC (*Computer-Assisted Survey Information Collection*) foi referida pela primeira vez por Couper (1998). Este autor designou-os por *process data* e destacou o potencial dos paradados como instrumentos de avaliação e melhoria da qualidade dos questionários. Múltiplas definições de paradados surgiram nos anos posteriores e, ainda que não exista uma definição definitiva na literatura (Kreuter, 2013), existe consenso que paradados são os dados que resultam do processo de resposta a um inquérito (Callegaro, 2010; Kreuter, 2018; Lynn & Nicolaas, 2010; Matjašič et al, 2018). Os paradados podem ser recolhidos do lado do servidor ou do lado do cliente. A recolha através do servidor possibilita apenas o acesso a informações relativas à página (*e.g.* mudança de página, *timestamp*, erro de autenticação), enquanto que a recolha do lado do cliente proporciona dados ao nível de cada pergunta como, por exemplo, número de cliques ou alterações de resposta (Callegaro, 2013). Os paradados podem também ser agrupados em dados do dispositivo e dados de navegação do questionário. No primeiro grupo, incluem-se o tipo de dispositivo, o sistema operativo, o *browser* utilizado e o endereço IP. Os dados de navegação do questionário podem ser o estado da autenticação, movimentos entre páginas, mensagens de erro e acesso a páginas de ajuda. Uma listagem mais exaustiva de exemplos de paradados em *web surveys*, pode ser encontrada em Callegaro (2013).

Não basta recolher paradados, é necessário maximizar o seu potencial de forma a melhorar a qualidade dos dados e a experiência do respondente (West, 2011). Os *timestamps* permitem saber quanto tempo esteve o respondente em cada página, as mensagens de erro possibilitam a identificação de questões potencialmente problemáticas

e o dispositivo utilizado pode influenciar positiva ou negativamente a experiência do indivíduo, bem como o seu tempo de resposta (Callegaro, 2010). Estudos que envolvam paradados podem ser enriquecidos através da inclusão de dados auxiliares. Os dados auxiliares são dados que não resultam da resposta ao questionário, mas que provêm de fontes externas (Callegaro, 2013). Este tipo de dados permite, por exemplo, relacionar o tempo de resposta ao questionário com a área de residência do respondente.

Os investigadores têm procurado estudar as possíveis relações entre as características do respondente, os paradados que resultam da sua resposta e o seu comportamento perante o inquérito. Alguns destes estudos incidem sobre o modo CAWI mas, sendo a literatura sobre o tema ainda escassa, estudos sobre inquéritos noutros modos de recolha foram também considerados. Bosch et al. (2019) investigaram o comportamento dos *millennials*[1] perante os inquéritos *online* e descobriram que estes são menos propensos a responder, que a maioria responde usando um dispositivo móvel e que as suas taxas de *break-off* não diferem das dos respondentes mais velhos. Couper & Kreuter (2013) estudaram o tempo de resposta a cada pergunta, num inquérito com recolha CAPI, individualmente, e procuraram relacioná-lo com as características do respondente. Descobriram que um aumento na idade do respondente leva a um aumento no tempo de resposta e que um nível de escolaridade completo mais alto diminui a duração das entrevistas. Shi et al. (2018) encontraram evidência de que o tempo de resposta é muito menor para homens e para pessoas com maior nível de escolaridade completo. Por sua vez, Yan & Tourangeau (2008) descobriram, tendo por base quatro inquéritos *web*, que os indivíduos com maior nível de educação completo demoram menos tempo a responder ao questionário e que indivíduos mais velhos demoram mais tempo. Relativamente à adesão aos questionários *web*, Kelfve et al. (2020) determinaram que mulheres, pessoas reformadas ou solteiras, indivíduos com menor nível de educação, com maior incidência de depressão e menos saúde têm maiores probabilidade de não aceder aos inquéritos na *internet* do que um indivíduo sem estas características.

[1] São denomidados de *millennials* os indivíduos nascidos entre 1982 e 2003 (Strauss & Howe, 1991)

3. Metodologia

3.1 Modelo de Poisson estimado por Pseudo-Máxima Verosimilhança

Na estimação da regressão do tempo de resposta (y), é possível utilizar um modelo de regressão linear múltipla, com a forma apresentada em (1):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \quad (1)$$

em que $E(u|x_1, x_2, \dots, x_k) = 0$ e os $x_1, x_2, \dots, x_k$ são os regressores.

No entanto a variável y é não negativa e a estimação de (1) pode conduzir a valores negativos para os valores ajustados.

Se o tempo de resposta for uma variável estritamente positiva, uma solução possível passa por logaritmizar a variável dependente, como em (2) (cf. Lancaster, 1990, p. 22 e p.41):

$$\log(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \quad (2)$$

em que $E(u|x_1, x_2, \dots, x_k) = 0$. A equação (2) permite valores negativos para $E(\log(y)|x_1, x_2, \dots, x_k)$. No entanto, tal como mencionado por Santos Silva & Tenreyro (2006), estimar $E(\log(y)|x_1, x_2, \dots, x_k)$ não é o mesmo que estimar $E(y|x_1, x_2, \dots, x_k)$, nem mesmo transformando o primeiro em $e^{E(\log(y)|x1,x2,\dots,xk)}$. Adicionalmente, a transformação logarítmica não pode ser aplicada quando existem observações com y = 0. Santos Silva & Tenreyro (2006) sugerem então a utilização do estimador de Pseudo-Máxima Verosimilhança de Poisson (PPML, do inglês *Poisson Pseudo-Maximum Likelihood*), que maximiza a função logaritmo de verosimilhança baseada na distribuição de Poisson. O modelo utilizado pelo PPML assume $E(y|x_1, x_2, \dots, x_k) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$. Como a função exponencial só tem ordenadas positivas, o PPML resolve os dois problemas apontados à transformação logarítmica.

Este estimador soluciona outro problema associado a variáveis dependentes de valor não negativo. Usualmente, este tipo de dados apresenta heteroscedasticidade muito forte, ou seja, a variância condicional de y depende muito dos regressores $x_1, x_2, \dots, x_k$. Assim, qualquer estimador de $\beta_0, \beta_1, \dots, \beta_k$ que não tenha em conta a heteroscedasticidade vai produzir estimativas muito diferentes do verdadeiro valor de $\beta_1, \beta_2, \dots, \beta_k$ em amostras pequenas. No PPML, temos $Var(y|x_1, x_2, \dots, x_k) = e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}$, que por sua vez é igual ao valor esperado de y, isto é:

$$E(y|x_1, x_2, \dots, x_k) = Var(y|x_1, x_2, \dots, x_k) = e^{\beta_0+\beta_1 x_1+\cdots+\beta_k x_k} \quad (3)$$

Assumindo a hipótese (3) como verdadeira, e considerando uma amostra aleatória de tamanho n $\{(y_i, x_{i1}, x_{i2}, \dots, x_{ik}), i = 1, \dots, n\}$, é possível estimar $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ resolvendo o conjunto de condições dado por (4), que é o estimador PPML:

$$\sum_{i=1}^{n} \left[y_i - e^{x_i'\tilde{\beta}}\right] x_i = 0 \quad (4)$$

em que $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})'$. Para o estimador $\tilde{\beta}$ ser consistente para β, é necessário que a média condicional seja bem especificada. Isto permite a aplicação do PPML a dados que não sejam Poisson, mesmo que não sejam inteiros. Note-se que para fazer inferência estatística é necessário utilizar erros padrão robustos (Santos Silva & Tenreyro, 2006).

3.2 Modelo Probit

As questões de investigação Q2 e Q3 do capítulo 1 requerem modelos com variável dependente binária. Existem modelos próprios para este tipo de variáveis, os modelos de resposta binária. São exemplos desses modelos o modelo de probabilidade linear (MPL), o modelo Logit e o modelo Probit. O modelo Probit tem, relativamente ao MPL, a vantagem de garantir que a probabilidade de y = 1 se encontra entre 0 e 1. Optar pelo modelo Probit ou pelo modelo Logit baseia-se apenas numa questão de preferência. Assim, no desenvolvimento da análise, optou-se por utilizar o modelo Probit. Esta subsecção tem por base o exposto em Wooldridge (2016).

A forma do modelo Probit, que tem como objetivo traduzir os efeitos de x na probabilidade de y = 1, é a apresentada em (5).

$$P(y = 1|x) = \Phi(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) \quad (5)$$

Onde $\Phi(.)$ é a função de distribuição da distribuição normal estandardizada. Assim, é possível garantir que a probabilidade de y = 1 está contida no intervalo entre 0 e 1, para qualquer valor das variáveis explicativas.

Os coeficientes $(\beta_0, \beta_1, \dots, \beta_k)$ em (5) não exprimem diretamente a magnitude do efeito dos regressores sobre a probabilidade, no entanto, o seu sinal indica-nos imediatamente se o efeito é positivo ou negativo. Para quantificar o efeito que uma variável explicativa tem sobre a variável dependente, é necessário calcular o Efeito

Parcial Médio (EPM), como em (6), no caso da variável x_j ser contínua, ou como em (7), se a variável x_j for discreta:

$$EPM_j = n^{-1}\sum_{i=1}^{n}[\phi(x_i'\hat{\beta})\hat{\beta}_j] \quad (6)$$

Em que $x_i = (x_{i1}, x_{i2}, \dots, x_{ik})'$ *e* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)'$ e $\phi(.)$ é a função densidade de probabilidade da variável aleatória normal estandardizada.

$$EPM_j = n^{-1}\sum_{i=1}^{n}\{\Phi[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j(x_j + 1) + \cdots + \hat{\beta}_k x_{ik}] - \Phi[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_k x_{ik}]\} \quad (7)$$

3.3 Teste RESET

A aplicação do teste RESET pretende detetar problemas de especificação da forma funcional no modelo estimado. No caso dos modelos de regressão linear, se o modelo estimado satisfizer a hipótese da média condicional igual a zero, ou seja, se o valor esperado do erro no modelo da população for 0, para qualquer valor assumido pelas variáveis explicativas, nenhuma função não linear das variáveis independentes deve ser estatisticamente significativa, quando adicionada ao modelo estimado original. No caso dos modelos não lineares, como é o caso do modelo Probit, se o modelo estimado for $E(y|\mathbf{x}) = G(\beta 0 + \beta 1 x1 + \cdots + \beta k x k)$, nenhuma função não linear das variáveis independentes deve ser estatisticamente significativa, quando adicionada ao modelo estimado original. De acordo com Wooldridge (2016), efetuamos o teste adicionando a função quadrática e cúbica de $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$ como regressores adicionais: $\hat{y}^2$ e $\hat{y}^3$. Se estes dois regressores não forem estatisticamente significativos o modelo não tem problemas de especificação. Se estes regressores forem estatisticamente significativos existe um problema de especificação.

4. Dados

Os dados utilizados neste estudo são resultado da administração do IUTICF. O responsável por esta operação estatística, ao nível nacional, é o INE, sendo um inquérito anual, por amostra. Realizou-se pela primeira vez em 2001 e, nos primeiros dois anos, estava integrado no Inquérito ao Emprego. É um inquérito harmonizado ao nível europeu,

e resultou de um plano de ação do Conselho Europeu para a produção de estatísticas comparáveis entre Estados-Membros sobre os indivíduos e os agregados domésticos e a sociedade de informação. O seu objetivo é *"a recolha e produção de indicadores sobre o acesso por parte dos agregados domésticos privados às Tecnologias da Informação e da Comunicação (TIC) e à utilização de TIC pelos indivíduos"* (Instituto Nacional de Estatística, 2019).

A amostra é selecionada de forma aleatória e cada alojamento permanece durante 4 anos, existindo uma rotação de ¼ da amostra anualmente. Anualmente, antes do início do período de recolha, é enviada uma carta ou email[2] para cada alojamento a referir a seleção para resposta e condições de participação. Apesar da carta informar sobre as características do indivíduo a ser selecionado, qualquer um dos residentes pode iniciar a resposta ao questionário, completando as informações de todos os residentes, como o nome e a data de nascimento. Depois, é selecionado o indivíduo cujo aniversário se celebrou há menos tempo à data da entrevista e que tem, a 31 de março desse ano, entre 16 e 74 anos.

O IUTICF realizou-se, nos três anos em estudo, em modos mistos: 2018 e 2019 em CAWI, CATI (*Computer Assisted Telephone Interview)* e CAPI (*Computer Assisted Personal Interview)* e 2020, em CAWI e CATI, sendo que a vertente presencial foi suspensa devido à COVID-19. Os dados são recolhidos num formulário *online* desenvolvido em .net integrado no sistema de gestão de inquéritos às Famílias (Gestão de Processos de Inquéritos por Entrevista). Ao nível das questões do inquérito, o formulário é igual para qualquer modo de resposta, diferindo apenas na formulação das questões iniciais de caracterização do alojamento e residentes devido ao facto de em CAWI ser de autopreenchimento.

4.1 Preparação dos dados

Os dados utilizados na presente análise têm várias origens, tendo sido selecionados os campos pertinentes de cada tabela para a análise e criadas novas bases de dados para o estudo. São elas:

1. Tabela de Paradados, com dados recolhidos do lado do servidor

[2] O canal do envio da comunicação depende de estratégias de comunicação pré-definidas.

2. Tabela de Apuramento Final (validação de resposta)
3. Tabela da idade, escolaridade e sexo do respondente
4. Tabela de localidade de residência
5. Tabela de classificação de Tipologia de Áreas Urbanas (TIPAU) das freguesias
6. Tabela do número de respostas por módulo

As referidas tabelas apresentavam dados brutos para a totalidade da amostra e, antes da análise dos dados, foi necessário prepará-los através da exclusão de algumas observações e da correção ou adaptação de alguns descritivos. Primeiro, para o ano de 2018, foi necessário corrigir um erro no ID de utilizador da tabela de apuramento final. Todos os indivíduos cujo ID de utilizador incluía o conjunto de letras "NE" foram afetados por uma substituição automática de "NE" para "Residência principal - Selecionado: Não Elegíveis". Foi necessário reverter essa substituição.

Em segundo lugar, foram considerados para o estudo os alojamentos utilizados como residência principal e, dentro destes, os respetivos indivíduos selecionados que responderam ao questionário, ou seja, que tiveram uma entrevista conseguida. Os dados correspondentes aos restantes indivíduos foram excluídos.

Para os anos de 2018 e 2019, a tabela das características dos respondentes podia apresentar mais do que um conjunto de características para o mesmo indivíduo, consequência de, num mesmo alojamento, poder ter existido resposta em mais de um modo. Foi então necessário verificar o modo de apuramento final do indivíduo e validar o conjunto de características recolhidas no mês correspondente a esse modo. Os conjuntos restantes foram excluídos.

Relativamente à tabela de paradados, foi crucial eliminar as observações que correspondiam a testes e ensaios. Utilizando a tabela de apuramento final, foi possível excluir esses casos.

Foi também necessário preparar novas bases de dados, para que a análise fosse possível. Para todas as questões de investigação, foram conjugadas as tabelas de validação de resposta, de características do respondente, de localidade do respondente e de TIPAU. As novas bases de dados foram criadas em *MS Access*.

Para a questão de investigação Q1, em especial, foi necessário criar alguns programas em *Visual Basic for Applications* (VBA) para trabalhar a tabela de paradados. Isolou-se a primeira e última ação de cada sessão, calculou-se depois o tempo decorrido entre essas

duas ações, perfazendo o tempo total de uma sessão. Depois, somou-se o tempo total de todas as sessões de um determinado utilizador de forma a chegar ao tempo total de resposta desse utilizador. Foram também adicionados à base de dados um campo com o número de respostas dadas por utilizador, que resultou da adição do número de respostas por módulo, um campo relativo ao número de insucessos no *login* do questionário *web* e três campos relativos ao tipo de dispositivo utilizado na resposta.

Para a questão de investigação Q2, foi necessário determinar todos os utilizadores que tinham interagido com o questionário *web*, ou seja, que tinham algum tipo paradados.

Relativamente à questão de investigação Q3, foi necessário verificar que indivíduos tinham paradados associados e, desses, quais tinham modo de apuramento final distinto de CAWI. Foi também determinada, utilizando programas em VBA, a última ação de cada indivíduo antes de abandonar o questionário *web*.

Posteriormente, foi necessário codificar as variáveis das novas bases de dados, como é apresentado na Tabela I.

Tabela I - Variáveis em Estudo

Variável	**Descrição**	**Codificação**	**Tipo**
idade	Idade do respondente.		Numérica
mulher	Sexo do respondente. O grupo base é o sexo masculino.	0 = Homem 1 = Mulher	Binária
basico	Nível máximo de escolaridade completa. O grupo base corresponde à não conclusão de qualquer tipo de escolaridade.	0 = Caso contrário 1 = Ensino Básico	Binária
secund		0 = Caso contrário 1 = Ensino Secundário	
superior		0 = Caso contrário 1 = Ensino Superior	
nresp		0 = Caso contrário 1 = Prefere não responder	
scorresp		0 = Caso contrário 1 = O nível completado não têm correspondência com os apresentados	

(continua)

Continuação da Tabela I

Variável	**Descrição**	**Codificação**	**Tipo**
amu	Classificação atribuída à freguesia de residência do respondente, segundo a Tipologia das Áreas Urbanas. O grupo base é a Área Predominantemente Urbana.	0 = Caso contrário 1 = Área Medianamente Urbana	Binária
apr		0 = Caso contrário 1 = Área Predominantemente Rural	
respostas	Número de respostas dadas no questionário.		Numérica
n_insucessos	Número de insucessos no *login* do questionário.		Numérica
d_movel	Tipo de dispositivo(s) utilizado(s) para responder ao questionário. O grupo base é o computador.	0 = Caso contrário 1 = Utilizou dispositivo móvel apenas	Binária
misto		0 = Caso contrário 1 = Utilizou computador e dispositivo móvel	
temposegundos	Tempo total, em segundos, da resposta ao questionário.		Numérica
tentarcawi	Variável que indica se respondente acedeu alguma vez ao questionário *web*.	0 = Não acedeu ao questionário *web* 1 = Acedeu ao questionário *web*	Binária
abandonarcawi	Variável que indica se respondente, depois de iniciar resposta na *web*, concluiu por	0 = Concluiu o questionário na *web*	Binária

	essa via ou por outra (CATI ou CAPI).	1 = Apesar de ter paradados, não concluiu na *web*	

4.2 Estatísticas Descritivas

As estatísticas descritivas dos dados em estudo variam de questão de investigação para questão de investigação e de ano para ano, com amostras distintas. A dimensão da amostra é designada por n. As estatísticas descritivas referentes ao estudo do tempo de resposta encontram-se na Tabela II.

Note-se que, para as variáveis binárias, a sua média representa a proporção de casos na amostra. Por exemplo, na Tabela II, a média da variável mulher, para o ano de 2018, é aproximadamente 0,50, ou seja, 50% dos respondentes eram mulheres.

Tabela II - Estatísticas Descritivas das variáveis da questão de investigação Q1

Variável	**Ano**	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
temposegundos	2018 (n=2187)	1404,29	706,45	106,12	6571,53
idade		48,18	14,75	16	74
mulher		0,50	0,50	0	1
basico		0,34	0,48	0	1
secund		0,24	0,43	0	1
superior		0,34	0,47	0	1
nresp		0,03	0,18	0	1
scorresp		0,02	0,13	0	1
amu		0,10	0,30	0	1
apr		0,09	0,29	0	1
respostas		31,35	16,76	1	56
n_insucessos		0,05	0,36	0	9
d_movel		0,13	0,34	0	1
misto		0,02	0,34	0	1

(continua)

Continuação da Tabela II

Variável	Ano	Média	Desvio Padrão	Mínimo	Máximo
temposegundos	2019 (n=1925)	1286,28	615,50	165,59	5466,85
idade		48,47	14,86	16	74
mulher		0,51	0,50	0	1
basico		0,32	0,47	0	1
secund		0,23	0,42	0	1
superior		0,36	0,48	0	1
nresp		0,04	0,20	0	1
scorresp		0,02	0,12	0	1
amu		0,09	0,29	0	1
apr		0,09	0,29	0	1
respostas		30,23	16,04	1	85
n_insucessos		0,33	1,09	0	19
d_movel		0,21	0,41	0	1
misto		0,02	0,14	0	1
temposegundos	2020 (n=1432)	1553,01	909,27	131,26	15025,55
idade		48,39	14,19	16	74
mulher		0,52	0,50	0	1
basico		0,29	0,45	0	1
secund		0,25	0,43	0	1
superior		0,40	0,49	0	1
nresp		0,04	0,20	0	1
scorresp		0,01	0,10	0	1
amu		0,10	0,30	0	1
apr		0,08	0,27	0	1
respostas		71,22	33,90	1	225

n_insucessos		0,45	1,22	0	13
d_movel		0,20	0,40	0	1
misto		0,03	0,17	0	1

A Tabela III sumariza as estatísticas descritivas relativas à adesão ao questionário *web*.

É de destacar que apenas 36%, 32% e 31% dos indivíduos que responderam ao inquérito, acederam ao mesmo via *web*.

Tabela III - Estatísticas Descritivas das variáveis da questão de investigação Q2

Variável	**Ano**	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
tentarcawi	2018 (n=6426)	0,36	0,48	0	1
idade		51,00	15,34	16	74
mulher		0,56	0,50	0	1
basico		0,50	0,50	0	1
secund		0,18	0,39	0	1
superior		0,20	0,40	0	1
nresp		0,01	0,11	0	1
scorresp		0,06	0,23	0	1
amu		0,14	0,35	0	1
apr		0,14	0,35	0	1
tentarcawi	2019 (n=6585)	0,32	0,47	0	1
idade		51,35	14,33	16	74
mulher		0,55	0,50	0	1
basico		0,47	0,50	0	1
secund		0,19	0,39	0	1
superior		0,20	0,40	0	1
nresp		0,01	0,12	0	1
scorresp		0,08	0,27	0	1

amu		0,14	0,35	0	1
apr		0,14	0,35	0	1

(continua)

Continuação da Tabela III

Variável	**Ano**	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
tentarcawi	2020 (n=5016)	0,31	0,46	0	1
idade		52,25	15,31	16	74
mulher		0,57	0,50	0	1
basico		0,45	0,50	0	1
secund		0,20	0,40	0	1
superior		0,22	0,41	0	1
nresp		0,02	0,12	0	1
scorresp		0,08	0,28	0	1
amu		0,14	0,35	0	1
apr		0,13	0,34	0	1

A Tabela IV apresenta as estatísticas descritivas referentes ao abandono do questionário *web*.

Tabela IV - Estatísticas Descritivas das variáveis da questão de investigação Q3

Variável	**Ano**	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
abandonarcawi	2018 (n=2323)	0,06	0,23	0	1
idade		48,18	14,81	16	74
mulher		0,51	0,50	0	1
basico		0,35	0,48	0	1
secund		0,24	0,43	0	1
superior		0,33	0,47	0	1
nresp		0,03	0,17	0	1

scorresp		0,02	0,14	0	1
amu		0,11	0,31	0	1
apr		0,09	0,29	0	1

(continua)

Continuação da Tabela IV

Variável	**Ano**	**Média**	**Desvio Padrão**	**Mínimo**	**Máximo**
abandonarcawi	2019 (n=2100)	0,07	0,26	0	1
idade		48,66	14,95	16	74
mulher		0,51	0,50	0	1
basico		0,33	0,47	0	1
secund		0,23	0,42	0	1
superior		0,35	0,48	0	1
nresp		0,04	0,19	0	1
scorresp		0,02	0,14	0	1
amu		0,10	0,30	0	1
apr		0,09	0,29	0	1
abandonarcawi	2020 (n=1545)	0,06	0,24	0	1
idade		48,51	14,26	16	74
mulher		0,52	0,50	0	1
basico		0,30	0,46	0	1
secund		0,25	0,43	0	1
superior		0,39	0,49	0	1
nresp		0,04	0,19	0	1
scorresp		0,01	0,11	0	1
amu		0,10	0,30	0	1
apr		0,08	0,27	0	1

Como apresentado na Tabela IV, apenas 6%, para 2018 e 2020, e 7%, para 2019, dos indivíduos que acederam ao questionário *web* terminaram a sua resposta através de outro modo.

5. Discussão de Resultados

Nesta secção, apresentam-se os coeficientes estimados para os modelos de tempo total de resposta ao questionário, de probabilidade de usar o questionário *web* e de probabilidade de abandono do questionário *web*. Os valores foram obtidos através da utilização do *software* Stata e aplicando a metodologia apresentada no terceiro capítulo. Em simultâneo, discutem-se os resultados obtidos, comparando-os com os resultados apresentados na literatura de outros autores.

5.1 Relação entre o tempo de resposta e as características do respondente

Pretendeu-se compreender as características dos utilizadores que tentaram aceder ao inquérito via *web*, ainda que possam ter concluído o questionário noutro modo, e compará-las às características dos indivíduos que, tendo respondido noutro modo, nunca acederam ao inquérito *web*. Na Tabela V são apresentados os resultados para o modelo do tempo total de resposta ao questionário.

As regressões dos três anos foram sujeitas ao teste RESET (resultados disponíveis no Anexo 1) e, inicialmente, não passaram o teste. Foi então necessário adicionar termos quadráticos e termos de interação e voltar a efetuar o teste. As regressões que incluem o regressor respostas^2 = (nº de respostas)^2 não tem problemas de especificação da forma funcional.

Relativamente ao impacto da idade no tempo de resposta, para 2018, existe evidência estatística de que mais um ano de idade leva a um tempo total de resposta, em segundos, superior em 0,2%, comparado a outro indivíduo em tudo o resto semelhante, mas um ano mais novo. Em 2019, existe evidência estatística de que um indivíduo um ano mais velho tem um tempo de resposta total em segundos superior em 0,3%, quando comparado com outro indivíduo em tudo o resto semelhante, mas um ano mais novo. Em 2020, mais um ano aumenta o tempo de resposta total em segundos em 0,5%,

comparativamente a um indivíduo em tudo o resto semelhante, mas um ano mais novo. Estas descobertas do efeito positivo da idade no tempo total de resposta estão em concordância com o trabalho de Yan and Tourangeau (2008) e Couper and Kreuter (2013), ainda que estes últimos tenham encontrado um aumento no tempo de respostas individuais, em vez de num inquérito completo.

Tabela V - Coeficientes estimados para o modelo do tempo de resposta

Variável (i)	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2018)**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2019)**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2020)**
constante	6,525 (0,101)	6,546 (0,101)	6,389 (0,190)
idade	0,002** (0,001)	0,003*** (0,001)	0,005*** (0,001)
mulher	0,004 (0, 021)	0,008 (0,021)	-0,009 (0,035)
basico	0,098 (0,087)	0,027 (0,085)	0,111 (0,180)
secund	0,108 (0,090)	-0,039 (0,088)	0,174 (0,183)
superior	0,094 (0,090)	-0,067 (0,088)	0,133 (0,186)
nresp	-0,032 (0,111)	-0,031 (0,107)	0,168 (0,196)
scorresp	0,154 (0,144)	0,051 (0,110)	0,296 (0,236)
amu	-0,065** (0,032)	-0,068* (0,036)	-0,007 (0,063)
apr	0,030 (0,036)	-0,010 (0,034)	0,029 (0,086)
respostas	0,035*** (0,003)	0,030*** (0,002)	0,015*** (0,002)
n_insucessos	0,074*** (0,020)	0,029*** (0,009)	0,013 (0,010)
misto	0,186** (0,091)	0,161** (0,073)	0,192 (0,182)

d_movel	-0,018 (0,026)	-0,092*** (0,027)	-0,084*** (0,032)
respostas^2	-0,0005*** (0,00004)	-0,0004*** (0,00003)	-0,0000009*** (0,000009)

(continua)

Continuação da Tabela V

Variável (i)	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2018)**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2019)**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2020)**
R^2	0,126	0,118	0,089

Nota: os asteriscos ***, **, * indicam significância a 1%, 5% e 10%, respetivamente.

Existe evidência estatística, para 2018, de que um respondente que resida numa Área Medianamente Urbana (AMU) tem um tempo total de resposta, em segundos, menor em 6,29%[3], comparativamente a um respondente em tudo o resto igual, mas que resida numa Área Predominantemente Urbana (APR). Em 2019, um respondente residente numa AMU tem um tempo total de resposta, em segundos, menor em 6,6%, comparativamente a um respondente em tudo o resto igual, mas que resida numa APR. Não existe evidência estatística deste efeito para a amostra de 2020. Este efeito vai contra a intuição de que quanto mais rural é o local de residência, maiores podem ser as dificuldades de utilização do computador e/ou da *internet*, conduzindo a maiores tempos de resposta. Pode no entanto sugerir que nas áreas citadinas, o ritmo de vida impele à resposta mais faseada. Não existe evidência estatística que revele qual o efeito associado às Áreas Predominantemente Rurais.

Existe evidência estatística de que um aumento em uma resposta dada, para o ano de 2018, aumenta em média, aproximadamente, 3,4%[4] o tempo total de resposta ao questionário, quando comparado a um indivíduo em tudo o resto igual, mas que respondeu a menos uma pergunta. Este efeito é de 2,92%, em 2019, e de 1,5%, em 2020.

[3] Para $\widehat{\beta_k}$ grandes, calcula-se a variação percentual da variável dependente y com a seguinte fórmula: $\%\Delta\hat{y} = 100 \times [\exp(\widehat{\beta_k}\Delta x_k) - 1]$ (Wooldridge, 2016)

[4] $\Delta\widehat{tempo segundos} \approx (\hat{\beta}_{respostas} + 2\hat{\beta}_{respostas^2} \cdot respostas)\Delta respostas$ (Wooldridge, 2016)

Existe evidência estatística de que um respondente com mais um insucesso no *login* tem um tempo total de resposta, em segundos, superior em 7,68%, para 2018 e em 2,94%, em 2019, quando comparado a um outro respondente em tudo o resto semelhante, mas com menos um insucesso no login. O efeito não é estatisticamente significativo na regressão referente a 2020. Este impacto positivo no tempo total de resposta pode ser explicado pelo facto das falhas no login atrasarem o processo de resposta. Podem também ser um indicador de que o respondente não domina a utilização de computadores, sendo mais demorado todo o processo de resposta.

Existe também evidência estatística de que um respondente que utilize, na sua resposta, tanto o dispositivo móvel como o computador (misto) tem um aumento de 20,44% no tempo de resposta, em 2018, e de 17,47%, em 2019, comparado a um indivíduo em tudo o resto igual, mas que utilizou apenas o computador. De entre os vários utilizadores que utilizaram computador e dispositivo móvel, a maioria utilizou primeiramente dispositivo móvel e só depois computador. Uma possível explicação para o acréscimo no tempo de resposta nos utilizadores que usaram os dois tipos de dispositivo pode ser algum tipo de dificuldade encontrada em responder num dispositivo móvel, atrasando o processo de resposta e conduzindo à mudança para computador.

Um respondente que utilize apenas o dispositivo móvel para a resposta, tem, segundo a evidência estatística, tempo total de resposta menor em 8,79%, para 2019, e em 8,06%, para 2020, quando comparado a um respondente em tudo o resto semelhante, mas que utilizou apenas o computador.

Não foi encontrada evidência estatística de que o nível de escolaridade completa ou o sexo do respondente impactem o tempo total de resposta ao questionário.

5.2 Relação entre a tentativa de resposta em CAWI e as características do respondente

Pretendeu-se compreender as características dos utilizadores que tentaram aceder ao inquérito via *web*, ainda que possam ter concluído o questionário noutro modo, e compará-las às características dos indivíduos que, apesar de terem respondido noutro modo, nunca acederam ao inquérito *web*.

Na Tabela VI são apresentados os coeficientes estimados para a regressão *tentarcawi*, nos anos de 2018, 2019 e 2020 e, entre parêntesis, os respetivos desvios-padrão. Também são apresentados na Tabela VI os efeitos parciais médios.

Tabela VI - Coeficientes estimados para o modelo *tentarcawi*

Variável(i)	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2018)**	**EPM**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2019)**	**EPM**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2020)**	**EPM**
constante	-0,203 (0,113)		-0,789 (0,117)		-1,438 (0,207)	
idade	-0,002 (0,001)	-0,001	0,001 (0,001)	0,0002	0,040*** (0,008)	0,012
mulher	-0,229*** (0,034)	-0,076	-0,181*** (0,034)	-0,057	-0,211*** (0,040)	-0,064
basico	0,176** (0,085)	0,058	0,191** (0,091)	0,060	0,118 (0,120)	0,036
secund	0,709*** (0,092)	0,235	0,677*** (0,097)	0,214	0,580*** (0,126)	0,175
superior	1,101*** (0,091)	0,365	1,109*** (0,096)	0,350	1,010*** (0,125)	0,305
nresp	1,969*** (0,198)	0,654	2,001*** (0,184)	0,632	1,641*** (0,199)	0,496
scorresp	-0,277** (0,117)	-0,092	-0,482*** (0,117)	-0,152	-0,638*** (0,156)	-0,193
amu	-0,252*** (0,050)	-0,084	-0,251*** (0,051)	-0,079	-0,229*** (0,060)	-0,069
apr	-0,320*** (0,052)	-0,106	-0,308*** (0,053)	-0,097	-0,334*** (0,064)	-0,101
idade^2	-		-		-0,0005*** (0,00008)	-0,0001
Pseudo R^2	0,108		0,115		0,135	

Nota: os asteriscos ***, **, * indicam significância a 1%, 5% e 10%, respetivamente.

Encontrou-se evidência estatística, para 2020, de que um indivíduo um ano mais velho tem um decréscimo médio na probabilidade de aceder ao questionário *web* em de 0,2[5] pontos percentuais, em comparação com um indivíduo em tudo o resto semelhante, mas um ano mais novo. Este efeito parece confirmar a crença generalizada de que os mais velhos evitam a utilização das novas tecnologias e optam por meios mais familiares de resposta.

Existe evidência estatística de que uma mulher tem, para 2018, um decréscimo médio de 7,6 pontos percentuais de probabilidade de tentar responder na *internet*, comparada a um outro respondente, em tudo o resto semelhante, mas homem. Em 2019, este decréscimo é de 5,7 pontos percentuais e em 2020 de 6,4 pontos percentuais.

Um indivíduo com o ensino básico como nível máximo de escolaridade completa tem um acréscimo médio de 5,8 pontos percentuais na probabilidade de aceder ao questionário *web*, comparativamente a um indivíduo em tudo o resto semelhante, mas que não tem nenhum nível de escolaridade completo. O acréscimo médio é de 6 pontos percentuais para 2019. O efeito não é estatisticamente significativo para a amostra de 2020.

Um indivíduo com o ensino secundário como nível máximo de escolaridade completa tem um acréscimo médio de 23,5 pontos percentuais na probabilidade de aceder ao questionário *web*, comparativamente a um indivíduo em tudo o resto semelhante, mas que não tem nenhum nível de escolaridade completo. O acréscimo médio é de 21,4 pontos percentuais para 2019 e de 17,5 pontos percentuais em 2020.

Um indivíduo com o ensino superior como nível máximo de escolaridade completa tem um acréscimo médio de 36,5 pontos percentuais na probabilidade de aceder ao questionário *web*, comparativamente a um indivíduo em tudo o resto semelhante, mas que não tem nenhum nível de escolaridade completo. O acréscimo médio é de 35 pontos percentuais para 2019 e de 30,5 pontos percentuais em 2020.

O aumento consistente da probabilidade de acesso ao questionário *web* associado ao aumento do nível máximo de escolaridade completa está em concordância com as descobertas feitas por Kelfve et al. (2020), que descobriu que pessoas com menores níveis de escolaridade têm menos probabilidade de aceder a um questionário na *internet*.

[5] $\Delta \widehat{tentarcawi} = [\sum_{i=1}^{n} \phi(\hat{\beta}_0 + \hat{\beta}_{idade} \cdot idade_i + \hat{\beta}_{mulher} \cdot mulher_i + \cdots + \hat{\beta}_{idade^2} \cdot idade_i^2) \times (\hat{\beta}_{idade} + 2\hat{\beta}_{idade^2} \cdot idade_i)] \div n$

Um indivíduo que tenha optado não responder sobre o seu nível máximo de escolaridade completa tem um acréscimo médio de 65,4 pontos percentuais na probabilidade de aceder ao questionário *web*, comparativamente a um indivíduo em tudo o resto semelhante, mas que não tem nenhum nível de escolaridade completo. O acréscimo médio é de 63,2 pontos percentuais para 2019 e de 49,6 pontos percentuais em 2020.

Um indivíduo que declare que o seu nível máximo de escolaridade completa não tem correspondência com nenhum dos atuais tem um decréscimo médio de 9,2 pontos percentuais na probabilidade de aceder ao questionário *web*, comparativamente a um indivíduo em tudo o resto semelhante, mas que não tem nenhum nível de escolaridade completo. O decréscimo médio é de 15,2 pontos percentuais para 2019 e de 19,3 pontos percentuais em 2020.

Para 2018, encontrou-se evidência estatística de que um residente numa Área Medianamente Urbana (AMU) tem um decréscimo médio de probabilidade de 8,4 pontos percentuais de aceder ao questionário *online*, comparativamente a um respondente em tudo o resto semelhante, mas que seja residente numa Área Predominantemente Urbana. Este decréscimo médio de probabilidade é de 7,9 pontos percentuais, em 2019, e de 6,9 pontos percentuais, em 2020.

Um indivíduo que resida numa Área Predominantemente Rural (APR) tem, segundo a evidência estatística, um decréscimo médio de probabilidade de aceder ao questionário web de 10,6 pontos percentuais, para 2018, de 9,7 pontos percentuais, para 2019 e de 10,1 pontos percentuais, para 2020, em comparação com um indivíduo semelhante, mas que resida numa Área Predominantemente Urbana.

Os efeitos acima mencionados, relativos à área de residência do respondente, estão de acordo com a perceção geral de que os indivíduos que habitam em zonas mais rurais têm maiores dificuldades de acesso a tecnologia e que, por essa razão, podem evitar responder pela *internet* ou podem mesmo não ter acesso à mesma.

5.3 Relação entre o abandono do questionário em CAWI e as características do respondente

Pretendeu-se estudar os respondentes que começaram o questionário na *web*, mas terminaram noutro modo, ou seja, os respondentes que têm paradados associados, mas que têm modo de apuramento final CATI ou CAPI, e compará-los aos que começaram e concluíram no modo CAWI. Na Tabela VII são apresentados os coeficientes estimados para as regressões relativas ao *abandonarcawi* em 2018, 2019 e 2020 e, entre parêntesis, os respetivos desvios-padrão. Também são apresentados na Tabela VII os efeitos parciais médios.

Tabela VII - Coeficientes estimados para o modelo *abandonarcawi*

Variável(i)	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2018)**	**EPM**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2019)**	**EPM**	$\widehat{\boldsymbol{\beta}}_{\mathbf{i}}$ **(2020)**	**EPM**
constante	-1,818 (0,380)		-1,676 (0,347)		-1,959 (0,522)	
idade	-0,005 (0,003)	-0,001	-0,00007	-0,00001	0,001	0,000
mulher	0,354*** (0,089)	0,039	0,166**	0,023	0,132	0,016
basico	0,402 (0,327)	0,045	0,276	0,039	0,407	0,049
secund	0,072 (0,340)	0,008	0,110	0,015	0,369	0,044
superior	0,091 (0,333)	0,010	-0,095	-0,013	0,0367	0,004
nresp	-0,357 (0,497)	-0,040	Omitido	-	Omitido	-
scorresp	1,145*** (0,378)	0,127	1,043***	0,146	1,246**	0,150
amu	0,263** (0,126)	0,029	0,272**	0,038	0,137	0,017
apr	-0,114 (0,161)	-0,013	-0,069	-0,010	0,039	0,005
Pseudo R^2	0,051		0,039		0,033	

Nota: os asteriscos ***, **, * indicam significância a 1%, 5% e 10%, respetivamente.

As regressões dos três anos foram sujeitas ao teste RESET (resultados disponíveis no Anexo 3) e passaram o teste, não tendo então problemas de especificação.

O regressor *nresp* foi automaticamente excluído da análise de 2019 e 2020 por problemas de colinearidade perfeita, pois na amostra estudada não está presente nenhum indivíduo que tenha escolhido não responder (*nresp*) sobre o seu nível máximo de escolaridade completo e simultaneamente abandonado o questionário.

Encontrou-se evidência estatística de que, para a amostra de 2018, uma mulher tem um aumento médio da probabilidade de abandonar o modo CAWI de 3,9 pontos percentuais, comparada com um respondente, em tudo o resto semelhante, mas homem. Para 2019, esse efeito é de 2,3 pontos percentuais. Não foi encontrado qualquer efeito relacionado com o sexo do respondente estatisticamente significativo para a regressão de 2020.

Para 2018, encontrou-se evidência estatística de que um indivíduo que declare que o seu nível de escolaridade não tem correspondência aos atuais, tem em média mais 12,7 pontos percentuais de probabilidade de abandonar o modo CAWI, comparativamente a um indivíduo em tudo o resto semelhante, mas sem qualquer nível de escolaridade completa. A evidência estatística aponta para um aumento médio dessa mesma probabilidade em 14,6 pontos percentuais, em 2019, e em 15 pontos percentuais, em 2020.

Um indivíduo residente numa Área Medianamente Urbana (AMU) tem, segundo a evidência estatística, em 2018, 2,9 pontos percentuais de acréscimo médio na probabilidade de abandonar o modo CAWI, comparativamente a um indivíduo em tudo o resto semelhante, mas que reside numa área predominantemente urbana. Em 2019, esse efeito é negativo, ou seja, um residente numa AMU tem, segundo a evidência estatística, 1 ponto percentual de decréscimo médio, na probabilidade de abandonar o modo CAWI, comparativamente a um indivíduo em tudo o resto semelhante, mas que reside numa área predominantemente urbana.

Não foram encontrados efeitos estatisticamente significativos relacionados com a escolaridade completa nem com a idade do respondente.

O estudo do *abandonarcawi* teve como grande limitação o baixo número de respondentes que após terem acedido ao questionário *web* o completaram por outro modo. Em 2018 e 2020, estavam nesta situação apenas 6% dos respondentes e em 2019, 7% dos respondentes. Na tentativa de colmatar essa limitação, foram estimados modelos de variável dependente binária cloglog, cuja utilização é aconselhada para distribuições de

dados assimétricas, com muitos 0 ou muitos 1 (Cameron e Trivedi, 2009), como é o presente caso. Os resultados desta estimação não resultaram em grandes melhorias na significância dos coeficientes e conduziram a resultados no geral parecidos com os apresentados nesta secção, logo optou-se por não apresentar as estimativas dos cloglog.

Tendo em vista a melhoria do processo de recolha de resposta, os resultados obtidos parecem sugerir, tal como na questão de investigação Q2, que a disponibilização de vários modos de resposta, ajustados sempre que possível ao perfil do respondente, é essencial para obter o maior número de respostas possível.

6. Conclusão

A crescente utilização de inquéritos realizados na *internet* torna imprescindível compreender o comportamento do respondente na sua interação com o mesmo. A crescente popularidade da recolha e utilização dos paradados pode ser aproveitada para um conhecimento mais aprofundado dessa interação. Adicionalmente, os dados auxiliares referentes às características sociodemográficas dos respondentes podem complementar essa análise de forma relevante.

O presente trabalho tem como objetivo melhorar a compreensão das interações dos utilizadores com o questionário *web*, tendo por base os dados recolhidos no IUTICF. Nesse sentido, foi possível encontrar evidência estatística de que o tempo de resposta ao questionário depende positivamente da idade do respondente, do número de insucessos no *login* e da utilização mista de dispositivo móvel e computador, pelo mesmo utilizador, para responder ao questionário. Em sentido oposto, a resposta através do dispositivo móvel diminui o tempo total de resposta. Em dois dos três anos em análise, os indivíduos residentes em áreas medianamente urbanas demoram menos tempo a responder, comparativamente a indivíduos semelhantes, mas residentes em áreas predominantemente urbanas. Não foi encontrada evidência estatística de que o nível de escolaridade completa e o sexo do respondente influenciem o tempo de resposta.

Esta informação pode contribuir para a otimização do processo de recolha de resposta de várias formas. Em primeira instância, pode-se tentar reduzir o número de falhas no *login*. Estas falhas podem dever-se à dificuldade de compreensão da comunicação inicial (carta ou email), sendo desejável testar os documentos e avaliar as possibilidades de

melhoria que incrementem a compreensão e legibilidade da carta. Outra origem poderá estar relacionada com a transição de modos, que origina tentativas de *login* em alojamentos afetos a outro modo de resposta. Neste campo, também se sugere que as mensagens que surgem ao respondente possam ser explicitas para evitar *logins* sucessivos. Relativamente à utilização mista de dispositivos na resposta, o efeito encontrado de acréscimo no tempo de resposta deve ser estudado de forma mais aprofundada, mas pode ter por base dificuldades sentidas na resposta num dispositivo móvel e posterior mudança para a resposta em computador. Nesse sentido, o INE pode aperfeiçoar a compatibilidade dos seus questionários com os dispositivos móveis, para que os utilizadores que lá começam o questionário não sintam necessidade de mudar para um computador. O efeito negativo da utilização do dispositivo móvel no tempo total de resposta reforça a importância do ponto anterior.

Foi também possível encontrar evidência de que, na altura de aceder ao questionário na *internet*, os homens têm maior probabilidade de o fazer. Quanto maior o nível de escolaridade completa do indivíduo, maior probabilidade tem de aceder ao questionário *web* e, em oposição, quanto mais rural for a sua área de residência, menor probabilidade tem de responder nesse modo. Assim, recomenda-se que o INE continue a sua estratégia de disponibilização de vários modos de resposta (CAPI, CATI, CAWI) e de estratégias de segmentação da amostra, de forma a alcançar a diversidade de selecionados, que pode assim responder de forma mais cómoda. Um aprofundamento da análise da variável idade, por exemplo, utilizando escalões é aconselhado para compreender melhor os seus efeitos no acesso ao questionário *online*.

Por fim, relativamente ao abandono do questionário *online*, o estudo foi limitado pela escassez de dados. No entanto, foi possível encontrar efeitos estatisticamente significativos relacionados com o sexo do respondente (as mulheres têm maior probabilidade de abandonar o questionário *online*) e com a área de residência, nomeadamente de que um indivíduo residente numa área medianamente urbana tem maior probabilidade de abandono, comparativamente com um indivíduo semelhante, mas residente numa área predominantemente urbana. Mais uma vez, é de destacar a importância de possibilitar os três modos de resposta alternativos, se o objetivo do INE é obter o maior número de respostas possível. As descobertas feitas suportam também o

objetivo de redução dos custos associados ao inquérito, pois pretendem promover um acesso mais generalizado e maior satisfação com o inquérito *web*.

É necessário referir que, apesar do INE escolher a sua amostra de forma aleatória, as transformações e exclusões de observações feitas tornam as amostras utilizadas neste estudo não aleatórias. Esta é uma das limitações que podem ser resolvidas num futuro trabalho.

Futuramente, seria relevante aplicar este tipo de estudo a outros inquéritos do INE, com o objetivo de compreender se os efeitos encontrados são transversais ou se estão relacionados com o tema do questionário. Mais, o presente estudo poderia ser alargado de forma a incluir a análise de mais paradados (e.g. tamanho do ecrã, número de alterações da resposta). Finalmente, o método de seleção da amostra do IUTICF permitiria também o estudo de dados de painel, com o prejuízo de que só um quarto da amostra se mantém durante os quatro anos consecutivos.

Bibliografia

Bosch, Oriol J., Melanie Revilla, and Ezequiel Paura. 2019. "Do Millennials Differ in Terms of Survey Participation?" *International Journal of Market Research* 61(4):359–365.

Callegaro, Mario. 2010. "Do You Know Which Device Your Respondent Has Used to Take Your Online Survey?" *Survey Practice* 3(6).

Callegaro, Mario. 2013. "Paradata in Web Surveys." 259–279 in *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter. John Wiley & Sons, Inc.

Callegaro, Mario, Katja Lozar Manfreda, and Vasja Vehovar. 2015. *Web Survey Methodology*. 1ª Ed. Londres: SAGE.

Cameron, Adrian Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata*. 1ª Ed. Texas: Stata Press.

Couper, Mick P. 1998. "Measuring Survey Quality in a CASIC Environment." *Proceedings of the Survey Research Methods Section* 41–49.

Couper, Mick P., and Frauke Kreuter. 2013. "Using Paradata to Explore Item Level Response Times in Surveys." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 176(1):271–286.

Daikeler, Jessica, Michael Bosnjak, and Katja Lozar Manfreda. 2020. "Web versus Other Survey Modes: An Updated and Extended Meta-Analysis Comparing Response Rates." *Journal of Survey Statistics and Methodology* 8(3):513–539.

Instituto Nacional de Estatística. 2019. "Documento Metodológico Do Inquérito à Utilização de Tecnologias Da Informação e Da Comunicação Das Famílias." *Instituto Nacional de Estatística*.

Kelfve, Susanne, Marie Kivi, Boo Johansson, and Magnus Lindwall. 2020. "Going Web or Staying Paper? The Use of Web-Surveys among Older People." *BMC Medical Research Methodology* 20(1).

Kreuter, Frauke. 2013. "Improving Surveys with Paradata: Introduction." 1–8 in *Improving Surveys with Paradata: Analytic Uses of Process Information*, edited by F. Kreuter. John Wiley & Sons, Inc.

Kreuter, Frauke. 2018. "Getting the Most out of Paradata." 193–198 in *The Palgrave Handbook of Survey Research*. Springer International Publishing.

Lancaster, T. 1990. *The econometric analysis of transition data* (No. 17). Cambridge: Cambridge University Press.

Lynn, Peter, and Gerry Nicolaas. 2010. "Making Good Use of Survey Paradata." *Survey Practice* 3(2).

Manfreda, Katja Lozar, Michael Bosnjak, Jernej Berzelak, Iris Haas, and Vasja Vehovar. 2008. "Web Surveys versus Other Survey Modes A Meta-Analysis Comparing Response Rates." *International Journal of Market Research* 50(1):79–103.

Matjašič, Miha, Vasja Vehovar, and Katja Lozar Manfreda. 2018. "Web Survey Paradata on Response Time Outliers: A Systematic Literature Review." *Metodološki Zvezki* 15(1):23–41.

Santos Silva, J. M. C., and Silvana Tenreyro. 2006. "The Log of Gravity." *Review of Economics and Statistics* 88(4):641–658.

Shi, Yi, Jun Feng, and Xiaoqin Luo. 2018. "Improving Surveys with Paradata: Analytic Uses of Response Time." *China Population and Development Studies* 2(2):204–223.

Strauss, William, and Neil Howe. 1991. *Generations: The History of America's Future, 1584 to 2069*. 1ª Ed. Nova Iorque: William Morrow and Company Inc.

West, Brady T. 2011. "Paradata in Survey Research." *Survey Practice* 4(4).

Wooldridge, Jeff M. 2016. *Introductory Econometrics: A Modern Approach*. 6ª Ed. Cengage Learning.

Yan, Ting, and Roger Tourangeau. 2008. "Fast Times and Easy Questions: The Effects of Age, Experience and Question Complexity on Web Survey Response Times." *Applied Cognitive Psychology* 22(1):51–68.

ANEXOS

Anexo 1 – Resultados dos testes RESET às regressões do tempo total de resposta

	Teste de Significância de $\hat{y}^2$	Modelo Original	Modelo com regressor adicional *respostas²*
2018	Prob > Qui²	0,000	0,692
2019	Prob > Qui²	0,000	0,248
2020	Prob > Qui²	0,007	0,179

Nota: foi excluído do teste o regressor $\hat{y}^3$ por não convergir.

Anexo 2 – Resultados dos testes RESET às regressões *tentarcawi*

	Teste de Significância Conjunta de $\hat{y}^2$ e $\hat{y}^3$	Modelo Original	Modelo com regressor adicional *idade²*
2018	Prob > Qui²	0,241	-
2019	Prob > Qui²	0,747	-
2020	Prob > Qui²	0,010	0,260

Anexo 3 – Resultados dos testes RESET às regressões *abandonarcawi*

	Teste de Significância Conjunta de $\hat{y}^2$ e $\hat{y}^3$	Modelo Original
2018	Prob > Qui²	0,672
2019	Prob > Qui²	0,951
2020	Prob > Qui²	0,400