GUSTAVO PIMENTEL BAAMONDE, RICARDO SEMIÃO E CASTRO

Analisando e prevendo o salário na Região Metropolitana de São Paulo

São Paulo

2020

GUSTAVO PIMENTEL BAAMONDE, RICARDO SEMIÃO E CASTRO

Analisando e prevendo o salário na Região Metropolitana de São Paulo

Trabalho realizado na Escola de Economia de Sâo Paulo - FGV para a obtenção de 7 créditos na matéria de Metodologia e Técnicas de Pesquisa - Projeto II

Fundação Getúlio Vargas

Escola de Economia de São Paulo - EESP

Programa de Graduação

Orientador: Emerson Marçal

São Paulo

2020

GUSTAVO PIMENTEL BAAMONDE, RICARDO SEMIÃO E CASTRO

Analisando e prevendo o salário na Região Metropolitana de São Paulo

Trabalho realizado na Escola de Economia de Sâo Paulo - FGV para a obtenção de 7 créditos na matéria de Metodologia e Técnicas de Pesquisa - Projeto II

Emerson Marçal
Orientador

São Paulo
2020

Dedico esse trabalho à Rafael Sanzio e Platão

Agradecimentos

Agradecemos ao nosso professor Vinicíus Lima e ao professor líder da matéria Emerson Marçal.

"My greatest strength as a consultant
is to be ignorant and ask a few questions"
(Peter Drucker)

Resumo

Esse estudo tem 2 objetivos principais. Primeiramente queremos entender qual é o efeito de diferentes variáveis no salário, estudando desde deficiências até raça. Há uma ampla literatura sobre isso, mas geralmente se analisa cada variável separadamente. Aqui. buscamos fazer uma pesquisa mais geral para ver se esses resultados ainda valem. O segundo objetivo foi criar um modelo para efetuar uma previsão salarial. Para fazer tal estudo, usamos os dados disponível do Censo de 2010 do IBGE na Região Metropolitana de São Paulo. Os resultados da primeira parte corroboram com os estudos feitos anteriormente, enquanto os da segunda mostram que para a previsão, modelos mais parcimoniosos são interessantes, especificamente obtivemos melhores resultados com um modelo *elastic-net*, sendo as variáveis educação e raça as que possuem maior coeficiente.

Palavras-chaves: Salário, previsão, FGV

Abstract

This study has 2 main objectives. First, we want to understand what's the effect of different variables on wage, analysing everything from disabilities to race. There exist a wide literature to talk about it, but they mainly analyse each variable separately. Here, we seek to do a more broad research to check it those results are still valid in a larger model. The second objective is to create a model to predict wage based on people's characteristics. To do this research, we used the data from the 2010 Brazilian Census of the Metropolitan Region of Sao Paulo. The results show that the first part corroborates with the existing literature, while the second show that the best model for prediction is obtained using an elastic net regularization, being race and education, the variables with the higher coefficients.

Key-words: Wage, prediction, FGV

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Quando falamos em uma sociedade capitalista, a palavra "dinheiro"provavelmente vêm à mente. Ele move a economia e as coisas em nossa volta. Ele financia a construção de obras públicas e privadas, possibilita as pessoas visitarem outros países e, acima de tudo, coloca comida em cima da mesa. Dessa forma, é *sine qua non* para as pessoas, pois influencia elas em sua tomada de decisão, como por exemplo, a escolha da profissão no futuro. Assim, entender quais e como as características de cada pessoa como idade, anos de educação, sexo, possuir deficiência, etc, afetam o salário é fundamental, especialmente para o longo prazo. Em outras palavras, saber qual será o salário e quais fatores o afetam é importante.

Com isso mente, fica claro que fazer uma análise sobre o salário é fundamental para conseguir entende-lo melhor. Entretanto, a maioria das pesquisas sobre esse tema buscam usa-lo para entender outros tópicos, como por exemplo, a relação entre discriminação de gênero no mercado de trabalho. São poucas as pesquisas que buscam analisa-lo como um todo.

Dessa forma, seguindo uma linha de pesquisa apresentada, temos dois objetivos com esse trabalho. Usando os dados disponibilizados pelo Censo de 2010 do IBGE na Região Metropolitana de São Paulo e do *software* R, iremos fazer uma análise das características impactam no salário. Ademais, iremos realizar uma previsão da remuneração esperada de uma pessoa dada essas características. Os modelos e hipóteses a serem usadas nesse processo serão discutidas mais adiante. Aqui, é importante salientar que a base de dados escolhida é ideal para efetuar nossa pesquisa, pois possui um grande número de observações e as variáveis necessárias para realizar a análise desejada.

O trabalho será dividido da seguinte maneira: na próxima seção faremos uma breve revisão bibliográfica, em seguida explicaremos a metodologia utilizada para realizar esta pesquisa. Na seção 4, discutiremos os dados e a escolha das variáveis. A seção 5 compreende os resultados das regressões e suas interpretações. Na conclusão, finalizamos o trabalho fazendo algumas considerações finais.

2 Revisão Bibliográfica

O tema salário não é novo na literatura, há vários trabalhos que querem analisa-lo. Entretanto, eles focam, em sua grande maioria, em um coeficiente especifico, ou seja, como determinada variável, seja ela escolaridade, idade, sexo ou deficiência (BAAMONDE; MARTINS, 2020), afeta o salário.

Há uma extensa literatura que busca analisar, em especial, o chamado *gender wage gap* , ou seja, um possível diferença salarial entre homens e mulheres, que varia desde uma análise brasileira (MADALOZZO, 2010) até uma comparação entre países (APPLETON; HODDINOTT; KRISHNAN, 1999). Elas mostram que mulheres, ainda que em menor grau que antigamente, ainda ganham proporcionalmente menos que homens no mercado de trabalho.

Pesquisas relacionadas ao *wage gap* não se limitam somente ao gênero. O artigo "Deficiência, Emprego e Salário no Mercado de Trabalho Brasileiro"da (BECKER, 2019), por exemplo, buscou estudar o efeito das deficiências no salário hora de trabalho no Brasil. Os resultados indicam que portar deficiência tem um impacto negativo no salário hora. Ou o artigo "O diferencial de salários formal-informal no Brasil: segmentação ou viés de seleção?"de (FILHO; MENDES; ALMEIDA, 2004) que analisou a diferença de salários entre o setor formal e informal brasileiro, mostrando que maiores salário no setor formal estão mais ligados a características não observáveis dos empregados do que do setor em si.

Já o tema de previsão salarial também não é novo na literatura, ainda que de forma reduzida quando comparada aos temas acima. Pierce (1999) buscou, através dos dados disponibilizados pelo National Compensation Survey (NCS), prever o nível de salário nos Estados Unidos. O autor separou a amostra em grupos e profissões para conseguir fazer uma melhor análise. Entretanto esse trabalho já está desatualizado, dado que foi realizado no final do século passado.

3 Metodologia

Neste trabalho, temos como objetivo fazer duas análises. A primeira consiste em fazer uma regressão múltipla e analisar quais e como as diferentes características das pessoas afetam seu rendimento, sendo isso medido através do valor do coeficiente. A segunda é a de fazer uma previsão do salário futuro de uma pessoa dada as questões como possuir deficiência ,educação, região de moradia, idade, etc. Sendo assim, buscamos nessa seção mostrar como pretendemos fazer essa análise, mostrando quais foram os métodos empregados e o porque de usarmos eles.

Antes de tudo, é de suma importância explicar que essa pesquisa teve uma abordagem mista, ainda que a sua maioria foi quantitativa, isso porquê, em sua grande extensão, estamos trabalhando com dados e os resultados numéricos. Entretanto, a abordagem qualitativa entra em cena quando escolhemos o tipo do modelo a ser usado, as variáveis a serem incluídas, assim como a interpretação dos resultados e as conclusões sobre ele. Dessa forma, podemos fazer uma análise mais objetiva e completa do problema aqui proposto.

Ademais, os dados do Censo de 2010 do IBGE da Região Metropolitana de São Paulo foram tratados no programa estatístico R. Ele foi utilizado considerando seu fácil manuseio e a disponibilidade de *packages* que nos ajudam a extrair os os dados necessários.

Escolhemos trabalhar com um região especifica do Brasil pelo simples fato de que não possuímos um servidor para conseguir trabalhar com a base de dados completa de todo país, uma possível expansão do trabalho seria aplicar o mesmo modelo para o país todo, na tentativa de checar se os resultados valem a nível nacional. Ademais, uma análise de um determinado local nos possibilita excluir qualquer tipo de diferença entre os Estados, por exemplo. A ideia aqui é que como diferentes regiões possuem economias muitas vezes distintas, o que pode afetar a conclusão de nosso estudo. Ainda seria interessante, porém, controlar por cidade, algo que não pudemos fazer.

Como dito anteriormente, primeiramente iremos analisar os efeitos de diferentes variáveis no salário. A regressão a ser efetuada será do tipo:

$$Ln(salário) = \beta_0 + \sum_{i=1}^{p} x_i \beta_i \tag{3.1}$$

Onde o nome de cada preditor x_i se encontra no apêndice A. O diferencial aqui é iremos adicionando um grupo de variáveis de cada vez a fim de compreender o efeito que a adição dela trás ao modelo. No final faremos uma análise completa com a equação final. Ademais, teremos um seção mais adiante para explicar o porquê que cada variável foi incluída e as hipóteses por trás desse modelo.

Além disso, como encontramos problemas de heterocedasticidade, fizemos a correção dos erros padrões dos estimadores através de erros robustos de White. Todos os resultados apresentados passaram por essa correção.

O nosso segundo objetivo é fazer uma previsão do salário, assim devemos primeiramente escolher o modelo a fim de conseguir ter a melhor previsão. Para tal, vamos utilizar diferentes métodos e comparar seus resultado. Como sempre, teremos um grupo de teste e outro de treinamento, sendo o método que nos dê o modelo com o menor Erro Quadrático Médio (EQM), o escolhido. Em nossa análise, escolhemos uma quebra de 90% da base para treino e 10% para teste, sendo que uma possível expansão de nosso trabalho seria aplicar os mesmos algorítimos mas para outras quebras, para checar se nossos resultados estão consistentes.

Nessa parte, utilizamos os seguintes algoritmos: regressão *ridge*, regressão *lasso*, e a combinação dos dois,*elastic net regularization*, também usamos o *backward selection*, *forward selection* e o *stepwise selection* (ou *sequential replacement*). A definição formal dos algorítimos consta no apêndice C. Optamos por escolher vários e depois compará-los, pois assim temos uma maior certeza que estamos escolhendo o modelo correto. Aqui é importante salientar que poderíamos ter usado o *best subset selection*, mas dado o grande número de variáveis com que estamos trabalhando e a baixa velocidade decidimos obter por fazer de varias formas a fim de chegar no melhor resultado.

4 Dados

A base escolhida é interessante pelo seu grande número de observações, que nos ajudam a ter resultados mais fidedignos com a realidade, e a grande disponibilidade de variáveis, contendo as principais características sociais e econômicas que julgamos necessárias. O processo de amostragem também tenta sem o mais livre de viés possível, o que será importante para garantir uma amostra i.i.d.

Para ambos os objetivos do trabalho, é interessante ter uma justificativa *apriori* para a inclusão das variáveis, isso será feito a seguir. Vale ressaltar que colocaremos, em parênteses, o código da variável segundo a classificação do dicionário do IBGE. Uma motivação geral para todas as variáveis categóricas incluídas são as diferenças nas médias condicionais apresentadas na tabela 3 no apêndice B, gostaríamos de fazer uma análise mais formal e ver se essas características realmente tem um efeito *ceteris paribus* diferente de 0 sobre o salário.

Primeiramente, temos que definir qual variável explicada em nosso modelo. Aqui, trabalhamos com a variável de rendimento no trabalho principal (código V6513). Escolhemos ela com o pensamento de que outras variáveis de rendimento, tal como rendimento familiar ou recebimento de transferências, como bolsa família, não ajudam explicar tanto o salário, ao menos em nível individual.

Partindo agora para as variáveis explicativas, iniciaremos com educação (código V6400). Foram criadas *dummies* para cada categoria de nível educacional. Ela se divide em 5 categorias: 1.Sem instrução e fundamental incompleto, 2. Fundamental completo e médio incompleto, 3. Médio completo e superior incompleto, 4. Superior completo e 5.Não determinado. Onde omitimos a categoria 1 para fins de evitar a multicolinearidade. Ela foi adicionada ao modelo pois há uma ampla literatura que afirma que quanto mais anos de estudo, maior será o salário desse indivíduo.

Em seguida, temos a variável raça(código V0606), também composta de diversas *dummies*, onde: 1.Branco, 2.Preta, 3.Amarela, 4.Parda, 5.Indígena e 9.ignorado, sendo a dummy de brancos a que foi omitida. Novamente, temos uma ampla literatura que discute diversas de ganhos entre diferentes raças (HECKMAN; LYONS; TODD, 2000).

Outro preditor que também se encontra no modelo é a idade (código V6036). Colocamos tanto ela como ela ao quadrado. Fizemos isso pois com base na figura 4, percebe-se esse tipo de relação. Ademais, adicionamos variáveis sobre o estado civil, são eles: 1.Casado, 2.Separado, 3.Divorciado, 4.Viúvo, 5.Solteiro. Há um extenso debate sobre as diferenças salárias entre estados civis, especialmente entre casados e não casados (CHUN; LEE, 2001), assim achamos importante adiciona-lá.

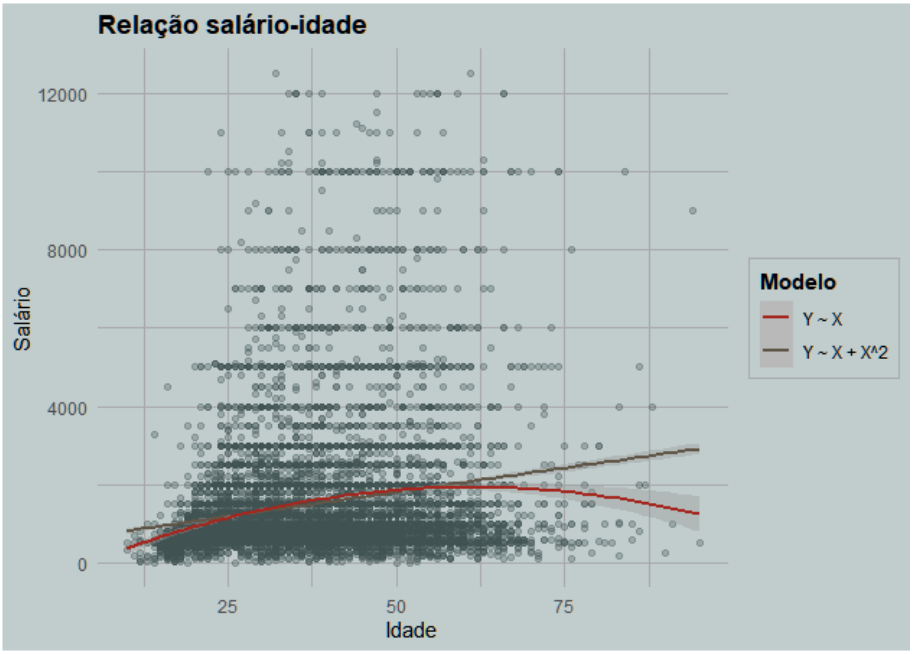


Figura 1 – Relação Salário-Idade

Adicionamos também variáveis que dizem respeito ao número de filhos (código V6633) e pessoas na família (código V5060). Isso porquê existe uma relação entre essas informações e o rendimento de uma pessoa, em especial a mulher (VIITANEN, 2014). Além disso, o modelo possui também dummies para diferentes religiões (código V6121), dado que esse é um fator que pode vir a afetar a variável explicada (EWING, 2000).

O modelo contem preditores sobre a presença de deficiências (código V0614-V0617), sendo elas 4: auditiva, motora, visual e mental. A literatura, ainda que reduzida nessa parte, nos informa que pessoas com deficiências têm salários mais baixos (GANNON; MUNLEY, 2009), logo inclui-las aqui dará uma análise ainda mais ampla.

Incluímos variáveis sobre a nacionalidade (código V0620) dado a existência de uma diferença entre o salário de nativos e imigrantes no Brasil (VILELA, 2008). Ademais, colocamos informações sobre recebimento de algum tipo de transferência de renda pois,

no geral, pessoas mais pobres (menor salário) são mais aptas a receber algum tipo de benefício.

Ademais, o modelo possui *dummies* para diferentes níveis de atividade e variáveis relacionadas ao trabalho, como horas trabalhadas (código V0653) ,tempo de deslocamento até o trabalho (código V0662) e se a pessoa volta para casa depois do turno (código V0661). Primeiramente, profissões diferentes possuem remuneração diferente, logo queremos captar o efeito dela. Em segundo lugar, o quanto uma pessoa trabalha tem uma relação positiva com seu salário, pois quanto mais se produz, mais se ganha. Em terceiro lugar, se deslocar para um serviço distante pode indicar a busca por melhores salários (GOULD, 2007).

Entretanto, existem algumas variáveis que gostaríamos de poder utilizar, mas infelizmente elas não estão disponíveis em nossa base de dados. Elencamos algumas delas a seguir.

Não temos uma boa variável de classe social. Ela nos possibilitaria analisar como cada um das variáveis analisadas variariam depende classe em que a pessoa se encontra.

Não temos informação para saber se determinado indivíduo possui conhecimentos de outras línguas, pois isso se refletiria em uma salário mais alto (LIWIŃSKI, 2019), e pode estar sendo capturado pela variável de nacionalidade, por exemplo.

O tempo que a pessoa está no emprego também pode influenciar seu salário, e podemos estar associando um salário baixo à características sociais quando na verdade é apenas dado a ser uma contratação recente.

Informação sobre quantos dos membros da família trabalham e quantos precisam ser sustentados também ajudaria a diminuir o viés de variável omitida.

Por último, como explicado na seção de metodologia, seria interessante controlar pelas cidades dentro da região, uma vez que elas apresentam uma série de características geográficas e políticas diferentes, o que afeta o estado das variáveis, como vemos no apêndice D.

5 Resultados

Nessa seção discutiremos os resultados encontrados nas regressões efetuadas. Iremos, assim como em todo o texto, dividir ela em duas partes, tendo em vista os dois objetivos do trabalho.

5.1 Primeiro objetivo

Vamos começar primeiramente analisando as regressões múltiplas feitas disponíveis na tabela 1

5.1 Primeira Regressão

Na primeira regressão utilizamos como variáveis explicativas as *dummies* de deficiência. Começamos com essas variáveis para ter uma atenção especial no efeito das deficiências no salário, mesmo que nosso objetivo fosse algo mais geral. Como podemos perceber, todos os coeficientes são significantes a 5%. Vale ressaltar aqui, que a variável "INTELECTO", assume valor 1 se a pessoa não tem deficiência mental [1], por isso ela está positiva, diferentemente das outras. Note que, por enquanto, o resultado corrobora com a literatura existente apresentada na revisão bibliográfica.

5.2 Segunda Regressão

A segunda regressão adicionou novas variáveis ao modelo, em especial, aquelas ligadas a escolaridade, nacionalidade, sexo, raça e idade. Como podemos perceber, todas elas são significantes, mas o mais impressionante é a mudança de coeficiente ligado as deficiências, onde tivemos uma variação considerável. Isso aconteceu pois o que antes era explicado apenas por elas, agora se distribui para outras informações como, por exemplo, raça.

[1] Ver o apêndice A para a explicação de todas as variáveis

Tabela 1 – Tabela com o modelo principal

	Variável dependente:			
	log(Salário)			
	(1)	(2)	(3)	(4)
CONSTANTE	6.702* (0.015)	5.257* (0.013)	5.193* (0.013)	5.488* (0.013)
VISÃO	−0.160* (0.006)	−0.094* (0.004)	−0.092* (0.004)	−0.086* (0.004)
AUDIÇÃO	−0.047* (0.011)	−0.031* (0.008)	−0.030* (0.008)	−0.039* (0.008)
LOCOMOÇÃO	−0.218* (0.011)	−0.120* (0.008)	−0.121* (0.008)	−0.106* (0.008)
INTELECTO	0.270* (0.015)	0.160* (0.011)	0.162* (0.011)	0.152* (0.010)
ZONA MUNICIPAL		−0.190* (0.006)	−0.177* (0.005)	−0.176* (0.005)
SEXOf		−0.261* (0.002)	−0.255* (0.001)	−0.210* (0.002)
RAÇAn		−0.126* (0.003)	−0.120* (0.002)	−0.110* (0.002)
RAÇAa		0.087* (0.006)	0.082* (0.005)	0.083* (0.005)
RAÇAp		−0.123* (0.002)	−0.116* (0.001)	−0.110* (0.001)
RAÇAi		−0.144* (0.018)	−0.138* (0.017)	−0.130* (0.017)
NACIONALIDADEn		0.214* (0.020)	0.205* (0.020)	0.196* (0.020)
NACIONALIDADEe		0.191* (0.011)	0.181* (0.010)	0.173* (0.010)
EDUCAÇÃO2		0.140* (0.002)	0.135* (0.002)	0.127* (0.002)
EDUCAÇÃO3		0.299* (0.002)	0.285* (0.002)	0.271* (0.002)
EDUCAÇÃO4		0.943* (0.008)	0.888* (0.005)	0.863* (0.005)
HORAS TRAB.		0.005* (0.0001)	0.005* (0.0001)	0.005* (0.0001)
RENDA FAM./C		0.00000* (0.00000)	0.00000* (0.00000)	0.00000* (0.00000)
IDADE		0.051* (0.0003)	0.052* (0.0003)	0.046* (0.0003)
IDADE^2		−0.0005* (0.00000)	−0.001* (0.00000)	−0.0004* (0.00000)
TEMPO DESLOC.			0.013* (0.001)	0.013* (0.001)
N TRABALHOS			0.026* (0.004)	0.028* (0.004)
APOSENTADORIA			0.058* (0.004)	0.056* (0.004)
BOLSA-FAMÍLIA			−0.224* (0.005)	−0.203* (0.005)
TRANSFERÊNCIAS			−0.049* (0.006)	−0.042* (0.006)
OUTROS			0.200* (0.005)	0.194* (0.005)
EXTRAS TOTAL			−0.00004* (0.00000)	−0.00004* (0.00000)
RETORNAR			0.126* (0.005)	0.130* (0.005)
ESTADO CIVIL d				−0.032* (0.003)
ESTADO CIVIL v				−0.046* (0.005)
ESTADO CIVIL s				−0.138* (0.002)
N FAMÍLIA				−0.004* (0.001)
IDADE FILHO				−0.005* (0.0002)
N FILHOS VIVOS				−0.030* (0.001)
RELIGIÃO	NÃO	SIM	SIM	SIM
ATIVIDADE	NÃO	SIM	SIM	SIM
Observações	539,647	539,647	539,647	539,647
R^2	0.003	0.590	0.608	0.616
R^2 ajustado	0.003	0.590	0.608	0.616
Erro padrão do resíduo	0.733 (df = 539641)	0.470 (df = 539576)	0.459 (df = 539566)	0.455 (df = 539558)
Estatística F	311.203* (df = 5; 539641)	11,105.250* (df = 70; 539576)	10,467.240* (df = 80; 539566)	9,856.424* (df = 88; 539558)

Nota: Apenas indicamos a presença das variáveis de *RELIGIÂO* e *ATIVIVADE*; p<0.1; p<0.05; **p<0.01

Perceba ainda, que a idade ao quadrado possui coeficiente negativo, o que reflete a ideia que o salário das pessoas tende a subir até determinada idade e então começar a decair, provavelmente causado pela entrada na aposentadoria, por exemplo. Novamente, assim como a literatura nos diz, pessoas com maior nível educacional e homens possuem coeficientes maiores.

5.3 Terceira Regressão

A terceira regressão contou com a adição de variáveis relacionadas a outros tipos de ganhos de renda. A análise mais significativa que podemos fazer e que corrobora com a revisão bibliográfica é que pessoas ganham mais quanto trabalham longe de casa. Isso está ligado a ideia de buscar melhores oportunidades mais distantes da residência do que ter um salário menor trabalhando perto.

O efeito positivo de APOSENTADORIA pode indicar que pessoas aposentadas preferem não trabalhar por salários pequenos, portando quem está empregado e aparece em nossa base tem salários maiores. O efeito negativo de BOLSA-FAMÍLIA e TRANSFERÊNCIAS pode ser explicado por estar capturando o efeito parcialmente omitido de classe social menos avantajada, ou algum apontar que a presença desses benefícios gere uma mudança de incentivos que diminui o salário. Porém qualquer conclusão necessitaria uma análise mais aprofundada acerca desses benefícios.

Ademais, adicionamos características como atividade que a pessoa exerce e sua religião. As atividades que pagam maior remuneração são "Serviços de arquitetura e engenharia e atividades técnicas relacionadas: testes e análises técnicas"e "Serviços financeiros". Com relação a religiosidade, para nossa surpresa as duas que mais se destacam são as variáveis que representam a religião "Espirita"e a "Religiosidade cristã não determinada". As informações sobre "ATIVIDADE"e "RELIGIÃO", por falta de espaço, se encontram na tabela 7. [2].Note que, diferentemente da grande variação que ocorreu com os coeficientes da primeira para a segunda regressão, aqui a mudança foi menor.

[2] As duas colunas do meio da tabela 7 representam os valores para as duas últimas regressões. Note que não adicionamos o nome de todas as atividades dada sua extensão, mas você pode conferir o que cada uma representa acessando o dicionário do IBGE disponível em seu site na internet (IBGE, 2010). O número apresentado no nome da variável representa o mesmo número no dicionário, ou seja, "ATIVIDADE14001"representa a atividade de número 14001.

5.4 Quarta Regressão

Na quarta e última regressão adicionamos preditores sobre estado civil e informação sobre a família, como número de filhos. Um resultado que segue a mesma linha da literatura é que pessoas casadas tendem a ganhar um maior salário quando comparadas a solteira. Mas perceba também que o resultado vai além disso e mostra que essa discrepância salarial ocorre com divorciados e viúvos também.

Uma outra mudança significativa foi a dos coeficientes relacionados a ocupação e religião. Agora as categorias que ganham um maior salário são "Serviços de arquitetura e engenharia e atividades técnicas relacionadas; testes e análises técnicas"e "Atividades de alimentação não especificadas"para ocupação e "Testemunha de Jeová"e pessoas da "Igreja Evangélica Batista"para religião.

5.5 Regressão de interações

Por último, permitimos que houvessem interações entre as variáveis do modelo. Escolhemos interagir as quatro variáveis de deficiência contra raça, sexo, educação, horas trabalhadas, e outros rendimentos. A interação com raça e sexo não foram significativas, um teste F sobre os coeficientes de todas as interações par-a-par das variáveis retornou uma estatística de 0.924 e um p-valor de 0.5406, logo não rejeitamos a hipótese nula de todos os coeficientes associados serem zero. A relação com horas trabalhadas também foi pouco significativa, indicando que deficientes não recebem igual por um aumento no esforço semanal.

Com educação, apenas obtivemos valores significantes na interação com deficiência intelectual e visual (5), a relação negativa com intelecto pode indicar que pessoas com essa deficiência tiram menor proveito da educação normal, e por isso recebem menores salários; a relação positiva com visão pode significar que ao contrário da deficiência intelectual, a visual não impede tanto o proveito da educação. O teste F sobre as interações da tabela retornou uma estatística de 5.552, e um p-valor de 0.00083

Além disso, relações com outros rendimentos foram significativo para todos menos visão, de acordo com a tabela 6, indicando que talvez pessoas com essa dificuldade a mais se beneficiem mais de programas de distribuição de renda que pessoas não deficientes. Esse

conjunto de quatro interações retornou uma estatística F de 8.5397, rejeitando a nula com um p-valor de $6.917e-07$.

Também fizemos interações entre raça, sexo e educação, onde a maioria foi significante (tabela 4). A relação com raça e sexo trouxe coeficientes positivos, o que significa que homens de raças não brancas tem uma desvantagem no "a mais"trazido pela interação da sua raça e sexo, porém, o intercepto de raças não brancas e de mulheres continua sendo negativo, de modo que esses grupos continuam piores que os homens brancos. As relações significativas com educação foram todas negativas, mostrando que mulheres e raças não brancas recebem menores ganhos por aumentar sua educação que homens brancos. O teste F sobre todos os coeficientes associados à estas interações retornou uma estatística de 244.2, e um p-valor $2.2e-16$.

5.6 Validade dos resultados - Hipóteses de Gauss-Markov

Nessa seção discutiremos as hipóteses que garantem a consistência e a eficiência dos nossos estimadores. Especificamente, sobre a "quarta regressão", assumimos:

Hipóteses fracas:

- O modelo é linear nos parâmetros
- A amostra é independente e identicamente distribuída. (ALBIERI; BIANCHINI, 2015) descreve a preocupação do IBGE em captar uma amostra i.i.d.
- Não existe multicolinearidade perfeita entre variáveis. Calculamos os *Variance Inflation Factors* e as únicas variáveis que apresentavam alto grau de correlação eram $IDADE$ e $IDADE^2$, mas nenhuma apresentava correlação perfeita.
- Estabilidade (ou variância finita da matriz de covariadas). Assumimos que a matriz de variância-covariância das variáveis explicativas converge para uma matriz definida: $plim(\frac{1}{n}X'X) = Q$

Hipóteses fortes:

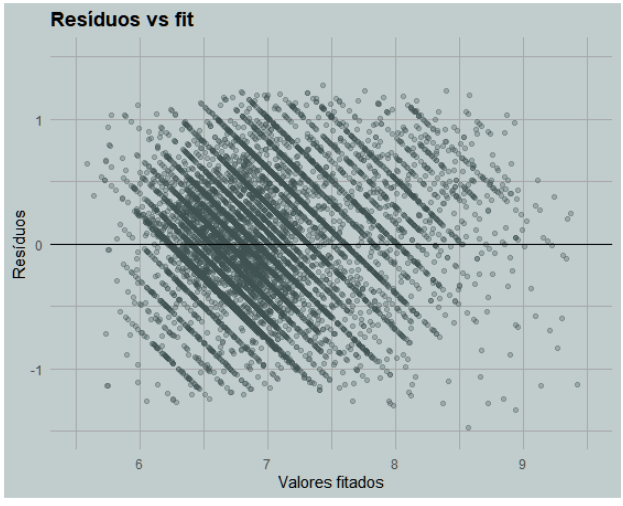
- Ortogonalidade. Como temos um grande número de observações, podemos substituir a hipótese de exogeneidade (garantidora de não viés), pela hipótese de ortogonalidade (garantidora de consistência). Como discutido na seção de dados, temos algumas

variáveis omitidas e elas implicariam em não-ortogonalidade. Porém, assumiremos que conseguimos capturar a maior parte das variáveis do modelo populacional, e que portanto, a interpretação de nossos estimadores como iguais aos parâmetros populacionais está errada, mas "por pouco".

- Não autocorrelação dos erros. Como sabemos que temos variáveis omitidas no erro, esse não será um termo 100% aleatório e temos motivos para desconfiar de autocorrelação.

- Homocedasticidade. Testamos essa hipótese com um teste de Breush-Pagan e White sobre os resíduos. Obtivemos uma estatística de teste de 174916 para o teste BP, portanto rejeitamos a hipótese nula de homocedasticidade com um p-valor $< 2.2e-16$, o teste de White rejeitou H_0 com o mesmo p-valor. A quebra dessa hipótese faz com que nosso modelo perca em eficiência, deixando de retornar estimadores $BLUE$. Além disso, nossos teste de hipótese são afetados, de modo que precisamos aplicar a correção de erros robustos de White. A heterocedasticidade provavelmente vem pelo fato de que a declaração do salário não é exatamente contínua, as pessoas declaram números redondos, logo valores de salário não redondos tendem a gerar resíduos maiores. Podemos ver isso através do gráfico 2.

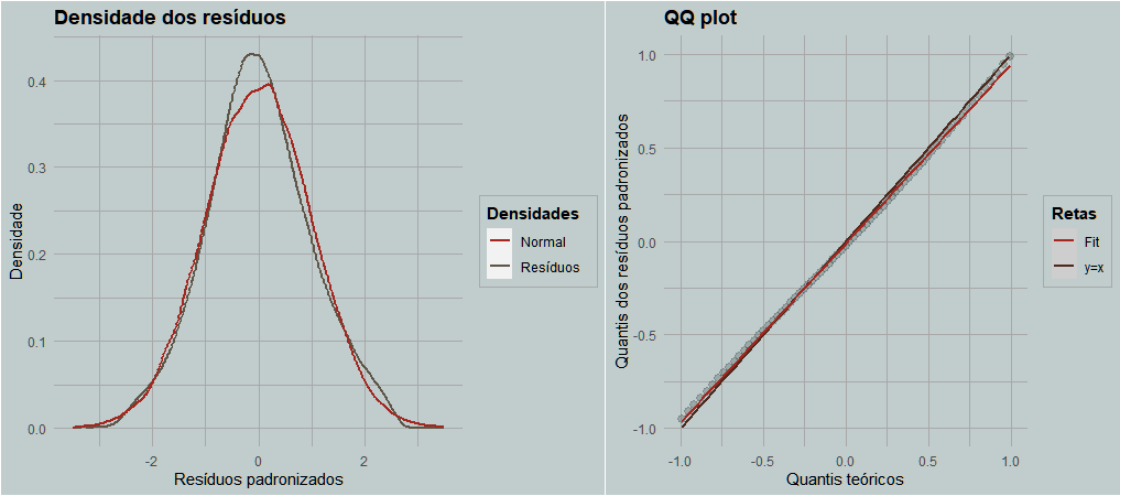
Figura 2 – Resíduos vs Fit



- Distribuição normal dos erros. Aplicamos um teste de Shapiro-Wilk sobre os resíduos, e obtivemos uma estatística de teste de 0.98909, rejeitando a hipótese nula de distribuição normal com um p-valor $< 2.2e-16$. A quebra dessa hipótese significa que nossos testes de hipótese terão sua validade questionável, essa é uma das

limitações de nosso trabalho. Provavelmente esse fato ocorre pelas variáveis omitidas no erro. Porém, o gráfico 3 traz um argumento visual de que nossa distribuição não está tão longe de uma normal. O primeiro painel mostra a densidade estimada dos erros contra uma normal, e o segundo painel é o plot quantil-quantil, que mostra a correlação entre os quantis de uma normal com os da distribuição do erro (vemos que essa correlação não é perfeita, mas é bem próxima de uma reta de 45º).

Figura 3 – Distribuição dos resíduos



5.2 Segundo objetivo

Como explicado na seção de metodologia, usamos as seguintes ferramentas de seleção: *foward selection*, *backward selection*, *stepwise selection*, *ridge*, *lasso*, e combinações entre as penalizações dos últimos dois.

A os três primeiros modelos de seleção trouxeram os mesmos resultados, o que faz sentido dado o nosso grande tamanho de amostra. Esses modelos apenas excluíram as variáveis de deficiência de audição, de visão, e de locomoção. O erro quadrático médio de previsão foi 0.2994626.

Depois fizemos os modelos de Ridge, Lasso, e várias combinações entre os dois, e reportamos a combinação que minimizou o EQM. O lâmbida do Lasso foi muito maior que o dos outros dois modelos reportados, de modo que foram zerados muitos coeficientes, incluindo todos os de deficiência. Já conseguimos perceber que por mais que pudemos chegar a uma conclusão sobre o efeito das deficiências sobre o salário, elas não parecem ter

um bom poder preditivo. O melhor EQM foi obtido na combinação das duas penalidades com um alpha de 0.2, obtendo uma estatística de 0.2892684, esse é então o modelo de previsão do salário recomendado por este artigo.

Uma limitação de nosso trabalho é não ter explorado modelos de previsão mais complexos, por exemplo, poderia ter sido escolhido um maior número de variáveis, e deixar a cargo de um algorítimo de *machine learning* para montar um modelo, que poderia ter retornado um EQM ainda menor.

Tabela 2 – Modelos de previsão de salário

	Variável dependente:				
	log(Salário)				
	Backward	Forward	Lasso	Ridge	$\alpha = 0.2$
CONSTANTE	5.455	5.457	6.533	5.491	5.508
VISÃO	−0.108	−0.108		-0.0015	-0.106
AUDIÇÃO	−0.048	−0.048		-0.0004	-0.045
LOCOMOÇÃO	−0.118	−0.118		-0.1074	-0.116
INTELECTO	0.179	0.179		0.0869	0.0856
ZONA MUNICIPAL	−0.222	−0.222		-0.220	-0.218
SEXOf	−0.208	−0.207	-0.172	-0.207	-0.218
RAÇAn	−0.152	−0.152		-0.151	-0.151
RAÇAa	0.097	0.097		0.095	0.090
RAÇAp	−0.151	−0.151		-0.150	-0.151
RAÇAi	−0.165	−0.165		-0.157	-0.151
NACIONALIDADEn	0.255	0.255		0.246	0.240
NACIONALIDADEe	0.216	0.216		0.210	0.207
EDUCAÇÃO2	0.155	0.155		0.153	0.151
EDUCAÇÃO3	0.325	0.325	0.161	0.323	0.322
EDUCAÇÃO4	0.988	0.987	0.944	0.988	0.988
HORAS TRAB.	0.006	0.006	0.004	0.006	0.0061
RENDA FAM./C	3.44e-7	3.47e-7	2.17e-7	3.43e-7	3.43e-7
IDADE	0.047	0.047	0.064	0.045	0.0445
IDADE^2	−0.0004	−0.0004		-0.0243	-0.000
TEMPO DESLOC.	0.009	0.008		0.008	0.0081
N TRABALHOS	0.083	0.083		0.082	0.0813
APOSENTADORIA	0.054	0.054		0.046	0.0428
BOLSA-FAMÍLIA	−0.249	−0.249		-0.246	-0.243
TRANSFERÊNCIAS	−0.066	−0.066		-0.063	-0.061
OUTROS	0.278	0.278	0.118	0.277	0.0277
EXTRAS TOTAL	−2.12e-5	−2.13e-5		-1.84e-5	-1.83e-5
RETORNAR	0.135	0.135		0.1329	0.1312
ESTADO CIVIL d	−0.083	−0.083		-0.080	-0.078
ESTADO CIVIL v	−0.028	−0.028		-0.025	-0.023
ESTADO CIVIL s	−0.075	−0.075		-0.075	-0.076
N FAMÍLIA	−0.011	−0.011		-0.010	-0.010
IDADE FILHO	−0.004	−0.004		-0.003	-0.003
N FILHOS VIVOS	−0.035	−0.035	-0.028	-0.033	-0.033
RELIGIÃO	SIM	SIM	NÃO**	SIM	SIM
ATIVIDADE	SIM	SIM	uma**	SIM	SIM
EQM	0.2994626	0.2994862	0.3209636	0.2892964	0.2892684
Lambda			0.04333	0.00025	0.00045

Nota: * Todas as atividades e religiões foram zeradas, fora a "Serviços domésticos"

6 Conclusão

Neste artigo, tendo em mente o debate sobre salário e uma literatura fragmentada sobre o tema, decidimos analisar tanto os fatores que o afetam, assim como fazer uma previsão dele. A pesquisa se deu com os dados do Censo de 2010 de IBGE na Região Metropolitana de São Paulo

Os principais resultados empíricos desse artigo corroboram, em sua maioria, com os já existentes na literatura atual, ou seja, maior educação, ser homem, ser casado e não ter deficiência proporciona um maior salário que sua contrapartida . Entretanto, mostramos também que diferentes atividades e religiões também tem um impacto significativo sobre o salário.

Outros resultados foram a relação quadrática do salário com idade; o maior salário associado a ser naturalizado ou estrangeiro (em contrapartida de nascido no Brasil) e; O menor salário em zonas rurais, para aposentados e para famílias maiores, entre outros resultados interessantes tratados na seção de resultados.

As interações trouxeram evidências que portadores de deficiências mentais aproveitam menos um acréscimo na educação, deficientes no geral aproveitam melhor programas de distribuição de renda e não sofrem discriminação racial diferentemente que não deficientes. Os grupos de mulheres com raças não brancas também tem um coeficiente especialmente menor do que homens brancos.

Infelizmente não foi possível tirar todo o viés de variável omitida, dado que algumas delas não estavam disponível na base de dados. Ainda sim, acreditamos que os nossos resultados são muito próximos da realidade.

Com relação a previsão, o modelo que nos propiciava o menor erro quadrático médio era o *elastic net* com $\alpha - 0.2$ $\lambda = 0.00045$. Ademais, percebemos que a presença de deficiências tem um baixo poder explicativo, ao passo que raça e nível de educação são os que tem os coeficientes maiores.

Possíveis novos rumos para a pesquisa após o trabalho estão associados com combater as limitações deste: principalmente no que tange obter as variáveis omitidas

elencadas, para dar maior validade às estimativas obtidas, mas também a utilização de maior poder computacional para rodar modelos mais complexos de previsão, selecionar diferentes quebras de treino na base, rodar dados com o país inteiro, explorar mais interações, etc.

Referências

ALBIERI, S.; BIANCHINI, Z. M. Principais aspectos de amostragem das pesquisas domiciliares do ibge-revisão 2015. *Rio de Janeiro: IBGE. Texto para Discussão*, n. 55, 2015.

APPLETON, S.; HODDINOTT, J.; KRISHNAN, P. The gender wage gap in three african countries. *Economic development and cultural change*, The University of Chicago Press, v. 47, n. 2, p. 289–312, 1999.

BAAMONDE, G.; MARTINS, P. *O impacto das deficiências no salário: uma análise do cenário paulista*. [S.l.], 2020.

BECKER, K. L. Deficiência, emprego e salário no mercado de trabalho brasileiro. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 49, n. 1, p. 39–64, 2019.

CHUN, H.; LEE, I. Why do married men earn more: productivity or marriage selection? *Economic Inquiry*, Wiley Online Library, v. 39, n. 2, p. 307–319, 2001.

EWING, B. T. The wage effects of being raised in the catholic religion: does religion matter? *American Journal of Economics and Sociology*, Wiley Online Library, v. 59, n. 3, p. 419–432, 2000.

FILHO, N. A. M.; MENDES, M.; ALMEIDA, E. S. d. O diferencial de salários formal-informal no brasil: segmentação ou viés de seleção? *Revista brasileira de economia*, SciELO Brasil, v. 58, n. 2, p. 235–248, 2004.

GANNON, B.; MUNLEY, M. Age and disability: explaining the wage differential. *Social Science & Medicine*, Elsevier, v. 69, n. 1, p. 47–55, 2009.

GOULD, E. D. Cities, workers, and wages: A structural analysis of the urban wage premium. *The Review of Economic Studies*, Wiley-Blackwell, v. 74, n. 2, p. 477–506, 2007.

HECKMAN, J. J.; LYONS, T. M.; TODD, P. E. Understanding black-white wage differentials, 1960-1990. *American Economic Review*, v. 90, n. 2, p. 344–349, May 2000. Disponível em: <https://www.aeaweb.org/articles?id=10.1257/aer.90.2.344>.

IBGE. *Censo Demográfico 2010: Características gerais da população*. Rio de Janeiro: [s.n.], 2010. Disponível em: <http://www.ibge.gov.br/home/estatistica/populacao/censo2000/populacao/censo2000_populacao.pdf>. Acesso em: 3 junho. 2020.

JAMES, G. et al. *An introduction to statistical learning*. [S.l.]: Springer, 2013. v. 112.

LIWIŃSKI, J. The wage premium from foreign language skills. *Empirica*, Springer, v. 46, n. 4, p. 691–711, 2019.

MADALOZZO, R. Occupational segregation and the gender wage gap in brazil: an empirical analysis. *Economia aplicada*, SciELO Brasil, v. 14, n. 2, p. 147–168, 2010.

PIERCE, B. Using the national compensation survey to predict wage rates. *Compensation and Working Conditions*, Bureau of Labor Statistics, v. 4, n. 4, p. 8–16, 1999.

VIITANEN, T. The motherhood wage gap in the uk over the life cycle. *Review of Economics of the Household*, Springer, v. 12, n. 2, p. 259–276, 2014.

VILELA, E. M. Imigração internacional e estratificação no mercado de trabalho brasileiro. Universidade Federal de Minas Gerais, 2008.

A Nome das variáveis

Variável	Explicação
VISÃO	1 se a pessoa tem problema de visão
AUDIÇÃO	1 se a pessoa tem problema de audição
LOCOMOÇÃO	1 se a pessoa tem problema de locomoção
INTELECTO	0 se a pessoa tem problema mental
ZONA MUNICIPAL	1 se a pessoa mora na zona rural
SEXO	1 se a pessoa é mulher
RAÇAn	1 se a pessoa é de raça preta
RAÇAa	1 se a pessoa é de raça amarela
RAÇAp	1 se a pessoa é de raça parda
RAÇAi	1 se a pessoa é de raça indígena
NACIONALIDADEn	1 se a pessoa é naturalizada brasileira
NACIONALIDADEe	1 se a pessoa é estrangeira
EDUCAÇÃO2	1 se a pessoa tem fundamental completo
EDUCAÇÃO3	1 se a pessoa tem médio completo
EDUCAÇÃO4	1 se a pessoa tem superior completo
HORAS TRABALHADAS	Número de horas trabalhadas na semana
RENDA FAM/C	Rendimento familiar per capita
IDADE	Idade cálculada em anos
$IDADE^2$	Idade ao quadrado
TEMPO DESLOCAMENTO	Tempo de deslocamente até o trabalho
N TRABALHOS	1 se a pessoa tinha mais de 2 trabalhos
APOSENTADORIA	1 se a pessoa recebe aposentadoria
BOLSA FAMILIA	1 se a pessoa recebe Bolsa Família
TRANSFERÊNCIAS	1 se a pessoa recebe de outros programas sociais
OUTROS	1 se a pessoa recebe outras fontes de renda
EXTRAS TOTAL	1 se a pessoa recebe Bolsa Família
RETORNAR	1 se a pessoa retorna do trabalho para a casa diariamente
ESTADO CIVIL.d	1 se a pessoa é divorciada
ESTADO CIVIL.v	1 se a pessoa é viúva
ESTADO CIVIL.s	1 se a pessoa é solteira
N FAMÍLIA	Número de pessoas na família
IDADE FILHO	Idade do último filho nascido até 31/07/2010
N FILHOS VIVOS	Total de filhos vivos
RELIGIÃO	Código da atividade exercida
ATIVIDADE	Código da regilião ou culto

B Alguns resultados do trabalho

Tabela 3 – Add caption

Característica	Valor	Salário médio
Zona municipal	Urbana	1083.72
	Rural	701.52
Sexo	Masc.	1223.21
	Fem.	920.4
Raça	Branca	1274.65
	Preta	863.43
	Amarela	1827.36
	Parda	821.04
	Indígena	890.84
Nacionalidade	Brasileiro	1072.92
	Naturalizado	2259.73
	Estrangeiro	1814.32
Educação	Fund. incompleto	744.24
	Fund. completo	837.47
	Médio incompleto	1048.33
	Médio completo	2683.98
Estado civil	Casado	1305.84
	Separado	1124.24
	Divorciado	1327.68
	Viúvo	908.77
	Solteiro	897.09
Trabalhos	1	1061.81
	2 ou mais	1648.8
Aposentado	Sim	1064.92
	Não	1282.86
Bolsa família	Sim	1089.01
	Não	562.37
Transferências	Sim	1080.91
	Não	899.1

Tabela 4 – Regressão de interação Sexo, raça e educação

Interação	Coeficiente (erro padrão)
SEXOf:RAÇAn	5.880e-02 (4.970e-03) *
SEXOf:RAÇAa	9.532e-03 (9.496e-03)
SEXOf:RAÇAp	3.547e-02 (2.814e-03) *
SEXOf:RAÇAi	1.168e-01 (3.498e-02) *
RAÇAn:EDUCAÇÃO2	-2.094e-02 (6.885e-03) **
RAÇAa:EDUCAÇÃO2	-2.472e-02 (1.844e-02)
RAÇAp:EDUCAÇÃO2	-1.883e-02 (3.912e-03) *
RAÇAi:EDUCAÇÃO2	-1.274e-01 (4.882e-02) **
RAÇAn:EDUCAÇÃO3	-6.865e-02 (5.907e-03) *
RAÇAa:EDUCAÇÃO3	1.702e-02 (1.486e-02)
RAÇAp:EDUCAÇÃO3	-6.854e-02 (3.375e-03) *
RAÇAi:EDUCAÇÃO3	-1.503e-01 (4.190e-02) *
RAÇAn:EDUCAÇÃO4	-1.949e-01 (9.120e-03) *
RAÇAa:EDUCAÇÃO4	-4.388e-02 (1.418e-02) **
RAÇAp:EDUCAÇÃO4	-1.948e-01 (5.246e-03) *
RAÇAi:EDUCAÇÃO4	-1.878e-01 (5.954e-02) **
SEXOf:EDUCAÇÃO2	-4.950e-02 (3.868e-03) *
SEXOf:EDUCAÇÃO3	-7.968e-02 (3.454e-03) *
SEXOf:EDUCAÇÃO4	-1.551e-01 (4.281e-03) *
Estatística F (p-valor)	244.2 (<2.2e-16)

Tabela 5 – Regressão de interação entre educação e deficiência ocular

Interação	Coeficiente (erro padrão)
EDUCAÇÃO2:VISÃO	2.783e-02 (1.140e-02) *
EDUCAÇÃO3:VISÃO	3.840e-02 (1.004e-02) *
EDUCAÇÃO4:VISÃO	2.900e-02 (1.318e-02) *
EDUCAÇÃO2:INTELECTO	-3.717e-02 (1.776e-01)
EDUCAÇÃO3:INTELECTO	-5.156e-02 (4.064e-02) *
EDUCAÇÃO4:INTELECTO	-7.115e-02 (5.645e-02) .
Estatística F (p-valor)	5.5525 (0.00083)

Tabela 6 – Regressão de interação entre ganhos totais e deficiência.

Interação	Coeficiente (erro padrão)
EXTRAS TOTAL:INTELECTO	2.360e-05 (8.529e-06) **
EXTRAS TOTAL:VISÃO	-1.292e-06 (1.928e-06)
EXTRAS TOTAL:AUDIÇÃO	3.147e-05 (7.232e-06) *
EXTRAS TOTAL:LOCOMOÇÃO	2.084e-05 (7.960e-06) **
Estatística F (p-valor)	8.5397 (6.917e-07)

Tabela 7 – Resultado Regressão Ocupação e Religião

Variável	(3)	(4)	(Teste t)	(p-valor)	
ATIVIDADE14001	-2,38E2	7,07E0	-33.701	< 2e-16	*
ATIVIDADE14999	-1,20E2	7,43E0	-16.171	< 2e-16	*
ATIVIDADE18000	2,76E2	7,51E1	3.680	0.000233	*
ATIVIDADE25001	5,73E3	4,13E2	13.868	< 2e-16	*
ATIVIDADE32999	3,08E4	5,79E3	5.316	1.06e-07	*
ATIVIDADE43000	-1,17E6	3,96E4	-29.482	< 2e-16	*
ATIVIDADE43999	-1,53E6	4,72E5	-3.250	0.001154	**
ATIVIDADE45020	-5,37E7	5,71E6	-9.407	< 2e-16	*
ATIVIDADE48030	-1,12E9	4,59E7	-24.336	< 2e-16	*
ATIVIDADE48042	-2,34E9	6,05E8	-3.866	0.000111	*
ATIVIDADE48071	-4,48E10	7,46E9	-6.006	1.90e-09	*
ATIVIDADE48080	-1,35E12	6,93E10	-19.449	< 2e-16	*
ATIVIDADE48100	-2,03E13	6,54E11	-31.079	< 2e-16	*
ATIVIDADE48999	1,65E13	3,63E12	4.533	5.81e-06	*
ATIVIDADE49030	7,44E14	4,59E13	16.213	< 2e-16	*
ATIVIDADE49999	-2,88E15	6,75E14	-4.258	2.06e-05	*
ATIVIDADE56011	-1,13E17	4,50E15	-25.168	< 2e-16	*
ATIVIDADE56999	-8,37E17	7,84E16	-10.677	< 2e-16	*
ATIVIDADE62000	1,97E19	5,98E17	32.991	< 2e-16	*
ATIVIDADE64000	2,17E20	5,09E18	42.723	< 2e-16	*
ATIVIDADE68000	5,08E20	7,76E19	6.547	5.88e-11	*
ATIVIDADE69000	1,20E22	5,00E20	24.015	< 2e-16	*
ATIVIDADE71000	2,64E23	7,92E21	33.306	< 2e-16	*
ATIVIDADE80000	-7,39E23	5,95E22	-12.435	< 2e-16	*
ATIVIDADE81011	-2,43E25	5,53E23	-43.872	< 2e-16	*
ATIVIDADE82001	6,35E24	6,34E24	1.002	0.316424	
ATIVIDADE82002	-1,55E27	6,55E25	-23.631	< 2e-16	*
ATIVIDADE84013	-3,65E27	6,53E26	-5.594	2.22e-08	*
ATIVIDADE84999	1,49E29	7,55E27	19.780	< 2e-16	*
ATIVIDADE85012	-1,60E30	5,58E28	-28.613	< 2e-16	*
ATIVIDADE85029	-8,34E30	7,70E29	-10.825	< 2e-16	*
ATIVIDADE85999	-1,65E32	6,22E30	-26.503	< 2e-16	*
ATIVIDADE86001	1,15E33	5,29E31	21.635	< 2e-16	*
ATIVIDADE86002	3,35E33	6,92E32	4.836	1.33e-06	*
ATIVIDADE96020	-1,12E35	5,71E33	-19.667	< 2e-16	*
ATIVIDADE97000	-2,98E36	3,38E34	-88.277	< 2e-16	*
ATIVIDADEOutros	9,13E35	2,42E35	3.779	0.000157	*
RELIGIÃO110	2,72E37	2,26E36	12.059	< 2e-16	*
RELIGIÃO240	2,73E38	6,38E37	4.279	1.88e-05	*
RELIGIÃO310	-3,48E39	3,57E38	-9.751	< 2e-16	*
RELIGIÃO320	-2,90E40	4,59E39	-6.330	2.46e-10	*
RELIGIÃO350	-3,02E41	6,35E40	-4.757	1.97e-06	*
RELIGIÃO450	-5,95E41	4,26E41	-1.395	0.163155	
RELIGIÃO490	1,75E43	3,26E42	5.359	8.35e-08	*
RELIGIÃO520	-2,01E43	6,77E43	-0.297	0.766784	
RELIGIÃO610	1,35E46	3,87E44	34.803	< 2e-16	*
RELIGIÃO850	4,34E46	6,14E45	7.062	1.65e-12	*
RELIGIÃOOutros	2,24E47	3,05E46	7.338	2.18e-13	*

C Algoritmo dos modelos

Como citado no trabalho, utilizamos diferentes modos de escolha de modelo, especialmente pra fazer a previsão do salário. Dessa forma, deixo aqui nessa parte o algoritmo para cada um dos modos utilizados. Eles foram retirados do James et al. (2013).

C.1 Forward stepwise selection

1. Denote M_0 como sendo o modelo nulo, que não tem preditores.
2. Para $k = 0, 1, \ldots, p-1$:
   - Considere todos os $p-k$ modelos que aumentam o preditor em M_k com um preditor adicional
   - Escolhe o melhor entres os $p-k$ modelos, e chame de M_{k+1}. Aqui o melhor é definido como o melhor RSS ou o maior R_2
3. Escolha o único modelo entre os $M_0, \ldots, M_p$ usando *cross-validated prediction error*, $C_p(AIC), BIC$, ou R^2 ajustado

C.2 Backward stepwise selection

1. Denote M_p como sendo o modelo nulo, que não tem preditores.
2. Para $k = p, p-1, \ldots, 1$:
   - Considere todos os k modelos que contêm todos menos um dos preditores em M_k, para o total de $k-1$ preditores
   - Escolhe o melhor entres os k modelos, e chame de M_{k-1}. Aqui o melhor é definido como o melhor RSS ou o maior R_2
3. Escolha o único modelo entre os $M_0, \ldots, M_p$ usando *cross-validated prediction error*, $C_p(AIC), BIC$, ou R^2 ajustado

C.3 Regressão Ridge

A regressão ridge consiste em:

$$minimizar_{\beta}\left[\sum_{i=0}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2\right] sujeito à \sum_{j=1}^{p}\beta_j^2 \leq s \tag{C.1}$$

C.4 Regressão Lasso

A regressão lasso consiste em:

$$minimizar_{\beta}\left[\sum_{i=0}^{n}(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij})^2\right] sujeito à \sum_{j=1}^{p}|\beta_j| \leq s \tag{C.2}$$

C.5 Elastic Net

A elastic net regulatization é uma combinação da lasso e da ridge, que consiste em achar o valor de $\beta, talque$:

$$\hat{\beta} \equiv \beta argmin\left(\|y-X\beta\|^2+\lambda_2\|\beta\|^2+\lambda_1\|\beta\|_1\right) \tag{C.3}$$

Onde no nosso caso, $\lambda_i = 1-\lambda_j, \forall j \neq i$

C.6 Sequential Replacement

O *sequential replacement* consiste em uma combinação do *forward* e do *backward stepwise selection*. Começamos sem nenhum preditor e vamos adicionando o melhor preditor, assim como no *forward*. Depois de adicionar a variável, removemos qualquer variável que não melhora nosso modelo, da mesma forma que o *backward*.

D Mapas espaciais dos dados

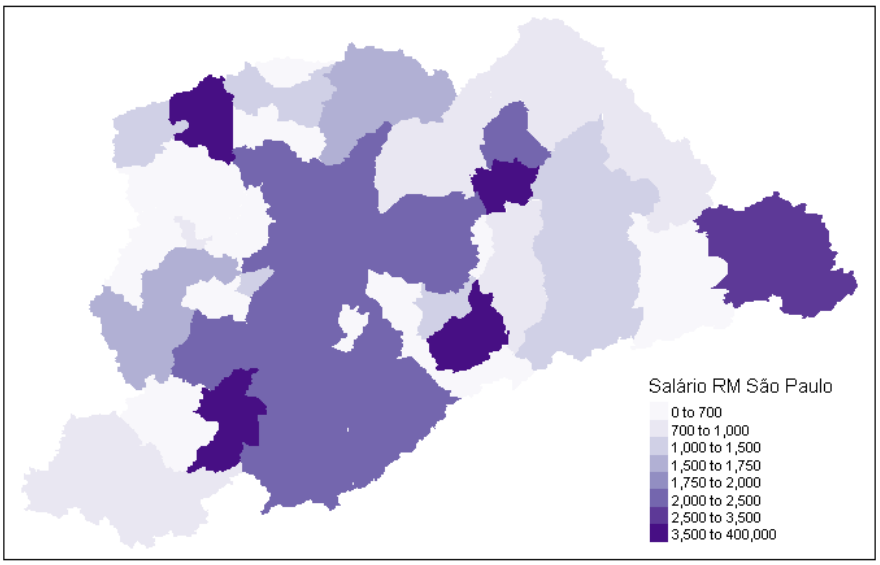


Figura 4 – Salário médio na RMSP

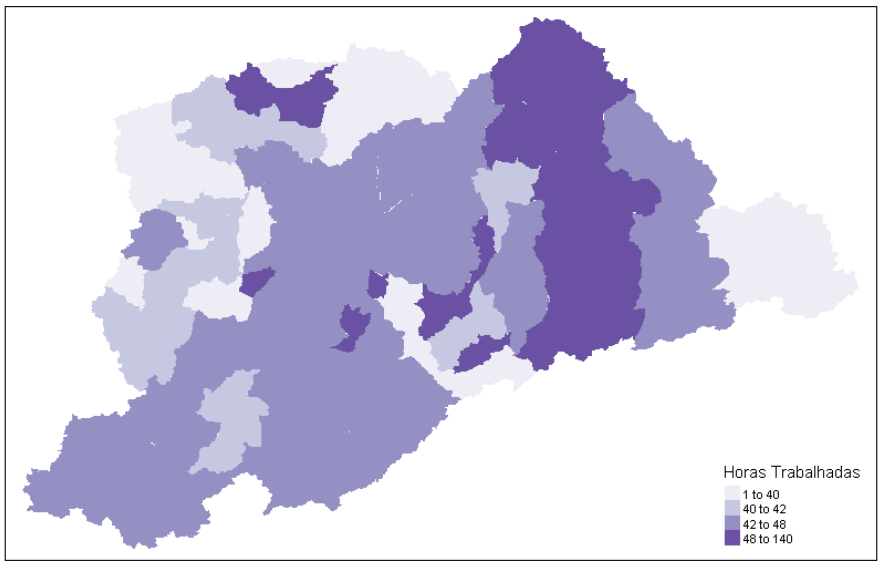


Figura 5 – Horas Trabalhadas na RMSP

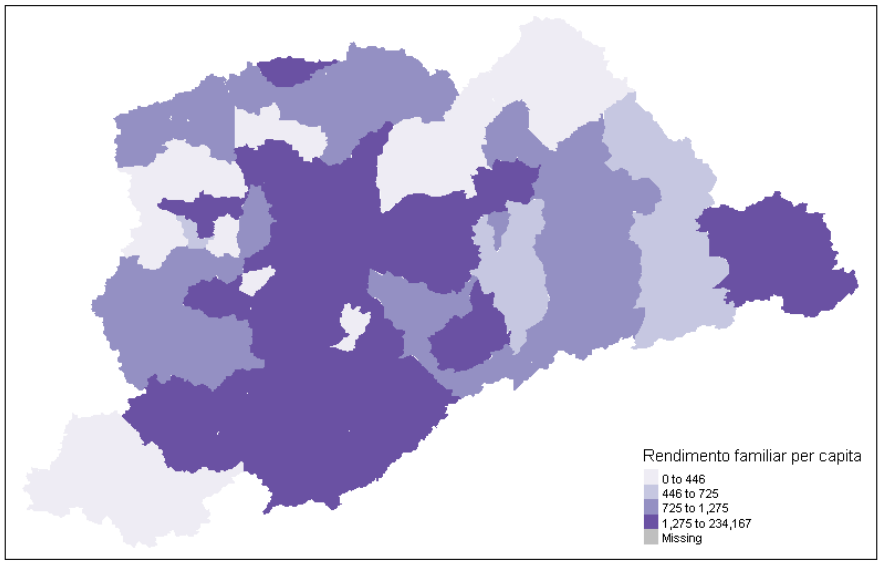


Figura 6 – Rendimento Familiar per capita na RMSP

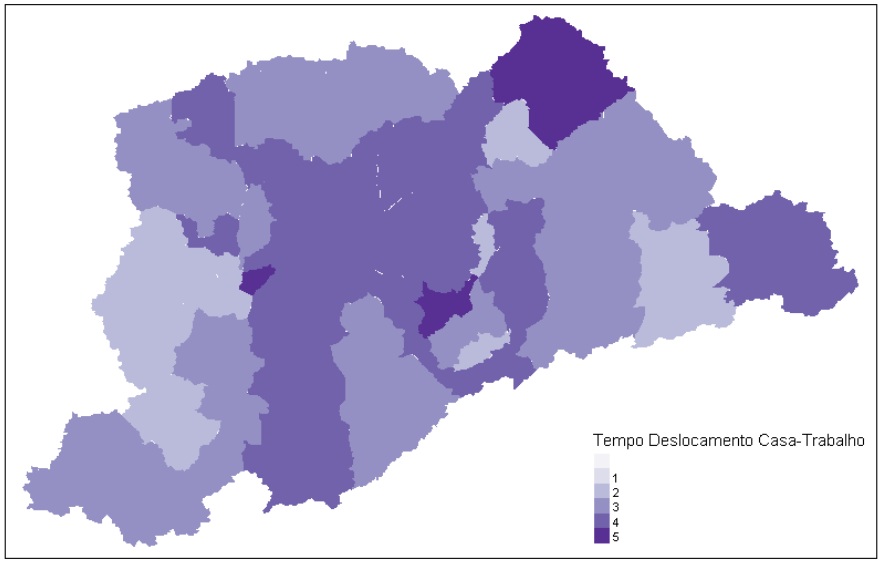


Figura 7 – Tempo de Deslocamento até o trabalho na RMSP