4 APLICAÇÃO A DADOS REAIS

Para exemplificar o uso do modelo de regressão linear normal com priori conjugada natural apresentado no capítulo 1, será desenvolvida uma aplicação baseada no artigo de Bonini e Pozzobon (2016), que fala de prêmio salarial e discriminação de gênero nos salários de alguns setores produtivos da região Sul do Brasil. Diferentemente do artigo original, nesta monografia serão analisados apenas os dados do setor de tecnologia de informação (TI) para o estado de Santa Catarina. Optou-se por essa simplificação de redução do modelo original pois o foco da monografia é apresentar o método de estimação.

Este capítulo está dividido da seguinte forma: uma breve introdução ao referencial teórico sobre mercado de trabalho e prêmio salarial, além do modelo para o problema são apresentados, seguida pela descrição dos dados da amostra. As estimativas utilizando o mesmo modelo de regressão normal com priori conjugada natural normal gama são comparadas com as estimativas de MQO e depois é feita a previsão, pelos dois métodos, para 10 elementos de fora da amostra. Ressalta-se, também, que a análise tem como objetivo verificar se o modelo estimado utilizando as técnicas bayesianas irá produzir resultados consistentes com os obtidos no trabalho original de Bonini e Pozzobon (2016) e por isso não foi realizada a interpretação econômica das estimativas de maneira mais aprofundada.

4.1 PRÊMIO SALARIAL E DISCRIMINAÇÃO DE GÊNERO NO MERCADO DE TRABALHO DE TI

A falta de representatividade feminina em carreiras nas áreas de ciência, tecnologia, engenharias e matemática (STEM[1]) tem sido discutida na literatura por suas implicações para as mulheres e para a sociedade como um todo. Por exemplo, a falta de mulheres participantes em estudos clínicos levou à posterior descoberta de que os efeitos de alguns medicamentos eram diferentes nas mulheres em comparação com os homens. Além disso, carreiras nas áreas de STEM usualmente tem um salário mais alto em comparação com as demais áreas, acentuando o gap salarial de gênero (KAHN; GINTHER, 2017). Bonini e Pozzobon (2016) apontam para a crescente demanda de profissionais nessa área que não é acompanhada pela oferta, o que gera a pressão para aumento dos salários.

Proposta por Becker na década de 50, a teoria do capital humano relaciona as variáveis intrínsecas aos indivíduos com a sua produtividade. Especificamente, a teoria assume que educação formal e treinamento dentro da empresa determinam atividades

[1] Do inglês *science, technology, engineering and mathematics.*

cognitivas que por sua vez impactam em habilidades e produtividade. A produtividade, de acordo com a teoria neoclássica, deve ser igual a remuneração do fator mão-de-obra, que é o salário. Sendo assim, existe uma ligação entre as habilidades da pessoa (mensuradas pela educação e treinamento) e a sua remuneração.

A teoria do capital humano entra em um arcabouço econométrico através das funções mincerianas, que permitem uma avaliação quantitativa do efeito das características dos trabalhadores sobre o salário. Essa abordagem pode ser espandida, usando a decomposição de Oaxaca, para avaliação do efeito que a discriminação que determinados grupos sofrem tem nos salários.

O modelo econométrico utilizado tem o mesmo formato do modelo de Bonini e Pozzobon (2016), que possui a seguinte forma funcional:

$$\begin{aligned}\ln(\text{salário}) = \beta_0 + \beta_1 \cdot \text{Sup. Inc.} + \beta_2 \cdot \text{Sup. Comp.} + \beta_3 \cdot \text{Pós-Grad.} \\ + \beta_4 \cdot \text{Idade} + \beta_5 \cdot \text{Idade}^2 + \beta_6 \cdot \text{T. Emp.} + \beta_7 \cdot \text{Gênero} + \varepsilon, \end{aligned} \tag{4.1}$$

onde:

- *ln(salário)* é o vetor com os logarítmos naturais dos salários, em reais, dos trabalhadores;
- β_0 é o coeficiente de intercepto, que representa a heterogeneidade devida a variáveis não observadas no modelo e comuns a todos indivíduos;
- *Sup. Inc.*, *Sup. Comp.* e *Pós-Grad.* são vetores de variáveis binárias (com valores 0 e 1) que identificam se a escolaridade máxima do indivíduo é superior incompleto, superior completo ou pós graduação, respectivamente. Quando as três variáveis são iguais a 0 significa que o grau de escolaridade máximo da pessoa é ensino médio;
- *Idade* e $Idade^2$ representam os vetores com as idades, em anos, e seus valores ao quadrado. O quadrado da idade é utilizado por assumir que os retornos marginais no salário são decrescentes para anos de vida;
- *T. Emp.* é o vetor com os tempos de emprego (do emprego atual) dos indivíduos, mensurados em meses;
- *Gênero* é um vetor com a variável dicotômica que é 0 para homens e 1 para mulheres. Uma vez que o gênero de um indivíduo não é uma "habilidade", o coeficiente associado a esta variável mede o grau de discriminação salarial entre homens e mulheres;
- É assumido que os termos do vetor ε são independentes, identicamente distribuídos e não correlacionados com as variáveis explicativas.

4.2 DESCRIÇÃO DA BASE DE DADOS

Bonini e Pozzobon (2016) utilizaram dados da Relação Anual de Informações Sociais (RAIS) do Ministério do Trabalho para o ano de 2011. Como mencionado no início do capítulo, no trabalho original foram analisados os três estados da Região Sul do Brasil, porém para esta monografia optou-se por utilizar uma versão reduzida da base de dados contendo apenas as informações dos trabalhadores de TI que trabalhavam no Estado de Santa Catarina. Ao todo, a amostra utilizada neste trabalho compreende 4328 trabalhadores. Destes, 10 foram selecionados para servirem de validação para previsão de dados fora da amostra, fazendo com que o modelo tenha sido estimado com base em 4318 observações.

A variável resposta é o logaritmo neperiano do salário hora. Sendo assim, os coeficientes estimados representam a variação salarial, em termos percentuais, entre os diferentes atributos (mantendo os demais constantes). Além do salário, outras três variáveis são quantitativas: tempo no emprego atual (em meses), idade (em anos) e idade ao quadrado (em anos ao quadrado). As estatísticas descritivas para as variáveis quantitativas (removendo as 10 observações que foram utilizadas para previsão fora da amostra) encontram-se na tabela (3):

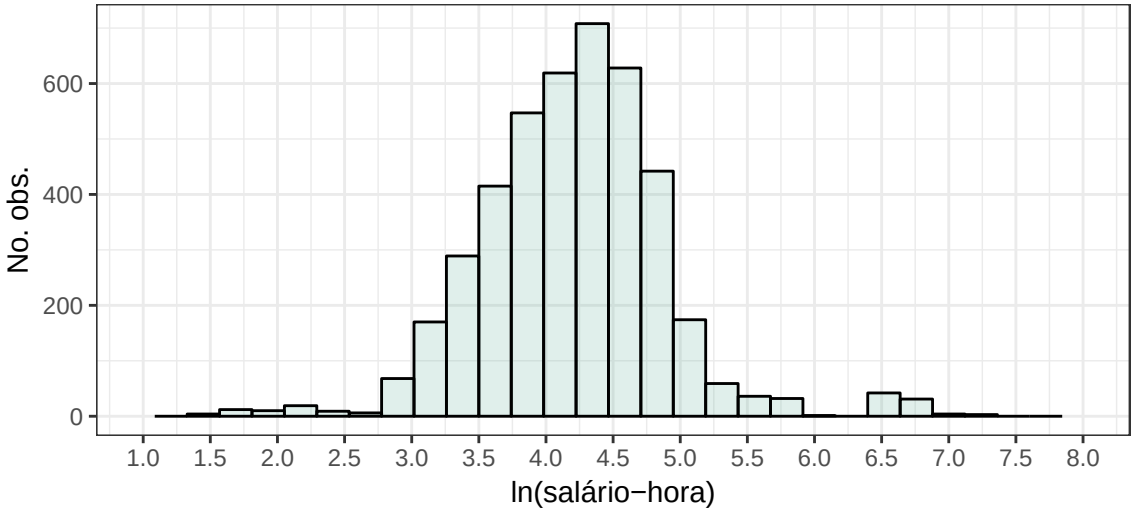
Tabela 3 – Estatísticas descritivas das variáveis quantitativas do modelo

	ln(Salário-Hora)*	Tempo de Emprego†	Idade**	(Idade)^2
N^o de Observações	4318	4318	4318	4318
Mínimo	1.38	0.10	17.00	289.00
1^o quartil	3.78	9.70	26.00	676.00
Mediana	4.22	16.90	29.00	841.00
Média	4.21	34.59	30.58	989.10
3^o quartil	4.60	37.10	34.00	1156
Máximo	7.32	433.40	72.00	5184
Desvio Padrão	0.69	54.26	7.35	527.32

* O salário-hora é medido em reais (R\$) de 2011 e ln(Salário-Hora) denota seu logaritmo na base neperiana.
† Tempo de emprego é medido em meses.
** Idade é medida em anos e, consequentemente, idade ao quadrado, denotada por (Idade)^2, tem como unidade anos ao quadrado.
Fonte – Elaboração própria utilizando dados da RAIS de trabalhadores de setores de TI no estado de Santa Catarina para o ano de 2011.

Tomando como base os valores dos logarítmos dos salários, os salários hora em reais variam de R$ 3.98 a R$ 1513.00, o que evidencia a heterogeneidade da variável resposta entre os indivíduos. O salário hora mediano é de R$ 68.23 enquanto que o salário médio é de R$ 89.40, indicando relativa simetria na distribuição da variável. O histograma do logaritmo dos salários está na Figura (4) e é possível perceber que de fato a distribuição está centralizada em torno da média de 4.2, com uma ligeira assimetria positiva (cauda para a direita), que é característica de dados de renda.

Figura 4 – Histograma do logaritmo do salário hora dos trabalhadores de TI do Estado de Santa Catarina para o ano de 2011



Fonte – Elaboração própria com base nos dados da RAIS.

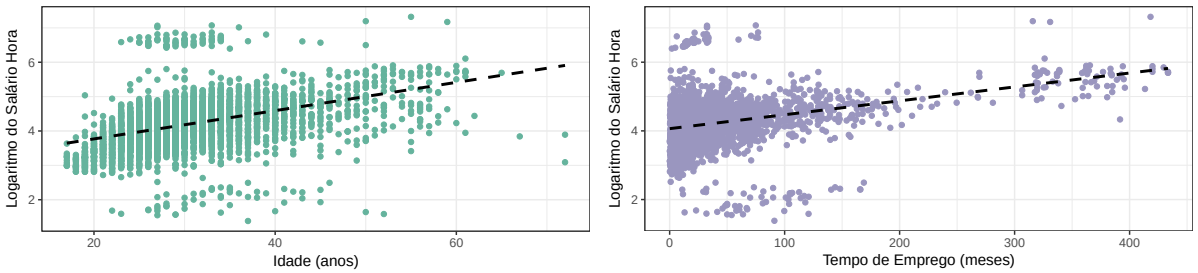
Com relação ao tempo de emprego, a maior parte dos trabalhadores (75%) tem no máximo 37 meses no serviço atual, o que corresponde a pouco mais de 3 anos, sendo que a mediana no tempo de emprego é de 16 meses e 27 dias. A idade dos trabalhadores também apresenta uma variação grande, indo de 17 a 72 anos. Embora o intervalo seja amplo, a mediana indica que até 50% dos trabalhadores tem no máximo 29 anos, evidenciando uma concentração de pessoas jovens nesta amostra.

Na figura (5) são apresentados os gráficos de dispersão tanto da idade como do tempo de emprego versus o logaritmo do salário-hora. Em ambos há uma tendência linear positiva, porém com grande concentração de valores menores no eixo x. Em particular, para a idade (Figura 5a), alguns outliers que apresentam pouca idade e elevados salários. Para o tempo de emprego (Figura 5b), há uma concentração ainda maior de baixos valores, porém há três pequenos grupos que se destacam: a) pessoas com pouco tempo de emprego e alto salário; b) pessoas com pouco tempo de emprego e c) salário bem abaixo da média e pessoas com muito tempo de emprego e altos salários. Em ambos gráficos da Figura (5) estão plotadas linhas tracejadas cujas inclinações e interceptos foram obtidos calculando a reta de regressão linear simples das variáveis.

As variáveis categóricas que compõe o modelo são as de escolaridade e gênero. A amostra é formada de pessoas com 4 níveis de escolaridade máximo: ensino médio, ensino superior incompleto, ensino superior e pós graduação. Para cada uma delas foi criada uma variável indicadora que é igual a 1 caso a pessoa tenha no máximo o grau de instrução correspondente e zero caso contrário, sendo que o ensino médio é a categoria base e portanto não é incluída no modelo para evitar multicolinearidade.

Ainda com relação às variáveis categóricas, para a variável gênero adotou-se como

Figura 5 – Gráficos de dispersão das variáveis explicativas quantitativas contra a variável explicada no modelo



(a) Distribuição da idade e logaritmo do salário hora

(b) Distribuição do tempo de emprego e logaritmo do salário hora

Fonte – Elaboração própria com base nos dados da RAIS.

categoria basal o sexo masculino, de forma que um homem que cursou no máximo o segundo ano de graduação teria o preenchimento de suas informações no banco de dados como descrito na tabela (4).

Tabela 4 – Exemplo de preenchimento do banco de dados

Variável	Gênero	Sup. Incompleto	Superior	Pós Graduação
Valor	0	1	0	0

Fonte – Elaboração própria.

Tabela 5 – Proporções nas categorias das variáveis categóricas incluídas no modelo

	Sim		**Não**	
	n	**%**	**n**	**%**
Homem*	3285	76,08	1033	23,92
Ensino superior incompleto†	697	16,14	3621	83,86
Ensino superior completo†	3201	74,13	1117	25,87
Pós graduação†	29	0,0067	4289	99,33

* A categoria "Homem" é considerada a categoria base para a variável do modelo "Gênero".

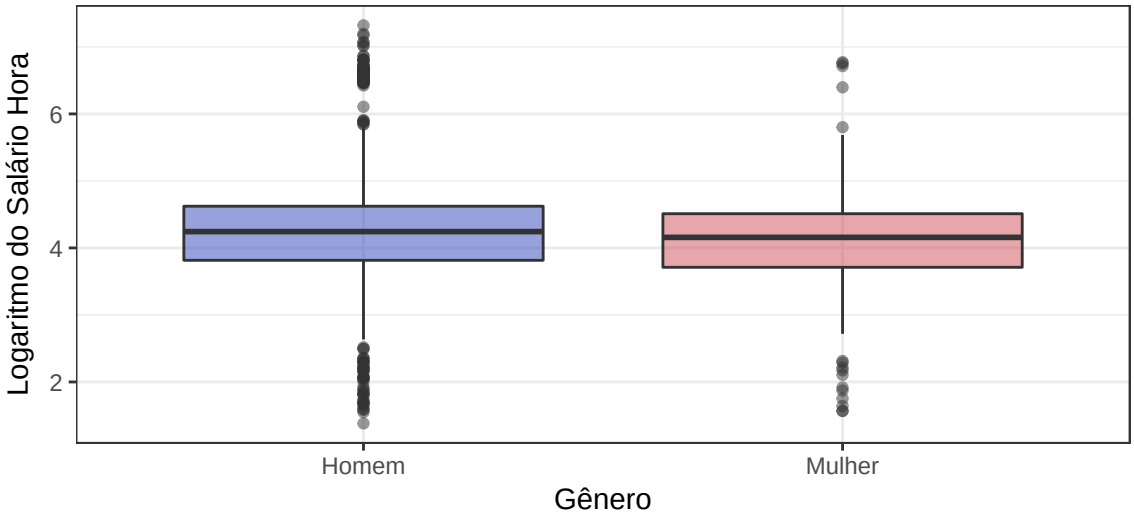
† As variáveis de escolaridade representam o grau de escolaridade máximo do indivíduo, isto é, uma pessoa que tenha cursado até o primeiro ano da faculdade terá "Sim" para ensino superior incompleto e "Não" para ensino superior completo e pós graduação. A categoria base da variável escolaridade é "Ensino Médio".

Fonte – Elaboração própria utilizando dados da RAIS de trabalhadores de setores de TI no estado de Santa Catarina para o ano de 2011.

As mulheres representam pouco menos de um quarto do total da amostra ($23,92\%$), o que vai ao encontro das evidências da literatura atual, que aponta diferenças significativas na proporção de mulheres que participam de carreiras STEM. Com relação à escolaridade, grande parte da amostra (74.12%) é formada de pessoas cujo maior grau de instrução é ensino superior completo, versus 0.01% com pós-graduação e 16.13% com ensino superior

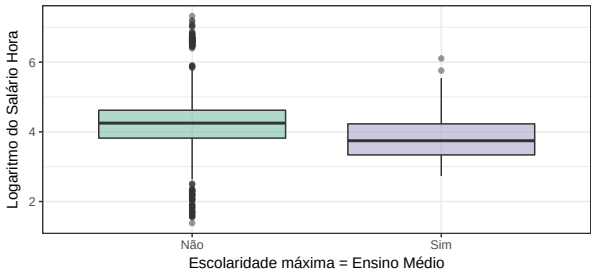
incompleto - os cerca de 10% restantes tem apenas ensino médio completo. A tabela (6) contém as proporções das variáveis categóricas na amostra.

Figura 6 – Boxplot do logaritmo do salário hora estratificado por gênero para trabalhadores de TI no Estado de Santa Catarina no ano de 2011

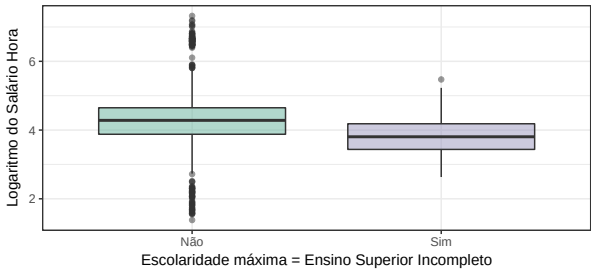


Fonte – Elaboração própria com base nos dados da RAIS.

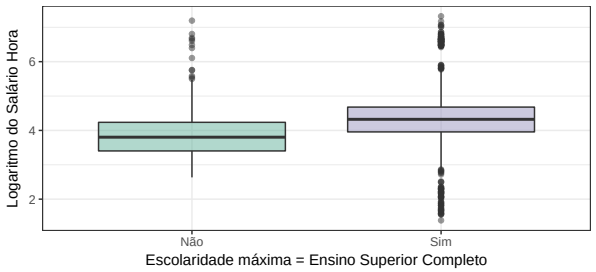
Figura 7 – Boxplots do logaritmo do salário hora por escolaridade máxima para trabalhadores de TI no Estado de Santa Catarina no ano de 2011



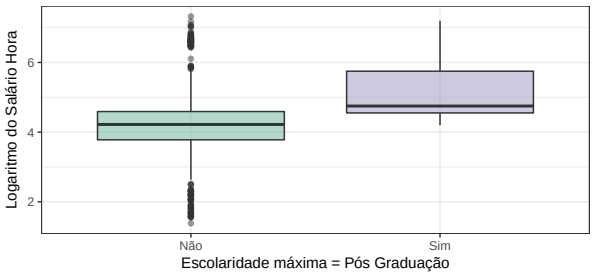
(a) Distribuição do logaritmo do salário hora para trabalhadores que cursaram até o ensino médio versus os demais trabalhadores



(b) Dist. do log. do salário hora para trabalhadores que cursaram até o ensino superior incompleto versus os demais trabalhadores



(c) Dist. do log. do salário hora para trabalhadores que cursaram até o final do ensino superior versus os demais trabalhadores



(d) Distribuição do logaritmo do salário hora para trabalhadores que finalizaram pós graduação versus os demais trabalhadores

Fonte – Elaboração própria com base nos dados da RAIS.