Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Análise da disparidade salarial entre homens e mulheres na Paraíba utilizando modelos para dados categóricos ordinais

Marina Rodrigues de Oliveira

João Pessoa, 2024

Marina Rodrigues de Oliveira

Análise da disparidade salarial entre homens e mulheres na Paraíba utilizando modelos para dados categóricos ordinais

Monografia apresentada ao curso de Ba - charelado em Estatística da Universidade Federal da Paraíba, como requisito funda - mental para obtenção do grau de Bacharel em estatística.

Orientadora: Ana Hermínia Andrade e Silva

Julho de 2024

Catalogação na publicação
Seção de Catalogação e Classificação

O48a Oliveira, Marina Rodrigues de.
Análise da disparidade salarial entre homens e mulheres na Paraíba utilizando modelos para dados categóricos ordinais / Marina Rodrigues de Oliveira. - João Pessoa, 2021.
52 f. : il.

Orientação: Ana Hermínia Andrade e Silva.
TCC (Graduação/Bacharelado em Estatística) - UFPB/CCEN.

1. Regressão. 2. Regressão logística ordinal. 3. Renda média salarial. 4. Sexo - Diferença salarial. I. Silva, Ana Hermínia Andrade e. II. Título.

UFPB/CCEN CDU 519.246.8(043.2)

Elaborado por Josélia Maria Oliveira da Silva - CRB-15/113

Universidade Federal da Paraíba
Centro de Ciências Exatas e da Natureza
Departamento de Estatística

Trabalho de Conclusão de Curso de Bacharelado em Estatística intitulado *Análise da disparidade salarial entre homens e mulheres na Paraíba utilizando modelos para dados categóricos ordinais* de autoria de Marina Rodrigues de Oliveira, aprovada pela banca examinadora constituída pelos seguintes professores:

Profª. Dra. Ana Hermínia Andrade e Silva
Universidade Federal da Paraíba - UFPB

Profª. Dra. Maria Lídia Coco Terra
Universidade Federal da Paraíba - UFPB

Profª. Dra. Tatiene Correia de Souza
Universidade Federal da Paraíba - UFPB

Coordenadora do Departamento de Estatística
Gilmara Alves Cavalcanti
DE/UFPB

João Pessoa, 13 de julho de 2021

DEDICATÓRIA

Eu dedico este trabalho a Deus que me fortaleceu dia após dia durante toda a minha graduação, em nenhum momento me deixando sozinha, sempre enviando pessoas para me ajudar.

AGRADECIMENTOS

Quero agradecer a Deus em primeiro lugar, sempre!

Gostaria de agradecer a minha mãe Regina, que sempre foi tão presente, a cada prova, a cada obstáculo, mesmo de longe não se fez menos presente, seja com palavras de incentivo, seja na parte financeira, se esforçando muito para que não faltasse nada, presente também em cada sorriso pela nova etapa. Agradecer a minha tia Mara que sempre se fazendo presente, além de demonstrar seu apoio, carinho e preocupação.

Agradecer a minha orientadora que esteve disponível para dúvidas, reuniões, sempre prontamente para ajudar e orientar da melhor forma e não olhando apenas para um aluno tentando se formar, mas também como um ser humano que passar por diversas situações, porém tentando de todas as formas seguir não desistindo e sempre utilizou palavras de incentivo que fizeram total diferença nesses meses de bastante dificuldade.

Gleizielle Nayane, Kelfanio Alvez, José Nataniel, Mateus Bitencourt e Ullyssis Rozendo agradecer por toda a paciência, por serem ombros amigos de verdade, quero agradecer pelos estudos em grupo, sempre estiveram a disposição para tirar dúvidas, explicar algo que não entendi e também agradecer pelos momentos de distração reunidos para conversar e jogar, sempre bem humorados. Agradecer aos professores do departamento, cada matéria concluída, cada período concluído não seria possível sem vocês, vocês têm o meu respeito, admiração e gratidão.

Agradecer Giovani Nucci, Manuel Ferreira, Juliana, Kleber Henrique, pessoal do departamento por todo o companheirismo. E agradecer uma queridíssima amiga Denise que sempre acreditou em mim e que a edução é o melhor caminho.

RESUMO

Diante de estatísticas do Instituto Brasileiro de Geografia e Estatística (IBGE), sabemos que em 2019 as mulheres no Brasil receberam 77,7% (cerca de 3/4) do rendimento dos homens, ao passo que o rendimento médio mensal dos homens era de R$2.555,00, o das mulheres era de R$1.985,00. Assim sendo, de suma importância analisar se essa diferença salarial entre os sexos está presente no estado da Paraíba e o quanto essa diferença influencia no salário do trabalhador. Utilizando modelos de regressão, modelos de regressão para dados categóricos, que possuem uma natureza ordinal, pode-se verificar que em todos os modelos testados a variável Sexo foi significativa e a estimativa do coeficiente foi negativa, como a categoria de referência é ser do sexo masculino, então ser do sexo feminino influência negativamente na Renda Média Salarial. Portanto pode-se concluir que existe diferença salarial entre os sexos no estado da paraíba, a renda média salarial do sexo feminino é 0,5951 vezes a renda média salarial do sexo masculino.

Palavras-chave: Regressão, Regressão Logística Ordinal, Renda Média Salarial, Diferença Salarial, Sexo.

ABSTRACT

Observing statistics from the Brazilian Institute of Geography and Statistics (IBGE), we know that in 2019 women received 77.7% (about 3/4) of men's salary, while the average monthly salary of men was R$2,555,00, the women's was R$1,985.00. Therefore, it is extremely important to analyze whether this salary difference between the sexes is present in the state of Paraíba and how much this difference influences the worker's salary. Using regression models, regression models for categorical data, which have an ordinal nature, it can be seen that in all models tested the variable Sex was significant and the coefficient estimate was negative, as the reference category is being of male sex, then being female negatively influences the Average Salary. Therefore, it can be concluded that there is a salary difference between the sexes in the state of Paraíba, the average salary for women is 0.5951 times the average salary for men.

Key-words: Regression, Ordinal Logistic Regression, Average salary, Salary Difference, Gender.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS

ACF - Função de Autocorrelação

AIC - Critério Akaike

BIC - Critério Bayesiano de Schwarz

CAGED – Cadastro Geral de Empregados e Desempregados

IBGE - Instituto Brasileiro de Geografia e Estatística

LGPD - Lei Geral de Proteção de Dados

MLG - Modelo de Regressão Generalizado

MQO - Mínimos Quadrados Ordinários

MTE – Ministério do Trabalho e Emprego

PDET – Programa de Disseminação das Estatísticas do Trabalho

QDPs - Quantidades Descritivas Populacionais

RAIS – Relação Anual de Informações Sociais

Sumário

1 INTRODUÇÃO

Ao observar alguns estudos do Instituto Brasileiro de Geografia e Estatística (IBGE), concatenando informações, estatísticas sociais e estatísticas de gênero, verificou-se que em 2019, as mulheres recebiam 77,7% (cerca de 3/4) do rendimento dos homens. Ao passo que o rendimento médio mensal dos homens era de R$2.555,00, o das mulheres era de R$1.985,00. Essa desigualdade é maior nos grupos ocupacionais com o maior rendimento, tais como diretores e gerentes e nos grupos de profissionais das ciências intelectuais, as mulheres receberam respectivamente 61,9% e 63,6% do rendimento dos homens.

O modelo estatístico utilizado para avaliar a relação de causa e efeito entre uma variável dependente e uma ou mais variáveis independentes (explicativas) é denominado modelo de regressão. Além de analisar a associação entre as variáveis, é possível realizar predições e inferências de modo a entender melhor o comportamento dos dados. Ao utilizar os modelos, algumas suposições devem ser verificadas, dessa maneira na prática é comum estimar variância constante aos erros, isto é, assumir pressupostos acerca da distribuição dos erros, quando estes não estão sendo verificados, resultados imprecisos podem ser obtidos possivelmente gerando interpretações equivocadas do problema em questão. Sendo assim, é preciso estudar novas estratégias inferenciais em modelos lineares de regressão.

Diante dessas estatísticas fica o questionamento se existe essa diferença salarial entre os sexos para o estado da Paraíba e em caso de positivo o quanto essa diferença influencia no salário do trabalhador.

1.1 Objetivos

Este trabalho tem como objetivo analisar a disparidade salarial entre homens e mulheres do estado da Paraíba utilizando modelos de regressão para verificar se existe diferença salarial entre os sexos e se houver avaliar o quanto essa diferença influencia no salário para ambos os sexos, no âmbito populacional do estado da Paraíba em 2016.

Plataforma Computacional

Na análise, manipulação e apresentação gráfica de dados a princípio foi utilizado o ambiente R, que é uma linguagem de programação de alto nível e tem a vantagem de ser distribuída gratuitamente. Para maiores detalhes ver (Ziegel 2003) e ver também http:www.r-project.org. Com base na necessidade de mais memória RAM do que a disponível na máquina de trabalho, devido ao tamanho da base de dados, passou-se a utilizar o *Google Colaboratory*, que é um ambiente virtual na nuvem do *Google* que permite escrever código Python pelo próprio navegador, sem necessidade de nenhuma configuração na

máquina e tem acessos gratuitos a unidade de processamento gráfico GPU's. Para mais detalhes sobre o ambiente do *Google Colaboratory* ver (Bisong 2019) e (Silva 2020).

2 METODOLOGIA

Em modelos para superpopulações supõe que $y_1, ..., y_N$ são a realização conjunta dos vetores aleatórios de $Y_1, ..., Y_N$. A distribuição conjunta de probabilidade de $Y_1, ..., Y_N$ é considerada um modelo de superpopulação marginal. De forma análoga, $x_1, ..., x_N$ pode ser considerada a realização conjunta de vetores aleatórios de $X_1, ..., X_N$. Sabendo que as matrizes de tamanho N x Q formadas com o vetores transpostos das observações das variáveis auxiliares correspondentes à todas as unidades da população x_U e a matriz correspondente X_U formada pelos vetores aleatórios geradores das variáveis auxiliares na população são definidas de forma análoga às matrizes y_U e Y_U. O modelo de superpopulação será denotado por $f(y_i; \theta)$. Desta forma, permitindo a especificação da distribuição conjunta combinando as variáveis da pesquisa e as variáveis auxiliares que aqui será representada por $f(y_i; \theta, \eta)$ a função de densidade de probabilidade conjunta de (y_U, x_U) em que η é o vetor de parâmetros. No caso em que toda a população for pesquisada, os dados observados serão $(y_1, x_1), ..., (y_N, x_N)$. Na hipótese de resposta completa, a única fonte de incerteza seria que $(y_1, x_1), ..., (y_N, x_N)$ é uma realização de $(Y_1, X_1), ..., (Y_N, X_N)$. Desta maneira os dados observados poderiam ser usados para fazer inferência sobre η, ϕ e θ, usando os procedimentos padrões. Então estaremos trabalhando com inferência descritiva para quantidades descritivas da população (QDPs), pois os dados são populacionais (Pessoa e Silva 1998).

2.1 Modelo regressão linear

Dada a importância em conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer sobre outras, uma das áreas da estatística que desperta grande interesse é a área de regressão. Os modelos de regressão destacam-se na análise de possíveis associações entre diferentes variáveis permitindo, assim, discutir o relacionamento entre elas, além de predizer valores para uma variável de interesse. Há uma grande aplicabilidade dos modelos lineares de regressão, sendo estes bastante utilizados nas mais diversas áreas. Para a utilização dos modelos lineares de regressão alguns pressupostos devem ser assumidos, como, por exemplo, a constância da variância (homoscedasticidade) e suposição distribucional dos erros (Correia de Souza 2003) (Silva 2017).

Nos casos em que as suposições estão sendo satisfeitas, utiliza-se o método de Mínimos Quadrados Ordinários - MQO, que consiste em minimizar a soma dos quadrados dos erros, para se obter as estimativas dos parâmetros que indexam o modelo linear, que é um dos objetivos centrais em modelagem de regressão fazer inferências sobre β, pois este vetor de parâmetros representa o efeito dos regressores considerados sobre a média da variável explicada. O modelo de regressão linear múltiplo pode ser expresso da seguinte forma:

$$y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_p x_{ip} + \epsilon_i, i = 1, ...N,$$

em que

- y_i é a i-ésima resposta;
- $x_{i2}...x_{ip}$ são os $p-1$ ($p < N$) regressores que influenciam na média da variável resposta $\mu_i = E(y_i)$;
- $\beta_1, ..., \beta_p$ são os parâmetros desconhecidos do modelo;
- ϵ_i é o i-ésimo erro aleatório.

O modelo de regressão linear múltiplo na forma matricial é dado por:

$$y = X\beta + \epsilon$$

- y é um vetor N x 1 de resposta;
- X é uma matriz $N - p$ ($p < N$) de regressores (posto$(X) = p$);
- β é um vetor p x 1 de parâmetros;
- ϵ_i é um vetor N x 1 de erros aleatórios.

Os pressupostos do modelo são:

(S0) O modelo estimado é o modelo correto;

(S1) $E(\epsilon_0) = 0$, $\forall i$;

(S2) (homoscedasticidade) $var(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$, $0 < \sigma^2 < \infty$, $\forall i$;

(S3) (auto-correlação) $cov(\epsilon_i \epsilon_s) = E(\epsilon_1 \epsilon_s) = 0$ $\forall i \neq s$;

(S4) Os únicos valores de $c_1, c_2, ..., c_p$ tais que $c_1 x_{i1} + c_2 x_{i2} + ... + c_p x_{ip} = 0$ $\forall i$ são $c_1 = c_2 = ... = c_k = 0$, ou seja, as colunas da matriz X são linearmente independentes, i.e., X tem posto completo: posto$(X) = p$ $(< i)$;

(S5) (normalidade) $\epsilon_i \sim$ Normal $\forall i$. Como a relação entre ϵ_i e y_i é linear, por conseguinte $y_i \sim$ Normal. Essa suposição é muitas vezes utilizada para estimação intervalar e testes de hipóteses;

- multicolinearidade: Posto(X) < p. Dizemos que há multicolinearidade exata se $\exists$ $c = (c_1, ..., c_p)' \neq 0$ tal que

$$c_1 x_1 + c_2 x_2 + ... + c_p x_p = 0.$$

Entretanto, em diversas situações práticas, tais pressupostos não são verificados, ou são verificados, mas seguem o modelo sem considerá-los (Correia de Souza 2003) (Silva 2017).

2.2 Modelo de regressão linear generalizado

O modelo de regressão linear generalizado (MLG) é definido como uma extensão do modelo de regressão linear, que foi descrito anteriormente. Considere $\mathbf{y}$ como sendo um vetor de observações contendo N componentes, que é assumido como uma realização de uma variável aleatória $\mathbf{Y}$, em que seus componentes são independentes e identicamente distribuídos com média μ (McCullagh e Nelder 2019). No caso do modelo linear ordinário é dado por:

$$\mu = \sum_{1}^{p} \mathbf{x}_j \beta_j,$$

em que os β's são parâmetros cujos valores são desconhecidos e são estimados a partir dos dados. Seja i indexado nas observações então:

$$E(Y_i) = \mu_i = \sum_{1}^{p} \mathbf{x}_j \beta_j, \qquad i = 1, ..., n,$$

em que x_{ij} é o valor da j-ésima covariável para a observação i. Em notação matricial em que μ é n x 1, $\mathbf{X}$ é n x p e β é p x 1 a medida pode ser expressa da seguinte forma:

$$\mu = \mathbf{X}\beta, \tag{1}$$

em que X é a matriz do modelo e β é o vetor de parâmetros.

Com a utilização dos MLG's é possível considerar outras distribuições de probabilidade além da distribuição Normal para a variável resposta, desde que pertença a família exponencial de distribuições, desta forma podendo flexibilizar a relação funcional entre a média e o preditor linear (Nelder e Wedderburn 1972). Ao redefinir a Equação (1) é possível fazer a transição para o MLG e desta forma produzir a estrutura em três partes específicas:

1. A componente aleatória: o componente de $\mathbf{y}$ é independente e segue distribuição Normal com $E(\mathbf{Y}) = \mu$ e variância constante σ^2;

2. A componente sistemática que define o preditor linear η, pode ser expressa por:

$$\eta = \sum_{1}^{p} x_j \beta_j; \tag{2}$$

3. A função de ligação que relaciona o componente aleatório e a componente sistemática:

$$\mu = \eta. \tag{3}$$

Essa generalização introduz um novo símbolo η para o preditor linear, então especificando dessa forma, μ e η são de fato idênticos na Equação (3). E se escrevermos:

$$\eta = g(\mu_i),$$

essa função $g(.)$ será chamada de função de ligação.

Na formulação do modelo linear, a variável resposta segue uma distribuição Normal na Equação (2) e a função de ligação é a função identidade. A partir do modelo linear generalizado é possível realizar duas extensões; primeira a distribuição na Equação (2) tem origem na família exponencial, porém não fica restrito somente na distribuição Normal; e segundo a função de ligação pode ser qualquer função monótona diferenciável.

Componente aleatória

Seja $y = (y_1, ..., y_N)^T$ um vetor de observações referente às realizações da variável aleatória $Y = (Y_1, ..., Y_N)^T$ independentes e identicamente distribuídas, com vetor de médias $\mu = (\mu_1, ..., \mu_N)$ com função de densidade na forma:

$$f(y_i; \theta_i, \phi) = \exp\{(y_i\theta_i - b(\theta_i))/a(\phi) + c(y_i, \phi)\}, \tag{4}$$

em que algumas funções específicas a(.), b(.) e c(.) são conhecidas, ϕ é o parâmetro de dispersão e θ_i é denominado parâmetro natural ou canônico, que caracteriza a distribuição em (4). Se ϕ é conhecido, a Equação (4) representa a família exponencial uniparamétrica. A função de log-verossimilhança é definida por:

$$l(y_i; \theta_i, \phi) = \frac{\{(y_i\theta_i - b(\theta_i)]\}}{a(\phi)} + c(y_i, \phi). \tag{5}$$

Escrevemos $l(\theta, \phi; y) = logf(y, \theta, \phi)$ para a função de verossimilhança considerando a função de θ e ϕ. A média e a variância de y podem ser derivadas facilmente a partir das

relações conhecidas:

$$E\left(\frac{\partial l}{\partial \theta_i}\right) = 0, \tag{6}$$

e

$$E\left(\frac{\partial^2 l}{\partial \theta_i^2}\right) + E\left(\frac{\partial l}{\partial \theta_i}\right)^2 = 0, \tag{7}$$

da Equação (4) temos que:

$$l(\theta_i; y_i) = \{y_i\theta_i - b(\theta_i)\}/a(\phi) + c(y_i, \phi),$$

uma vez que:

$$\frac{\partial l}{\partial \theta_i} = \{y_i - b'(\theta_i)\}/a(\phi), \tag{8}$$

e

$$\frac{\partial^2 l}{\partial \theta_i^2} = -b''(\theta_i)/a(\phi), \tag{9}$$

em que a diferenciação é feita com respeito à θ.

Pode-se mostrar que $E\left(\frac{\partial l}{\partial \theta_i}\right) = 0$, e de (8) temos que:

$$E\frac{\{y_i - b'(\theta_i)\}}{a(\phi)} = 0,$$

logo,

$$E(Y_i) = \mu_i = b'(\theta_i).$$

Pode-se também mostrar que na Equação (7) a partir de (8) e (9), temos que:

$$E\left(-\frac{b''(\theta_i)}{a(\phi)}\right) + E\left(\frac{y_i - b'(\theta_i)}{a^2(\phi)}\right)^2 = 0,$$

$$-\frac{b''(\theta_i)}{a(\phi)} + \frac{1}{a(\phi)}E(y_i - E(y_i))^2 = 0,$$

$$\frac{1}{a(\phi)}Var(Y_i) = \frac{b''(\theta_i)}{a(\phi)},$$

então $var(Y) = b''(\theta_i)a(\phi)$, que pode ser também escrita na forma $Var(Y_i) = a(\phi)V_i$, em que $V_i = \dfrac{d\mu}{d\theta_i}$ é chamada de variância.

Componente sistemática

A componente sistemática é formada pela estrutura linear de um modelo de regressão em que $\eta = X\beta$, $\eta = (\eta_1, ..., \eta_N)^T$, $\beta = (\beta_, ..., \beta_p)^T$ e X é uma matriz modelo de dimensão n x p ($p < n$) conhecida, de posto p. A função linear η dos parâmetros

desconhecidos β é chamada de preditor linear e corresponde à parte sistemática de um MLG.

Funções de ligação

A função de ligação relaciona o preditor linear η ao valor esperado μ de um dado y, esta por sua vez relaciona o componente aleatório á componente sistemática. No caso do modelo linear a média e o preditor linear são iguais, pois a ligação é a função identidade. É plausível que ambos η e μ podem assumir qualquer valor na reta real. Entretanto, quando estamos lidando com contagens, como na distribuição Poisson, o $\mu > 0$, logo a ligação identidade não é tão atrativa, em partes porque η pode assumir valores negativos, enquanto μ não pode assumir valores abaixo de 0. Logo,

$$\mu 1 - k = g^{-1}(\eta_k) \quad ou \quad \eta = g(\mu_k) \quad k = 1, ..., N,$$

sendo que $g(.)$ é uma função monótona e diferenciável.

2.3 Estatísticas suficientes e ligação canônica

Cada uma das distribuições que anteriormente foram citadas como importantes têm uma função de ligação especial para qual existe uma estatística suficiente igual a dimensão de β no preditor linear $\eta = \sum x_j \beta_j$. Essa ligação canônica como são chamadas ocorrem quando:

$$\theta = \eta,$$

em que θ é o parâmetro canônico. As ligações canônicas para as principais distribuições estão disponíveis na Tabela 1:

Tabela 1: Principais distribuições dos MLG's e suas respectivas ligações canônicas

Distribuições	Ligações canônicas
Normal	$\eta = \mu$
Poisson	$\eta = log(\mu)$
Binomial	$\eta = log(\pi)/(1-\pi)$
Gama	$\eta = \mu^{-1}$
Inversa Gaussiana	$\eta = \mu^{-2}$

Para as ligações canônicas, a estatística suficiente é $X^T V$ em notação de vetor, com componentes:

$$\eta = \sum_i x_{ij} Y_j, \qquad j = 1, ..., p,$$

em que o somatório é feito sobre as observações.

2.4 Modelos para dados categóricos ordinais

Modelos de regressão logística ordinal

Em geral, pelo uso das classes de modelos de regressão para dados ordinais, é possível trabalhar com dados que tenham natureza ordinal. Dois modelos em particular, as probabilidades proporcionais e os modelos de risco proporcionais são mais utilizados na prática devido a simplicidade de sua interpretação. Esses modelos são mostrados como extensões multivariadas de modelos lineares generalizados (McCullagh 1980).

Quando a variável resposta y possui uma ordenação entre as suas categorias, o uso do modelo logístico para respostas ordinais tem interpretações mais simples. A regressão logística para respostas ordinais se baseia no uso da probabilidade acumulada de Y. Dessa forma, a probabilidade considerada agora, é a de que o valor de Y recai numa faixa de interesse, j. Então, dada uma categoria j de interesse:

$$P(Y \leq j) = \pi_1 + ... + \pi_j, \quad j = 1, ..., k,$$

a probabilidade acumulada reflete a ordenação entre as categorias da variável dependente. Sucede-se que:

$$P(Y \leq 1) \leq P(Y \leq 2) \leq ... \leq P(Y \leq j) = 1.$$

Os *logits* para probabilidade acumulada são:

$$logit[P(Y \leq j)] = ln\left[\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right] = ln\left[\frac{\pi_1 + ... + \pi_j}{\pi_{j+1} + ... + \pi_k)}\right] \quad j = 1, ..., k-1. \quad (10)$$

De acordo com (Agresti 2018), um modelo para *logit* acumulativo se parece com um modelo de regressão binária, na qual as categorias de 1 a j se combinam para formar uma única categoria e as outras $j+1$ a k formam uma segunda categoria. Para apenas um variável independente x, tal modelo *logit* cumulativo pode ser escrito da seguinte forma:

$$logit[P(Y \leq j)] = \beta_{0j} + \beta_1 x, \quad j = 1, ..., k-1, \quad (11)$$

Na Equação (11), β não possui um índice j, indicando que o efeito da variável x é descrito por apenas um parâmetro para todas as categorias. Neste modelo, o intercepto é o parâmetro que diferencia o modelo de uma categoria a outra, como se pode perceber na equação acima o índice j no parâmetro β_0. Segundo (Ananth e Kleinbaum 1997), o modelo é invariante quando a codificação das categorias é invertida (a k-ésima categoria

passa a ser a primeira, a primeira passa a ser a j-ésima, a segunda passa a ser a penúltima e assim por diante). (Agresti 2018), afirma que nesse caso porém, os sinais dos β's ficam invertidos. Incorporando de forma separada os argumentos do modelo $\beta_j's$ e então aplicar o método convencional de verossimilhança para avaliar a hipótese:

$$H_0 : \beta_j = \beta, \qquad j = 1, ..., k-1;$$

$$H_1 : \beta_j \neq \beta, \qquad j = 1, ..., k-1.$$

Modelos para dados categóricos ordinais

No modelo de regressão logístico ordinal *logit* a probabilidade de se observar uma classe inferior ou igual a k do conjunto das K classes da variável dependente, para um determinado vetor de observações das variáveis independentes X, é dada por:

$$P(Y_j \leq k|x) = \pi_1 + \pi_2 + ...\pi_k, \quad (j = 1, ..., N; k = 1, ..., K),$$

em que,

$$\pi_1 = P(Y_j = 1), \pi_2 = P(Y_j = 2), ..., \pi_k = P(Y_j = k).$$

(Marôco 2018)

Fazendo a analogia ao modelo de regressão logística, *Logit* $[P(Y_j \leq k)]$, que é:

$$Logit[P(Y_j \leq k|x)] = Ln\left(\frac{P(Y_j \leq k|x)}{1 - P(Y_j \leq k|x)}\right) = \tag{12}$$

$$Logit[P(Y_j \leq k|x)] = Ln\left(\frac{P(Y_j \leq k|x)}{P(Y_j > k|x)}\right) =$$
$$= \alpha_k + X^*\beta \quad (k = 1, ..., K-1) =$$

$$= e^{-d\beta}. \tag{13}$$

em que, α_k representa o parâmetro de locação das $(k = 1, ..., K-1)$, β é o vetor dos coeficientes de regressão e X^* representa é a matriz das variáveis independente.

O razão de chances acumuladas, igual para todas as classes, é dada por:

$$OR_k = \frac{P(Y \leq k|x = x + d)/P(Y > k|x = x + d)}{P(Y \leq k|x = x)/P(Y > k|x = x)} \tag{14}$$

O razão de chances é β-proporcional à distância d entre os dois pontos da variável independente, se $d = 1$, as chances de observar uma classe inferior ou igual a k.

Assumindo que existe uma variável continua (η), e que a variável Y que resulta do corte em K-classes ordinais de η. O modelo estrutural da relação é dado por:

$$\eta_j = x_j\beta + \epsilon_j \quad (j = 1, ..., N), \tag{15}$$

em que

$$Y_j = k \quad se \quad \alpha_{k-1} \leq \eta \leq \alpha_k. \tag{16}$$

Existem vários modelos de regressão logística usados quando a resposta possui ordenação, tais como o modelo de regressão *probit* que supõe que a variável dependente é de tipo normal (Della Lucia et al. 2013) e o modelo de regressão ordinal com função de ligação *log-log* (Marôco 2018).

De uma forma linear generalizada, recorrendo à função de ligação, o modelo pode escrito como:

$$Link(P[Y \leq k]) = \alpha_k - X^*\beta \tag{17}$$

com possíveis funções de ligação:

Função de ligação	
Probit	$\Phi^{-1}(P[Y \leq K])$
Logit	$Ln\left[\frac{P[Y \leq k]}{P[Y > k}\right]$
Log-log	$Ln(-Ln(1 - P[Y \leq k]))$

2.5 Seleção de variáveis

O teste de t é um dos testes apropriados para a seleção de variáveis. Calculando as médias e as variâncias dos coeficientes obtidos em cada regressão estimada, temos:

$$\hat{\beta}_0 = \sum_{j=1}^{k} \frac{\hat{\beta}_{0j}}{j},$$

$$\hat{\beta} = \sum_{j=1}^{k} \frac{\hat{\beta}_j}{j},$$

$$\sigma^2(\hat{\beta}_0) = \sum_{j=1}^{k} \frac{(\hat{\beta}_{0j} - \hat{\beta}_0)^2}{j^2},$$

$$\sigma^2(\hat{\beta}) = \sum_{j=1}^{k} \frac{(\hat{\beta}_j - \hat{\beta})^2}{j^2},$$

então, podemos definir a estatística de teste, para cada parâmetro, como sendo:

$$t_{\beta_0} = \left(\frac{\hat{\beta}_0}{\sigma(\hat{\beta}_0)} \right) / \sqrt{N},$$

$$t_{\beta} = \left(\frac{\hat{\beta}}{\sigma(\hat{\beta})} \right) / \sqrt{N}.$$

2.6 Seleção do melhor modelo

O Critério Akaike (AIC) é um método proposto por Akaike em 1974 tem como ideia básica selecionar um modelo que seja parcimonioso, ou seja, que esteja bem ajustado e tenha um número reduzido de parâmetros. Como o logaritmo da função de verossimilhança L(β) cresce com o aumento do número de parâmetros do modelo, uma proposta razoável é encontrar o modelo com menor valor para a função:

$$AIC = 2L(\hat{\beta}) + 2p,$$

em que p denota o número de parâmetros. Em função do desvio do modelo, tem-se:

$$AIC = D(y; \hat{\mu}) + 2p,$$

em que $D(y; \hat{\mu})$ denota o desvio do modelo e p o número de parâmetros.

O Critério bayesiano de Schwarz (BIC) é o critério de Informação Bayesiano, proposto por Schwarz em 1978 é dado por:

$$BIC = 2log f(x_N|\theta) + p \log N,$$

em que $f(x_N|\theta)$ é o modelo escolhido, p é o número de parâmetros a serem estimados e N é o número de observações da população.

3 RESULTADOS

3.1 Análise dos dados

Nesta seção, serão apresentadas estatísticas descritivas das variáveis utilizadas para realizar a modelagem da Renda Média Salarial do trabalhador no estado da Paraíba, com intuito de analisar a disparidade salarial entre homens e mulheres em 2016. Estão disponíveis ao público apenas dados até 2016, pois uma lei promulgada em 14 de agosto de 2018 em atenção às disposições da Lei Geral de Proteção de Dados (Lei 13.709/2018 – "LGPD") e do artigo 31 da Lei de Acesso à Informação que restringiu o acesso a dados pessoais constantes nas referidas bases governamentais. (Ramos e Gomes s.d.).

Neste cenário, o banco de dados foi montado a partir da Relação Anual de Informações Sociais (RAIS) e o Cadastro Geral de Empregados e Desempregados (CAGED) no âmbito de trabalho e rendimento - pessoas de 18 anos ou mais de idade no ano de 2016 para o estado da Paraíba. Um conjunto de 819849 informações que estão descritas por seis variáveis na Tabela 2.

Apesar da variável Renda Média Salarial inicialmente ser uma variável contínua, o salário do trabalhador em reais, no banco de dados da RAIS está disponível de forma categórica. A primeira categoria refere-se a pessoa que recebe até meio salário mínimo (i.e., em média recebe até R$440,00 por mês, pois no ano de referência o salário mínimo era R$880,00), a segunda categoria refere-se a pessoas que recebem entre meio e um salário mínimo e assim por diante até a ultima categoria referente a pessoas que recebem mais que 20 salários mínimos, a variável possui 12 categorias, que estão disponíveis na Tabela 2. Observando uma análise descritiva da proporção de pessoas em cada categoria da variável Renda na Tabela 3 percebemos que a proporção de mulheres analfabetas e que não concluíram até 5º ano do ensino fundamental é maior que a dos homens no estado da Paraíba em 2016. Em todos os testes o nível de significância adotado foi de 5%.

Tabela 2: Identificação das variáveis do banco de dados em estudo.

Variável	Descrição
Renda Média Salarial	1 = até meio salário mínimo;
	2 = meio até 1 salário mínimo;
	3 = 1 até 1 e meio salários mínimo;
	4 = 1 e meio até 2 salário mínimo;
	5 = 2 até 3 salários mínimo;
	6 = 3 a 4 salários mínimo;
	7 = 4 a 5 salários mínimo;
	8 = 5 a 7 salários mínimo;
	9 = 7 a 10 salários mínimo;
	10 = 10 a 15 salários mínimo;
	11 = 15 até 20 salários mínimo;
	12 = mais de 20 salários mínimo.
Sexo	1 = Masculino, 2 = Feminino.
Faixa Etária	1 = até 17 anos;
	2 = 18 a 24 anos;
	3 = 25 a 29 anos;
	4 = 30 a 39 anos;
	5 = 40 a 49 anos;
	6 = 50 a 59 anos;
	7 = 60 a 64 anos;
	8 = 65 anos ou mais;
	99 = Ignorado.
Grau de Escolaridade	1 = analfabeto;
	2 = até o 5º ano incompleto do ensino fundamental;
	3 = 5º ano completo do ensino fundamental;
	4 = do 6º ao 9º ano incompleto do ensino fundamental;
	5 = ensino fundamental completo;
	6 = ensino médio incompleto;
	7 = ensino médio completo;
	8 = educação superior incompleta;
	9 = educação superior completa;
	10 = mestrado completo;
	11 = doutorado completo;
	99 = ignorado.
Qnt de Horas Trabalhadas Semanalmente	
Raça	1 = Indígena;
	2 = Branca;
	4 = Preta/Negra;
	6 = Amarela;
	8 = Parda;
	99 = Ignorado.

Tabela 3: Porcentagem da quantidade de pessoas nas categorias da variável Renda Média Salarial por Sexo.

	1	2	3	4	5	6	7	8	9	10	11	12
Masculino	0,68	11,20	48,64	15,87	12,05	3,59	1,69	2,29	1,55	1,32	0,59	0,54
Feminino	0,85	15,12	47,03	11,10	12,30	5,32	2,36	2,31	1,46	1,34	0,50	0,31
Total	0,75	12,81	47,98	13,91	12,15	4,30	1,96	2,30	1,51	1,33	0,55	0,45

Tabela 4: Porcentagem da quantidade de pessoas nas categorias da variável Raça/Cor.

	Indígena	Branca	Preta/Negra	Amarela	Parda	Não Respondido
Masculino	0,137	20,449	2,858	0,759	44,168	31,629
Feminino	0,094	17,771	1,359	0,547	29,108	51,121
Total	0,120	19,349	2,242	0,672	37,981	39,637

Tabela 5: Porcentagem da quantidade de pessoas nas categorias da variável Faixa Etária.

	1	2	3	4	5	6	7	8	99
Masculino	0,003	0,175	14,125	16,045	32,671	20,661	14,907	1,412	0,001
Feminino	0,003	0,152	11,728	15,089	31,470	22,171	18,051	1,335	0,001
Total	0,003	0,166	13,140	15,652	32,178	21,282	16,198	1,380	0,001

Tabela 6: Porcentagem da quantidade de pessoas nas categorias da variável Escolaridade.

	1	2	3	4	5	6	7	8	9	10	11
Masculino	1.618	6.604	3.701	6.958	13.745	6.528	43.954	2.650	13.643	0.402	0.197
Feminino	0.167	1.117	1.043	2.752	13.159	3.465	45.711	4.038	27.499	0.757	0.292
Total	1.022	4.350	2.609	5.230	13.504	5.270	44.676	3.221	19.335	0.548	0.236

3.2 Modelo de regressão linear

Primeiramente foram propostos modelos de regressão univariados, considerando Renda como variável resposta e as variáveis Sexo, Faixa Etária, Grau de Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor como variáveis independentes. No modelo univariado considerando Renda como variável resposta sendo explicada pela variável Sexo, a mesma foi significativa e a estimativa dos coeficientes do modelo e seus respectivos p-valores estão disponíveis na Tabela 7.

No modelo univariado considerando Renda como variável resposta e a variável Faixa Etária como variável independente apenas as classes Faixa Etária entre 18 a 24 anos e idade ignorada não foram significativas para o modelo com p-valores 0,9373 e 0,8030 respectivamente e a estimativa dos coeficientes estão na Tabela 8.

No modelo univariado considerando Renda como variável resposta sendo explicada pela variável Grau de Escolaridade, todas as categorias foram significativas como é possível observar na Tabela 9. No modelo univariado com a variável Raça/Cor apenas a pessoa se declarar branca e amarela não foram significativas para o modelo com p-valores 0,9538 e 0,1449 respectivamente e as estimativa dos coeficientes estão disponíveis na Tabela 11.

Em seguida foi proposto um modelo de regressão linear múltiplo considerando

Tabela 7: Estimativas dos coeficientes, p-valor do modelo univariado com a variável Sexo

	estimativa	p-valor	
(Intercepto)	4,422	$2\text{x}10^{-16}$	***
Sexo 2	$-0,565$	$2\text{x}10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Tabela 8: Estimativas dos coeficientes, p-valor do modelo univariado com a variável Faixa Etária

	estimativa	p-valor	
(Intercepto)	1,3324	$1{,}51\text{x}10^{-05}$	***
Faixa Etária 3	1,8136	$4{,}86\text{x}10^{-09}$	***
Faixa Etária 4	2,3218	$6{,}41\text{x}10^{-14}$	***
Faixa Etária 5	2,7213	$2\text{x}10^{-16}$	***
Faixa Etária 6	3,1423	$2\text{x}10^{-16}$	***
Faixa Etária 7	3,9273	$2\text{x}10^{-16}$	***
Faixa Etária 8	5,4294	$2\text{x}10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Tabela 9: Estimativas dos coeficientes, p-valor do modelo univariado com a variável Grau de Escolaridade

	estimativa	p-valor	
(Intercepto)	47,200	$2\text{x}10^{-16}$	***
Escolaridade 2	$-0,9227$	$1{,}63\text{x}10^{-11}$	***
Escolaridade 3	$-1,1840$	$4{,}09\text{x}10^{-16}$	***
Escolaridade 4	$-1,4342$	$2\text{x}10^{-16}$	***
Escolaridade 5	$-1,9191$	$2\text{x}10^{-16}$	***
Escolaridade 6	$-1,7500$	$2\text{x}10^{-16}$	***
Escolaridade 7	$-2,0941$	$2\text{x}10^{-16}$	***
Escolaridade 8	$-1,8166$	$2\text{x}10^{-16}$	***
Escolaridade 9	$-0,7117$	$1{,}81\text{x}10^{-08}$	***
Escolaridade 10	0,8578	$4{,}52\text{x}10^{-05}$	***
Escolaridade 11	1,7450	$1{,}40\text{x}10^{-09}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Tabela 10: Estimativas dos coeficientes, p-valor do modelo univariado com a variável Quantidade de Horas Trabalhadas

	estimativa	p-valor	
(Intercepto)	5,286	$2\text{x}10^{-16}$	***
Quantidade de horas trabalhadas	$-0,027$	$2\text{x}10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Tabela 11: Estimativas dos coeficientes, p-valor do modelo univariado com a variável Raca/Cor

	estimativa	p-valor	
(Intercepto)	4,736	$2\text{x}10^{-16}$	***
Raça 4	$-0,396$	$7{,}56\text{x}10^{-06}$	***
Raça 8	$-0,295$	$2\text{x}10^{-16}$	***
Raça 9	$-0,909$	$2\text{x}10^{-16}$	***
Raça 99	$-1,114$	$2\text{x}10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Renda como variável resposta e como variáveis independentes: Sexo, Faixa Etária, Grau de Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, adotando o nível de significância de 5%. Podemos observar os coeficientes estimados do modelo e seus respectivos p-valores na Tabela 12.

Para este modelo final procedemos uma análise residual e de diagnóstico para testar a normalidade foi realizado o teste de Lilliefors, obtendo p-valor de $< 2,2\text{x}10^{-16}$. Dessa forma, a 5% de significância, a hipótese nula foi rejeitada, ou seja, há evidências para afirmar que a suposição de normalidade dos erros está sendo violada. Para uma compreensão visual observe a Figura 1. Para analisar a suposição de homoscedasticidade foram aplicados os testes de Breush-Pagan e de Goldfeld-Quandt, obtendo p-valores $< 2,2\text{x}10^{-16}$, ao nível de significância de 5% a hipótese de homoscedasticidade foi rejeitada, ou seja, os erros apresentam heteroscedasticidade. Também é possível observar a Função de Autocorrelação (ACF) na Figura 1 que os pontos do gráfico não estão distribuídos de forma aleatória, logo temos evidências para afirmar que a variância dos erros do modelo não é constante. Ao observar a Figura 1 alguns lag's (defasagens) ultrapassam os limites do intervalo de confiança, apresentando correlações acima de 0,4. Desse modo, concluímos que a suposição de autocorrelação dos erros está sendo violada.

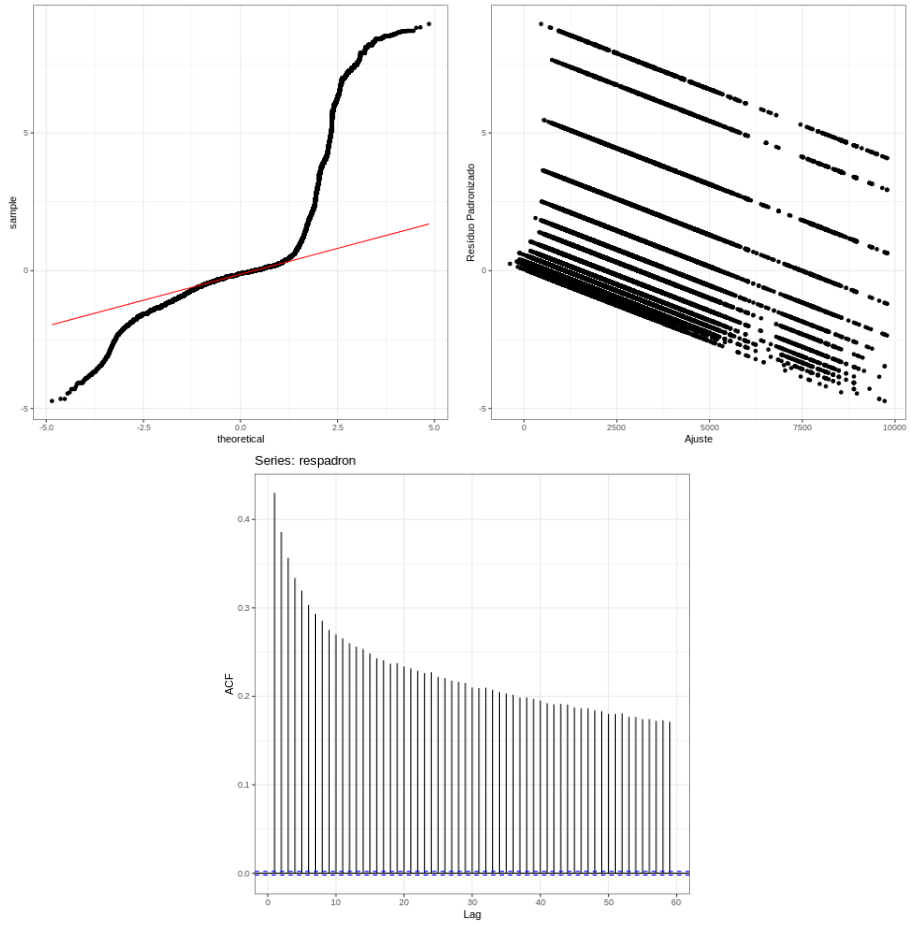
Então, ao realizar a análise de resíduos do modelo de regressão normal linear, a suposição ($S5$) de normalidade dos erros foi violada como visto no teste de hipóteses realizado de Lilliefros, além da análise gráfica feita a partir do Q-QPlot. A variância dos erros não é constante, ou seja, a suposição de homoscedasticidade ($S2$) também foi violada, como visto nos testes de Breush-Pagan e de Goldfeld-Quandt. De acordo com o gráfico ACF, a autocorrelação está presente, pois todos os lag's ultrapassam o intervalo de confiança. O poder explicativo do modelo foi baixo, já que o $R^2 = 0,2074$, isto é, o modelo explica apenas $20,74\%$ da variabilidade total dos dados. Podemos concluir então que este modelo é pouco útil para explicar os dados.

Tabela 12: Estimativa dos coeficientes, p-valor do modelo final de regressão normal linear.

	estimativa	p-valor	
(Intercepto)	$-66,499$	0,0188	*
Sexo 2	$-408,213$	$2\text{x}10^{-16}$	***
Faixa Etária 3	$-71,832$	$2\text{x}10^{-16}$	***
Faixa Etária 5	217,710	$2\text{x}10^{-16}$	***
Faixa Etária 6	415,184	$2\text{x}10^{-16}$	***
Faixa Etária 7	974,671	$2\text{x}10^{-16}$	***
Faixa Etária 8	1803,318	$2\text{x}10^{-16}$	***
Escolaridade 2	203,877	$2\text{x}10^{-16}$	***
Escolaridade 3	189,746	$5{,}61\text{x}10^{-15}$	***
Escolaridade 4	276,125	$2\text{x}10^{-16}$	***
Escolaridade 5	785,250	$2\text{x}10^{-16}$	***
Escolaridade 6	377,473	$2\text{x}10^{-16}$	***
Escolaridade 7	572,048	$2\text{x}10^{-16}$	***
Escolaridade 8	1096,990	$2\text{x}10^{-16}$	***
Escolaridade 9	2401,145	$2\text{x}10^{-16}$	***
Escolaridade 10	4119,390	$2\text{x}10^{-16}$	***
Escolaridade 11	7042,073	$2\text{x}10^{-16}$	***
Qtd Hora Contr	15,976	$2\text{x}10^{-16}$	***
Raça Cor 2	91,737	$1{,}54\text{x}10^{-12}$	***
Raça Cor 8	27,279	0,0291	*
Raça Cor 99	302,349	$2\text{x}10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Figura 1: QQ-plot dos resíduos padronizados, o ajuste x os resíduos padronizados e ACF dos resíduos padronizados do modelo linear



3.3 Modelos para dados ordinais

Depois dos modelos de regressão linear, foi proposto um modelo de regressão ordinal com função de ligação *probit* considerando Renda como variável resposta e como variáveis independentes: Sexo, Faixa Etária, Grau de Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, adotando o nível de significância de 5%, apenas a categoria 2 variável Faixa Etária (18 a 24 anos) não foi significativa para o modelo com p-valor $0,7422$.

Ao observar a Tabela 13 podemos ver que o modelo de regressão *probit* conseguiu estimar apenas as categorias 2, 3, 4, 10 e 12 da variável Renda e sua porcentagem de acerto foi de 48,45%, logo errou em 51,55% das estimativas.

Ao observar a Tabela 15 podemos ver que o modelo de regressão *log-log* conseguiu estimar apenas as categorias 1, 2, 3, 4 e 8 da variável Renda e sua porcentagem de acerto foi de 48,21%, logo errou em 51,79% das estimativas.

Observando a Tabela 17 o modelo com melhor desempenho foi o modelo de Regressão Logística Ordinal com valores de AIC e BIC respectivamente iguais a 2485144,52

Tabela 13: Matrix de confusão modelo de regressão ordinal probit

	2	3	4	10	12
1	480	5462	195	0	1
2	1408	100958	2602	2	20
3	478	384164	8689	2	34
4	32	108076	5924	0	32
5	15	88058	11504	4	46
6	2	27726	7511	2	48
7	1	11651	4392	4	35
8	2	13547	5202	4	68
9	0	7521	4741	5	121
10	0	4869	5683	4	351
11	0	1731	2577	5	188
12	0	1500	2057	6	109

Tabela 14: Estimativa da exponencial dos coeficientes do modelo final de regressão ordinal com função de ligação *probit*.

	β	*odds*	p-valor	
Sexo 2	$-0,2788$	1,3188	$2x10^{-16}$	***
Faixa etária 3	0,7826	0,4572	$2x10^{-16}$	***
Faixa etária 4	1,0799	2,9444	$2x10^{-16}$	***
Faixa etária 5	1,2885	3,6273	$2x10^{-16}$	***
Faixa etária 6	1,4067	4,0825	$2x10^{-16}$	***
Faixa etária 7	1,6646	5,2835	$2x10^{-16}$	***
Faixa etária 8	1,9351	6,9247	$2x10^{-16}$	***
Faixa etária 99	1,3843	3,9920	$2x10^{-16}$	***
Escolaridade 2	0,2275	0,7965	$2x10^{-16}$	***
Escolaridade 3	0,1937	0,8239	$2x10^{-16}$	***
Escolaridade 4	0,2675	0,7653	$2x10^{-16}$	***
Escolaridade 5	0,4955	0,6092	$2x10^{-16}$	***
Escolaridade 6	0,2506	0,7783	$2x10^{-16}$	***
Escolaridade 7	0,4600	0,6313	$2x10^{-16}$	***
Escolaridade 8	0,8690	0,4194	$2x10^{-16}$	***
Escolaridade 9	1,5795	4,8525	$2x10^{-16}$	***
Escolaridade 10	2,1948	8,9782	$2x10^{-16}$	***
Escolaridade 11	2,6419	14,0398	$2x10^{-16}$	***
Quantidade Hora Trabalhadas	0,0189	0,9812	$2x10^{-16}$	***
Raça/Cor 2	0,1270	0,8807	0,0002	***
Raça/Cor 4	0,0788	0,9242	0,0248	*
Raça/Cor 6	0,0934	0,9108	0,0119	*
Raça/Cor 8	0,1466	0,8636	$2x10^{-16}$	***
Raça/Cor 99	0,1877	0,8289	$2x10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

Tabela 15: Matrix de confusão modelo de regressão ordinal *log-log*

	1	2	3	4	8
1	82	1002	4980	74	0
2	137	2035	101988	828	2
3	34	907	390165	2261	0
4	0	213	112337	1514	0
5	0	75	96580	2971	1
6	0	15	32966	2308	0
7	0	9	14463	1611	0
8	0	6	16664	2152	1
9	0	2	10288	2095	3
10	0	0	8190	2708	9
11	0	0	3062	1436	3
12	0	0	2439	1230	3

Tabela 16: Estimativa da exponencial dos coeficientes do modelo final de regressão ordinal com função de ligação *log-log*.

	β	*odds*	p-valor	
Sexo 2	$-0,2225$	1,2492	$2x10^{-16}$	***
Faixa Etária 3	0,4494	0,6380	$2x10^{-16}$	***
Faixa Etária 4	0,6574	0,5182	$2x10^{-16}$	***
Faixa Etária 5	0,8207	0,4401	$2x10^{-16}$	***
Faixa Etária 6	0,9096	0,4026	$2x10^{-16}$	***
Faixa Etária 7	1,0876	2,9671	$2x10^{-16}$	***
Faixa Etária 8	1,2612	3,5296	$2x10^{-16}$	***
Faixa Etária 99	1,0505	2,8591	$2x10^{-16}$	***
Escolaridade 2	0,1901	0,8269	$2x10^{-16}$	***
Escolaridade 3	0,1813	0,8342	$2x10^{-16}$	***
Escolaridade 4	0,2298	0,7947	$2x10^{-16}$	***
Escolaridade 5	0,2972	0,7429	$2x10^{-16}$	***
Escolaridade 6	0,2191	0,8032	$2x10^{-16}$	***
Escolaridade 7	0,3791	0,6844	$2x10^{-16}$	***
Escolaridade 8	0,6557	0,5191	$2x10^{-16}$	***
Escolaridade 9	1,3280	3,7735	$2x10^{-16}$	***
Escolaridade 10	1,8815	6,5633	$2x10^{-16}$	***
Escolaridade 11	2,5317	12,5748	$2x10^{-16}$	***
Quantidade Hora Trabalhadas	0,0293	0,9711	$2x10^{-16}$	***
Raça/Cor 2	0,1409	0,8686	$2x10^{-16}$	***
Raça/Cor 4	0,1671	0,8461	$2x10^{-16}$	***
Raça/Cor 6	0,1784	0,8366	$2x10^{-16}$	***
Raça/Cor 8	0,2196	0,8028	$2x10^{-16}$	***
Raça/Cor 99	0,2802	0,7556	$2x10^{-16}$	***

Significância: 0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

e 2485574. O segundo melhor modelo neste critério foi modelo de regressão com função de ligação *probit* e o modelo com pior desempenho foi o modelo de regressão linear clássico, como esperado, já que dos quatro modelos testados é o único que não comporta a característica ordinal da variável resposta y Renda Média Salarial.

Note que em todos os modelos que a variável Sexo foi significativa a estimativa do coeficiente foi negativa. Ccomo a categoria de referência é ser do sexo masculino, então ser do sexo feminino influência negativamente na Renda Média Salarial.

Observando a Tabela 17 podemos ver as taxas de acerto e erro na previsão das classes pertencentes a variável Renda Média Salarial utilizando os modelos *probit*, *logit* e *log-log*. Nota-se que o modelo de regressão logística ordinal (*logit*) teve o melhor desempenho, com a maior taxa de acerto, cerca de 49,65% e consequentemente a menor taxa de erro.

Tabela 17: AIC, BIC, taxas de acertos e erros dos modelos propostos.

Modelo	AIC	BIC	Acerto	Erro
Normal linear	14930360	14717893	-	-
Probit	2496472	2496902	0,4845	0,5155
Logit	2485144	2485574	0,4965	0,5035
Log-log	2576875	2574462	0,4821	0,5179

Análise do modelo com melhor

Na Tabela 18 são mostrados os coeficientes, a significância de cada variável no modelo final, adotando o nível de significância de $0,05$, todas as variáveis e todas classes também foram significativas sendo elas: Sexo, Faixa Etária, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor.

No modelo de regressão logística ordinal a interpretação é dada da seguinte forma:

- Fixadas as variáveis Faixa Etária, Escolaridade, Quantidade de Horas Trabalhadas, Raça/Cor, $exp(-(-0,5189)) = 1{,}6801$, ou seja uma pessoa do sexo feminino recebe 68,01% do salário da pessoa do Sexo masculino;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, $exp(-(-0,1437) = 1{,}1545$, ou seja, o indivíduo que pertence a Faixa Etária 2, tem entre 18 a 24 anos a 15,15% de chance de receber até meio salario do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, o indivíduo que pertence a Faixa Etária 3, ou seja tem entre 25 a 29 anos tem

$exp(1,4900) = 4{,}3771$ vezes a chance de receber até meio salário mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, o indivíduo que pertence a Faixa Etária 4, ou seja tem entre 30 a 39 anos tem $exp(2,0139) = 7{,}49$ vezes a chance de receber até meio salario mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, o indivíduo que pertence a Faixa Etária 5, ou seja tem entre 40 a 49 anos tem $exp(2,3847) = 10{,}8558$ vezes a chance de receber até meio salário mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas e Raça/Cor, o indivíduo que pertence a Faixa Etária 6, ou seja tem entre 50 a 59 anos tem $exp(2,5927) = 13{,}3658$ vezes a chance de receber até meio salário mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas, Raça/Cor, o indivíduo que pertence a Faixa Etária 7, ou seja tem entre 60 a 64 anos tem $exp(3,0351) = 20{,}8030$ vezes a chance de receber até meio salário mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Escolaridade, Quantidade de Horas Trabalhadas, Raça/Cor, o indivíduo que pertence a Faixa Etária 9, ou seja tem entre 65 anos ou mais tem $exp(3,5758) = 35{,}7232$ vezes a chance de receber até meio salário mínimo do que uma pessoa que tem até 17 anos;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 2 da variável grau de Escolaridade, ou seja, tem até o 5^{o} ano incompleto do ensino fundamental tem $exp(0,3954) = 1{,}4850$, 48,50% de chance de receber até meio salário mínimo do que uma pessoa analfabeta;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 3 da variável grau de Escolaridade, ou seja, tem até 5^{o} ano completo do ensino fundamental tem $exp(0,3412) = 1{,}4066$, 40,66% de chance de receber até meio salário mínimo quando comparado a uma pessoa analfabeta;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 4 da variável grau de Escolaridade, ou seja, tem até do 6^{o} ao 9^{o} ano incompleto do ensino fundamental tem $exp(0,4693) = 1{,}5989$,

59,89% chance de receber até meio salário mínimo quando comparado a uma pessoa analfabeta;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 5 da variável grau de Escolaridade, ou seja, tem até ensino fundamental completo tem $exp(0,7762) = 2{,}1732$ vezes chance receber até meio salário mínimo quando comparado a uma pessoa analfabeta;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 6 da variável grau de Escolaridade, ou seja, tem até ensino médio incompleto tem $exp(0,4345) = 1{,}5442$, 54,42% vezes a chance de receber até meio salário mínimo quando comparado a uma pessoa analfabeta;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a pessoa que tem a classificação 7 da variável grau de Escolaridade, ou seja, tem até ensino médio completo de receber até meio salário mínimo é $exp(0,8039) = 2{,}2342$ vezes a chance de uma pessoa analfabeta receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a chance de uma pessoa que tem a classificação 8 da variável grau de Escolaridade, ou seja, tem até educação superior incompleta receber até meio salário mínimo é $exp(1,5552) = 4{,}7360$ vezes a chance de uma pessoa analfabeta receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a chance de uma pessoa que tem a classificação 9 da variável grau de Escolaridade, ou seja, tem até educação superior completa receber até meio salário mínimo é $exp(2,9419) = 18{,}9518$ vezes a chance de uma pessoa analfabeta receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a chance de uma pessoa que tem a classificação 10 da variável grau de Escolaridade, ou seja, tem até mestrado completo receber até meio salário mínimo é $exp(4,2263) = 98{,}4634$ vezes a chance de uma pessoa analfabeta receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Quantidade de Horas Trabalhadas e Raça/Cor a chance de uma pessoa que tem a classificação 11 da variável grau de Escolaridade, ou seja, tem até doutorado completo receber até meio salário mínimo é $exp(5,0510) = 156{,}1786$ vezes a chance de uma pessoa analfabeta receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Escolaridade, quantidade de horas a chance de uma pessoa que se define da Raça/Cor 2, ou seja, branca receber até meio salário

mínimo é $exp(0,2395) = 1{,}2706$, 27,06% de chance de uma pessoa que se declara indígena receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Escolaridade, quantidade de horas a chance de uma pessoa que se define da Raça/Cor 4, ou seja, preta/negra receber até meio salário mínimo é $exp(0,1503) = 1{,}1621$, 16,21% de chance de uma pessoa que se declara indígena receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Escolaridade, quantidade de horas a chance de uma pessoa que se define da Raça/Cor 6, ou seja, amarela receber até meio salário mínimo é $exp(0,1971) = 1{,}3103$, 31,03% vezes a chance de uma pessoa que se declara indígena receber até meio salário mínimo;

- Fixadas as variáveis Sexo, Faixa Etária, Escolaridade, quantidade de horas a chance de uma pessoa que se define da Raça/Cor 8, ou seja, parda receber até meio salário mínimo é $exp(0,2703) = 1{,}3272$, 32,72% de chance de uma pessoa que se declara indígena receber até meio salário mínimo;

Devido a suposição de razão de chances proporcionais assumida para o modelo de regressão logística ordinal melhor ajustado aos dados, as mesmas conclusões obtidas quanto a chance do indivíduo que recebe até meio salário mínimo, as mesmas conclusões também para as demais categorias da variável Renda Média Salarial.

Tabela 18: Estimativa da exponencial dos coeficientes do modelo final de regressão ordinal com função de ligação *logit*.

	β	*odds*	p-valor	
Sexo 2	$-0,5189$	1,6801	$2x10^{-16}$	***
Faixa etária 2	$-0,1437$	1,1545	0,0030	**
Faixa etária 3	1,4900	4,4370	$2x10^{-16}$	***
Faixa etária 4	2,0139	7,4925	$2x10^{-16}$	***
Faixa etária 5	2,3847	10,8558	$2x10^{-16}$	***
Faixa etária 6	2,5927	13,3658	$2x10^{-16}$	***
Faixa etária 7	3,0351	20,8030	$2x10^{-16}$	***
Faixa etária 8	3,5758	35,7232	$2x10^{-16}$	***
Faixa etária 99	2,6079	13,5705	$2x10^{-16}$	***
Escolaridade 2	0,3954	1,4850	$2x10^{-16}$	***
Escolaridade 3	0,3412	1,4066	$2x10^{-16}$	***
Escolaridade 4	0,4693	1,5989	$2x10^{-16}$	***
Escolaridade 5	0,7762	2,1700	$2x10^{-16}$	***
Escolaridade 6	0,4345	1,5441	$2x10^{-16}$	***
Escolaridade 7	0,8039	0,4476	$2x10^{-16}$	***
Escolaridade 8	1,5552	4,7360	$2x10^{-16}$	***
Escolaridade 9	2,9419	18,9518	$2x10^{-16}$	***
Escolaridade 10	4,2263	68,4634	$2x10^{-16}$	***
Escolaridade 11	5,0510	156,1786	$2x10^{-16}$	***
Quantidade Hora Trabalhadas	0,0326	0,9679	$2x10^{-16}$	***
Raça/Cor 2	0,2395	1,2706	0,0001	***
Raça/Cor 4	0,1503	1,1622	0,0138	*
Raça/Cor 6	0,1971	1,2178	0,0023	**
Raça/Cor 8	0,2703	1,3103	$2x10^{-16}$	***
Raça/Cor 99	0,2831	1,3272	$2x10^{-16}$	***

Significância: 0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1

4 CONCLUSÕES

Portanto o primeiro modelo proposto o modelo de regressão normal linear não foi adequado aos dados, em seguida propondo os modelos de regressão para dados O modelo de regressão *probit* conseguiu estimar uma porcentagem de acerto de 48,45%, logo errou em 51,55% das estimativas. E quando foi aplicado o modelo de regressão ordinal com função de ligação *loglog* obteve uma porcentagem de acerto de 48,21%, e errou 51,79% das estimativas. Notou-se que o modelo de regressão logística ordinal (*logit*) teve o melhor desempenho, com a maior taxa de acerto, cerca de 49,65% e consequentemente a menor taxa de erro.

Concluímos que existe diferença salarial entre os sexos no estado da Paraíba. A partir do modelo de regressão logístico, que teve o melhor desempenho tivemos a seguinte

interpretação: a pessoa do sexo feminino ter aproximadamente 68,01% de chance de ter Renda Média Salarial tal quando comparado a uma pessoa do sexo masculino. A partir desses resultados ficam outros questionamentos interessantes para trabalhos futuros, como qual motivo do analfabetismo em mulheres ser mais alto que o dos homens e o motivo de não estudarem mais, em uma pesquisa por amostragem em João Pessoa e em outras cidades do estado da Paraíba, será porque começaram uma família muito cedo a necessidade de trabalhar em algo e/ou cuidar do lar, ou por algum tipo de proibição ou impedimento ou até mesmo falta de incentivo e informação. Desta forma pesquisando os principais obstáculos. Também é possível fazer uma comparação da diferença salarial na cidade de João Pessoa com o restante do estado da Paraíba. Verificar como a diferença salarial foi afetada com a pandemia de 2020/2021, observando também o estado da Paraíba com os demais estados da região nordeste, fazendo um levantamento dos principais fatores que podem causar essa diferença.

Referências

Agresti, Alan (2018). *An introduction to categorical data analysis*. John Wiley & Sons.

Ananth, Cande V e David G Kleinbaum (1997). "Regression models for ordinal responses: a review of methods and applications." Em: *International journal of epidemiology* 26.6, pp. 1323–1333.

Bisong, Ekaba (2019). *Building machine learning and deep learning models on Google cloud platform: A comprehensive guide for beginners*. Apress.

Correia de Souza, Tatiene (2003). "Inferência em modelos heteroscedásticos na presença de pontos de alavanca". Diss. de mestr. Universidade Federal de Pernambuco.

Della Lucia, Suzana Maria et al. (2013). "Ordered probit regression analysis of the effect of brand name on beer acceptance by consumers". Em: *Food Science and Technology* 33, pp. 586–591.

Marôco, João (2018). *Análise Estatıstica com o SPSS Statistics.: 7ª edição*. ReportNumber, Lda.

McCullagh, Peter (1980). "Regression models for ordinal data". Em: *Journal of the Royal Statistical Society: Series B (Methodological)* 42.2, pp. 109–127.

McCullagh, Peter e John A Nelder (2019). *Generalized linear models*. Routledge.

Nelder, John Ashworth e Robert WM Wedderburn (1972). "Generalized linear models". Em: *Journal of the Royal Statistical Society: Series A (General)* 135.3, pp. 370–384.

Pessoa, Djalma Galvão Carneiro e Pedro Luis Nascimento Silva (1998). "Análise de dados amostrais complexos". Em: *São Paulo: Associação Brasileira de Estatıstica* 1.

Ramos, Lara Castro Padilha e Ana Virgınia Moreira Gomes (s.d.). "Lei geral de proteção de dados pessoais e seus reflexos nas relações de trabalho". Em: *Scientia Iuris* 23.2 (), p. 127.

Silva, Ana Hermınia Andrade (2017). "Essays on data transformation and regression analysis". Tese de dout.

Silva, Martony Demes da (2020). "Aplicação da Ferramenta Google Colaboratory para o Ensino da Linguagem Python". Em: *Anais da IV Escola Regional de Engenharia de Software*. SBC, pp. 67–76.

Ziegel, Eric R (2003). "Modern applied statistics with S". Em: *Technometrics* 45.1, p. 111.

B – APÊNDICES 1

Script do R pelo Python

```
%load_ext rpy2.ipython

%%R
install.packages("gridExtra")

%%R
library("ggplot2")
library("RColorBrewer")
library(gridExtra)

%%R
load("/content/drive/MyDrive/PB2016.Rda")

PB2016 <- PB2016[which(PB2016$`Faixa Remun M dia (SM)` != 99), ]
PB2016$`Raca Cor`[which(PB2016$`Raca Cor` == 9)] <- 99
# posicao <- which(PB2016$Munic pio == 250750) # 361094

########## AN LISE DESCRITIVA ##########

%%R
# variavel sexo

homens <- length(which(PB2016$`Sexo Trabalhador` == 1))
mulheres <- length(which(PB2016$`Sexo Trabalhador` == 2))

descritivas <- c(homens/(homens+mulheres),mulheres/(homens+mulheres) )
descritivas

%%R
library(xtable)
# variavel sexo
raca <- PB2016$`Ra a Cor`
homens_raca <- PB2016$`Ra a Cor`[which(PB2016$`Sexo Trabalhador` == 1)]
mulheres_raca <- PB2016$`Ra a Cor`[which(PB2016$`Sexo Trabalhador` == 2)]

ttt <- (table(raca)/length(raca))*100
hhh <- (table(homens_raca)/length(homens_raca))*100
mmm <- (table(mulheres_raca)/length(mulheres_raca))*100
xtable(rbind(ttt,hhh,mmm), digits = 3)

######### Tabelas #########

%%R
# TABELA Ph faixa remunerada m dia do sexo masculino
# TABELA Pf faixa remunerada m dia do sexo feminino
renda_media <- PB2016$`Faixa Remun M dia (SM)`

renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100


xtable(rbind(renda_mediat,renda_mediah,renda_mediam), digits = 2)
```

```
%%R
library(xtable)
# levels(as.factor(PB2016$`Faixa Et ria `))
#df_fm <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])
#df_ff <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])

# TABELA feh faixa et ria do sexo masculino
# TABELA fef faixa et ria do sexo feminino
df_ft <- table(as.factor(PB2016$`Faixa Et ria `))/length(PB2016$`Faixa Et ria `)*100
df_fm <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])
        /length(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])*100
df_ff <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])
        /length(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])*100

xtable(rbind(df_ft, df_fm, df_ff), digits = 3)

%%R
# TABELA feh faixa et ria do sexo masculino
# TABELA fef faixa et ria do sexo feminino
df_fe <- data.frame(table(PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 1)]))

# EA2005 Escolaridade ap s 2005 do sexo masculino
# Pf Escolaridade ap s 2005 do sexo feminino
EA20051 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 1)]
EA20052 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 2)]


# TABELA Ph Escolaridade ap s 2005 do sexo masculino
# TABELA Pf Escolaridade ap s 2005 do sexo feminino
# levels(as.factor(PB2016$`faixa et ria `))
table(as.factor(PB2016$`Escolaridade ap s 2005`))

%%R
renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediat

%%R

renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100

df_renda_mediat <- data.frame(renda_mediat, x =seq(0,11))
hist1 <- ggplot(data = df_renda_mediat, aes(x = x, y = renda_mediat)) +
         geom_bar(color = 'black', stat="identity")  +
         theme(axis.text.x = element_text(color="black", size=10))
#          geom_bar(data = data.frame(x = seq(0,11), y = renda_mediah),
            width = 0.4, stat = "identity", fill = "white") +
#          geom_bar(data = data.frame(x = seq(0,11), y = renda_mediam),
            width = 0.2, stat = "identity", fill = "black")



hist1


%%R
```

```
# HISTOGRAMA Ph faixa remunerada m dia do sexo masculino
# HISTOGRAMA Pf faixa remunerada m dia do sexo feminino
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100
#x <- c("At 1/2","1/2 at 1","1 at 1e1/2","1e1/2 at 2","2 at 3","3 at 4","4 at 5",
"5 at 7","7 at 10","10 at 15","10 at 20",">20")
renda_mediah
#df_renda_mediahm <- data.frame(renda_mediah, x = seq(0,11), renda_mediam)
#df_renda_mediahm

%%R
# HISTOGRAMA Ph faixa remunerada m dia do sexo masculino
# HISTOGRAMA Pf faixa remunerada m dia do sexo feminino
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100
#x <- c("At 1/2","1/2 at 1","1 at 1e1/2","1e1/2 at 2","2 at 3","3 at 4","4 at 5",
    "5 at 7","7 at 10","10 at 15","10 at 20",">20")
df_renda_mediahm <- data.frame(renda_mediah, x = seq(0,11), renda_mediam)
# Gr fico de barras simples:
bar1 <- ggplot(data = df_renda_mediahm, aes(x = x, y = renda_mediah)) +
          xlab("Renda m dia salarial em classes") +
          ylab("Propor  o de trabalhadores homens") +
          geom_bar(color = 'black', stat="identity") +
          theme(axis.text.x = element_text(color="black", size=10))+
          scale_x_continuous(breaks = seq(0, 11, 2)) +
          scale_y_continuous(breaks = seq(0, 50, 3))

bar2 <- ggplot(data = df_renda_mediahm, aes(x = x, y = renda_mediam)) +
          xlab("Renda m dia salarial em classes") +
          ylab("Propor  o de trabalhadores mulheres") +
          geom_bar(color = 'black', stat="identity")  +
          theme(axis.text.x = element_text(color="black", size=10))+
          scale_x_continuous(breaks = seq(0, 11, 2)) +
          scale_y_continuous(breaks = seq(0, 50, 3))

grid.arrange(bar1, bar2, ncol = 2, nrow = 1)

%%R
table(QHC2/length(QHC2))*100

Tabelas

%%R
# TABELA Ph faixa remunerada m dia do sexo masculino
# TABELA Pf faixa remunerada m dia do sexo feminino
renda_media <- PB2016$`Faixa Remun M dia (SM)`

renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100


xtable(rbind(renda_mediat,renda_mediah,renda_mediam), digits = 2)



%%R
library(xtable)
# levels(as.factor(PB2016$`Faixa Et ria`))
```

```
#df_fm <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])
#df_ff <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])

# TABELA feh faixa et ria do sexo masculino
# TABELA fef faixa et ria do sexo feminino
df_ft <- table(as.factor(PB2016$`Faixa Et ria `))/length(PB2016$`Faixa Et ria `)*100
df_fm <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])
        /length(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 1)])*100
df_ff <- table(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])
        /length(PB2016$`Faixa Et ria `[which(PB2016$`Sexo Trabalhador ` == 2)])*100

xtable(rbind(df_ft, df_fm, df_ff), digits = 3)

%%R
# TABELA feh faixa et ria do sexo masculino
# TABELA fef faixa et ria do sexo feminino
df_fe <- data.frame(table(PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 1)]))

# EA2005 Escolaridade ap s 2005 do sexo masculino
# Pf Escolaridade ap s 2005 do sexo feminino
EA20051 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 1)]
EA20052 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador ` == 2)]


# TABELA Ph Escolaridade ap s 2005 do sexo masculino
# TABELA Pf Escolaridade ap s 2005 do sexo feminino
# levels(as.factor(PB2016$`faixa et ria `))
table(as.factor(PB2016$`Escolaridade ap s 2005`))

%%R
renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediat

%%R

renda_mediat <- (table(renda_media)/length(renda_media))*100
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100

df_renda_mediat <- data.frame(renda_mediat, x =seq(0,11))
hist1 <- ggplot(data = df_renda_mediat, aes(x = x, y = renda_mediat)) +
         geom_bar(color = 'black', stat="identity") +
         theme(axis.text.x = element_text(color="black", size=10))
#          geom_bar(data = data.frame(x = seq(0,11), y = renda_mediah),
            width = 0.4, stat = "identity", fill = "white") +
#          geom_bar(data = data.frame(x = seq(0,11), y = renda_mediam),
            width = 0.2, stat = "identity", fill = "black")



hist1


%%R
# HISTOGRAMA Ph faixa remunerada m dia do sexo masculino
# HISTOGRAMA Pf faixa remunerada m dia do sexo feminino
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100
```

```
#x <- c("At 1/2","1/2at 1","1at 1e1/2","1e1/2at 2","2at 3","3at 4","4at 5",
"5at 7","7at 10","10at 15","10at 20",">20")
renda_mediah
#df_renda_mediahm <- data.frame(renda_mediah, x = seq(0,11), renda_mediam)
#df_renda_mediahm

%%R
# HISTOGRAMA Ph faixa remunerada m dia do sexo masculino
# HISTOGRAMA Pf faixa remunerada m dia do sexo feminino
renda_mediah <- (table(ph)/length(ph))*100
renda_mediam <- (table(pf)/length(pf))*100
#x <- c("At 1/2","1/2at 1","1at 1e1/2","1e1/2at 2","2at 3","3at 4","4at 5",
"5at 7","7at 10","10at 15","10at 20",">20")
df_renda_mediahm <- data.frame(renda_mediah, x = seq(0,11), renda_mediam)
# Gr fico de barras simples:
bar1 <- ggplot(data = df_renda_mediahm, aes(x = x, y = renda_mediah)) +
          xlab("Renda m dia salarial em classes") +
          ylab("Propor  o de trabalhadores homens") +
          geom_bar(color = 'black', stat="identity") +
          theme(axis.text.x = element_text(color="black", size=10))+
          scale_x_continuous(breaks = seq(0, 11, 2)) +
          scale_y_continuous(breaks = seq(0, 50, 3))

bar2 <- ggplot(data = df_renda_mediahm, aes(x = x, y = renda_mediam)) +
          xlab("Renda m dia salarial em classes") +
          ylab("Propor  o de trabalhadores mulheres") +
          geom_bar(color = 'black', stat="identity")  +
          theme(axis.text.x = element_text(color="black", size=10))+
          scale_x_continuous(breaks = seq(0, 11, 2)) +
          scale_y_continuous(breaks = seq(0, 50, 3))

grid.arrange(bar1, bar2, ncol = 2, nrow = 1)

%%R
table(QHC2/length(QHC2))*100

%%R
faixa_h <- (table(EA20051)/length(EA20051))*100
faixa_m <- (table(EA20052)/length(EA20052))*100

df_EA20051 <- data.frame(faixa_h, faixa_m, x = seq(1,11))
names(df_EA20051)


%%R
library(ggplot2)
# TABELA EA20051 Escolaridade ap s 2005 do sexo masculino
# TABELA EA20052 Escolaridade ap s 2005 do sexo feminino
#df_EA2005 <- data.frame(table(PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador` == 1
#                    table(PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador` == 2)]))
EA20051 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador` == 1)]
EA20052 <- PB2016$`Escolaridade ap s 2005`[which(PB2016$`Sexo Trabalhador` == 2)]

faixa_h <- (table(EA20051)/length(EA20051))*100
faixa_m <- (table(EA20052)/length(EA20052))*100

df_EA20051 <- data.frame(faixa_h, faixa_m, x = seq(1,11))

bar_df_EA20051 <- ggplot(data = df_EA20051, aes(x = x, y = faixa_h)) +
```

```
          ylab("N mero de trabalhadores mulheres") +
          xlab("Escolaridade") +
          geom_bar(color = 'black', stat="identity") +
          theme(axis.text.x = element_text(color="black", size=10))#+
#          scale_x_continuous(breaks = seq(0, 12, 2)) +
#          scale_y_continuous(breaks = seq(0, 50, 3))

bar_df_EA20052 <- ggplot(data = df_EA20051, aes(x = x, y = faixa_m)) +
          ylab("N mero de trabalhadores mulheres") +
          xlab("Escolaridade") +
          geom_bar(color = 'black', stat="identity") +
          theme(axis.text.x = element_text(color="black", size=10))#+
#          scale_x_continuous(breaks = seq(0, 12, 2)) +
#          scale_y_continuous(breaks = seq(0, 50, 3))


grid.arrange(bar_df_EA20051, bar_df_EA20052, ncol = 2, nrow = 1)

########## MODELO NORMAL LINEAR ##########

%load_ext rpy2.ipython

%%R
load("/content/drive/MyDrive/PB2016.Rda")
# posicao <- which(PB2016$Munic pio == 250750) # 361094

# jp_PB2016 <- data.frame(PB2016[posicao, ])

%%R
# Modelo completo
PB2016$`Faixa Et ria`[which(PB2016$`Faixa Et ria` == 99)] <- 1
PB2016$`Faixa Et ria`[which(PB2016$`Faixa Et ria` == 3)] <- 1
PB2016$`Faixa Et ria`[which(PB2016$`Faixa Et ria` == 2)] <- 1
PB2016$`Ra a Cor`[which(PB2016$`Ra a Cor` == 6)] <- 1
PB2016$`Ra a Cor`[which(PB2016$`Ra a Cor` == 4)] <- 1


#
# as.factor(`Faixa Et ria`)99 85.052    784.684   0.363   0.7164
# as.factor(`Faixa Et ria`)3 -32.829     52.176  -0.629   0.5292
# as.factor(`Faixa Et ria`)2 50.7903    52.5653   0.966 0.333928
# as.factor(`Ra a Cor`)6   90.3539    65.5490   1.378 0.168075
# as.factor(`Ra a Cor`)4   14.1736    27.3282   0.519    0.604

modelo_complt <- lm(`Faixa Remun M dia (SM)` ~ as.factor(`Sexo Trabalhador`) +
                      as.factor(`Faixa Et ria`) +  as.factor(`Escolaridade ap s 2005`) +
                      `Qtd Hora Contr` + as.factor(`Ra a Cor`), data = PB2016)
summary(modelo_complt)



########## MODELO PARA DADOS CATEGORICOS ORDINAIS ##########
# **PACKAGE MASS FUNCTION POLR**

%load_ext rpy2.ipython

%%R
install.packages("MASS")
install.packages("sure")
```

```
%%R
load("/content/drive/MyDrive/PB2016.Rda")
PB2016 <- PB2016[(which(PB2016$`Faixa Remun M dia (SM)` != 99)), ] # 11774
PB2016$`Ra a Cor`[which(PB2016$`Ra a Cor` == 9)] <- 99
nrow(PB2016)

# probit

%%R

PB2016$`Faixa Et ria`[which(PB2016$`Faixa Et ria` == 2)] <- 1

#                                          Estimate  Std. Error   t value   Pr(>|t|)
# as.factor(`Faixa Et ria`)2             -0.00367368  0.02665775   -0.1378   0.8903914
modelo_polr <- polr(as.factor(`Faixa Remun M dia (SM)`) ~ as.factor(`Sexo Trabalhador`) +
                      as.factor(`Faixa Et ria`) +  as.factor(`Escolaridade ap s 2005`) +
                      `Qtd Hora Contr` + as.factor(`Ra a Cor`), data = PB2016,
                       method = "probit")

%%R
install.packages("lmtest")

%%R
library(lmtest)
round(coeftest(modelo_polr),4)

%%R
length(as.numeric(PB2016$`Faixa Remun M dia (SM)`)) # 819849
length(modelo_polr$fitted.values) #9838188
# para as 12 vari veis 9838188/819849 = 12

fit <- fitted(modelo_polr)


fit_df <- as.data.frame(fit)

%%R
# O que antes era probabilidade agora    classificacao.
y_hat <- NULL
for(i in 1:nrow(fit)){
    coluna <- NA
    for(j in 1:ncol(fit)){
      if(max(fit[i, ]) == fit[i,j]) coluna <- j
    }
  y_hat[i] <- coluna
}
levels(as.factor(y_hat))

%%R
install.packages("xtable")

%%R
library(xtable)
table(PB2016$`Faixa Remun M dia (SM)`, y_hat)
xtable(table(PB2016$`Faixa Remun M dia (SM)`, y_hat))
```

```
%%R
mean(nova == y_hat)

%%R
BIC(modelo_polr)
# 2496902

%%R
install.packages("PResiduals")

%%R
install.packages("ggplot2")

%%R
library(PResiduals)
pres <- presid(modelo_polr)
# residuos vs covariancia
# plot e Q-Q Plot
library(ggplot2)


%%R
p1 <- ggplot(data.frame(x = PB2016$`Faixa Remun M dia (SM)`, y = pres), aes(x, y)) +
geom_point(color = "#444444", shape = 19, size = 2, alpha = 0.5) +
geom_smooth(color = "red", se = FALSE) +
ylab("Probability-scale residual")

p2 <- ggplot(data.frame(y = pres), aes(sample = y)) +
stat_qq(distribution = qunif, dparams = list(min = -1, max = 1), alpha = 0.5) +
stat_qq_line(colour = "red", dparams = list(min = -1, max = 1), alpha = 0.5) +
xlab("Sample quantile") +
ylab("Theoretical quantile")

grid.arrange(p1, p2, ncol = 2) # Figure 1


# probit

%%R

PB2016$`Faixa Et ria`[which(PB2016$`Faixa Et ria` == 2)] <- 1

#                                          Estimate  Std. Error   t value   Pr(>|t|)
# as.factor(`Faixa Et ria`)2             -0.00367368  0.02665775   -0.1378   0.8903914
modelo_polr <- polr(as.factor(`Faixa Remun M dia (SM)`) ~ as.factor(`Sexo Trabalhador`) +
                      as.factor(`Faixa Et ria`) +  as.factor(`Escolaridade ap s 2005`) +
                      `Qtd Hora Contr` + as.factor(`Ra a Cor`), data = PB2016,
                       method = "probit")

%%R
install.packages("lmtest")

%%R
library(lmtest)
round(coeftest(modelo_polr),4)

%%R
length(as.numeric(PB2016$`Faixa Remun M dia (SM)`)) # 819849
length(modelo_polr$fitted.values) #9838188
```

```
# para as 12 varieis 9838188/819849 = 12

fit <- fitted(modelo_polr)


fit_df <- as.data.frame(fit)

%%R
# O que antes era probabilidade agora    classificacao.
y_hat <- NULL
for(i in 1:nrow(fit)){
    coluna <- NA
    for(j in 1:ncol(fit)){
      if(max(fit[i, ]) == fit[i,j]) coluna <- j
    }
  y_hat[i] <- coluna
}
levels(as.factor(y_hat))

%%R
install.packages("xtable")

%%R
library(xtable)
table(PB2016$`Faixa Remun M dia (SM)`, y_hat)
xtable(table(PB2016$`Faixa Remun M dia (SM)`, y_hat))


%%R
mean(nova == y_hat)

%%R
BIC(modelo_polr)
# 2496902

%%R
install.packages("PResiduals")

%%R
install.packages("ggplot2")

%%R
library(PResiduals)
pres <- presid(modelo_polr)
# residuos vs covariancia
# plot e Q-Q Plot
library(ggplot2)


%%R
p1 <- ggplot(data.frame(x = PB2016$`Faixa Remun M dia (SM)`, y = pres), aes(x, y)) +
geom_point(color = "#444444", shape = 19, size = 2, alpha = 0.5) +
geom_smooth(color = "red", se = FALSE) +
ylab("Probability-scale residual")

p2 <- ggplot(data.frame(y = pres), aes(sample = y)) +
stat_qq(distribution = qunif, dparams = list(min = -1, max = 1), alpha = 0.5) +
stat_qq_line(colour = "red", dparams = list(min = -1, max = 1), alpha = 0.5) +
xlab("Sample quantile") +
```

```
ylab("Theoretical quantile")

grid.arrange(p1, p2, ncol = 2) # Figure 1

%%R
library(sure)
set.seed(2021)
sres <- resids(modelo_polr)


%%R
set.seed(101)
sres <- resids(modelo_polr)
# Residual-vs-covariate plot and Q-Q plot
library(ggplot2) # needed for autoplot function
p3 <- sure::autoplot.resid(resids(modelo_polr), what = "covariate",
              x = PB2016$`Faixa Remun M dia (SM)`)
p4 <- ggplot(data.frame(sres), aes(sample = sres)) +
  stat_qq() + stat_qq_line(colour = "red")

grid.arrange(p3, p4, ncol = 2) # Figure



%%R
modelo_polr_logistic <- polr(as.factor(`Faixa Remun M dia (SM)`) ~
                        as.factor(`Sexo Trabalhador`) +
                       as.factor(`Faixa Et ria`) + as.factor(`Escolaridade ap s 2005`) +
                       `Qtd Hora Contr` + as.factor(`Ra a Cor`), data = PB2016,
                      method = "logistic")

%%R
length(as.numeric(PB2016$`Faixa Remun M dia (SM)`)) # 819849
length(modelo_polr_logistic$fitted.values) #9838188
# para as 12 vari veis 9838188/819849 = 12

fit1 <- fitted(modelo_polr_logistic)


fit1_df <- as.data.frame(fit1)

%%R
# O que antes era probabilidade agora    classificacao.
y_hat1 <- NULL
for(i in 1:nrow(fit1)){
    coluna <- NA
    for(j in 1:ncol(fit1)){
      if(max(fit1[i, ]) == fit1[i,j]) coluna <- j
    }
  y_hat1[i] <- coluna
}
levels(as.factor(y_hat1))

%%R
install.packages("xtable")

%%R
library(xtable)
table(PB2016$`Faixa Remun M dia (SM)`, y_hat)
```

```
xtable(table(PB2016$`Faixa Remun M dia (SM)`, y_hat))


%%R
mean(nova != y_hat1)

%%R
1-0.484466

%%R
library(lmtest)
round(coeftest(modelo_polr_logistic),4)

%%R
BIC(modelo_polr_logistic)
#2485574

%%R
install.packages("PResiduals")

%%R
library(PResiduals)
pres_logistic <- presid(modelo_polr_logistic)
# residuos vs covariancia
# plot e Q-Q Plot
library(ggplot2)

%%R
p1_logistic <- ggplot(data.frame(x = PB2016$`Faixa Remun M dia (SM)`,
        y = pres_logistic), aes(x, y)) +
    geom_point(color = "#444444", shape = 19, size = 2, alpha = 0.5) +
    geom_smooth(color = "red", se = FALSE) +
    ylab("Escala de probabilidade residual")

p2_logistic <- ggplot(data.frame(y = pres_logistic), aes(sample = y)) +
stat_qq(distribution = qunif, dparams = list(min = -1, max = 1), alpha = 0.5) +
stat_qq_line(colour = "red", dparams = list(min = -1, max = 1), alpha = 0.5) +
xlab("Amostra do quantil") +
ylab("Quantil te rico")

grid.arrange(p1_logistic, p2_logistic, ncol = 2) # Figure 1

loglog

%%R
modelo_polr_loglog <- polr(as.factor(`Faixa Remun M dia (SM)`) ~
                        as.factor(`Sexo Trabalhador`) +
                      as.factor(`Faixa Et ria`) +  as.factor(`Escolaridade ap s 2005`) +
                      `Qtd Hora Contr` + as.factor(`Ra a Cor`), data = PB2016,
                    method = "loglog")
summary(modelo_polr_loglog)

# Intercepts:
#                 Value     Std. Error t value
#440|664.4         0.7457    0.0332     22.4345
#664.4|1104.4      1.7831    0.0333     53.4843
#1104.4|1544.4     3.2955    0.0334     98.6499
#1544.4|2204.4     3.8787    0.0334    115.9940
#2204.4|3084.4     4.6768    0.0335    139.5266
```

```
#3084.4|3964.4      5.1443    0.0336    153.0999
#3964.4|5284.4      5.4406    0.0337    161.5624
#5284.4|7484.4      5.9324    0.0339    175.2374
#7484.4|11004.4     6.4498    0.0342    188.8099
#11004.4|15404.4    7.3137    0.0352    207.9393
#15404.4|17608.8    8.1204    0.0372    218.0293

#Residual Deviance: 2573958.24
#AIC: 2574032.24
```

```
%%R
length(as.numeric(PB2016$`Faixa Remun M dia (SM)`)) # 819849
length(modelo_polr_loglog$fitted.values) #9838188
# para as 12 vari veis 9838188/819849 = 12

fit2 <- fitted(modelo_polr_loglog)


fit2_df <- as.data.frame(fit2)
```

```
%%R
# O que antes era probabilidade agora    classificacao.
y_hat2 <- NULL
for(i in 1:nrow(fit2)){
    coluna <- NA
    for(j in 1:ncol(fit2)){
      if(max(fit2[i, ]) == fit2[i,j]) coluna <- j
    }
  y_hat2[i] <- coluna
}
levels(as.factor(y_hat2))
```

```
%%R
library(xtable)
table(PB2016$`Faixa Remun M dia (SM)`, y_hat2)
xtable(table(PB2016$`Faixa Remun M dia (SM)`, y_hat2))


%%R
mean(nova == y_hat2)
```

```
%%R
BIC(modelo_polr_loglog)
# 2574462
```

```
%%R
library(PResiduals)
pres_loglog <- presid(modelo_polr_loglog)
# residuos vs covariancia
# plot e Q-Q Plot
library(ggplot2)
```

```
%%R
p1_loglog <- ggplot(data.frame(x = PB2016$`Faixa Remun M dia (SM)`,
        y = pres_loglog), aes(x, y)) +
        geom_point(color = "#444444", shape = 19, size = 2, alpha = 0.5) +
        geom_smooth(color = "red", se = FALSE) +
        ylab("Probability-scale residual")
```

```
p2_loglog <- ggplot(data.frame(y = pres_loglog), aes(sample = y)) +
        stat_qq(distribution = qunif, dparams = list(min = -1, max = 1),
        alpha = 0.5) +
        stat_qq_line(colour = "red", dparams = list(min = -1, max = 1),
        alpha = 0.5) +
        xlab("Sample quantile") +
        ylab("Theoretical quantile")

grid.arrange(p1_loglog, p2_loglog, ncol = 2) # Figure 1

%%R
# Fit models with various link functions to the simulated data
# Construct Q-Q plots of the surrogate residuals for each model
set.seed(1056)  # for reproducibility
p1_probit <- autoplot.polr(modelo_polr, nsim = 100, what = "qq")
p2_logistic <- autoplot.polr(modelo_polr_logisitc, nsim = 100, what = "qq")
p3_loglog <- autoplot.polr(modelo_polr_loglog, nsim = 100, what = "qq")
#p4 <- autoplot(fit.cloglog, nsim = 100, what = "qq")

# bottom left plot is correct model
gridExtra::grid.arrange(p1_probit, p2_logistic, p3_loglog, ncol = 2)
```