UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE INFORMÁTICA E ESTATÍSTICA

Disciplina: INE 5649-03238 (20212) - Técnicas Estatísticas de Predição

Professor: Luiz Ricardo Nakamura

Alunos: Mariany Ferreira da Silva, 19200646 e Pedro Henrique Dias Nobrega, 19100876

ANÁLISE DE DADOS DOS CRIMES OCORRIDOS NOS ESTADOS UNIDOS COM BASE NA POPULAÇÃO

Introdução, justificativa e objetivos

Nos últimos anos, surgiram diversas iniciativas nas áreas de segurança pública e justiça criminal para tratar de questões relacionadas à eficiência das operações policiais, controle do crime organizado, controle de operações policiais, policiamento comunitário e outras iniciativas dignas de censo e melhor avaliação**.** Universidades e centros de pesquisa tentam ajudar no diagnóstico e controle da violência urbana. Tendo isso em vista esse trabalho objetiva realizar uma análise explorando dados disponibilizados pela Universidade da Califórnia, Irvine levantando o seguinte questionamento: O Número populacional, índice de renda per capita, escolaridade e percentual de desemprego são relevantes quando olhamos para o número total de crimes violentos e não violentos de uma comunidade? Para responder essa pergunta elaboramos 6 hipóteses:

1. Quanto maior a população, maior o número de crimes
2. Quanto maior o índice de renda per capita, menor o número de crimes
3. Quanto maior a % de pessoas sem ensino básico, maior o número de crimes
4. Quanto maior a % de pessoas sem ensino médio, maior o número de crimes
5. Quanto maior a % de pessoas com ensino superior, menor o número de crimes
6. Quanto maior a % de pessoas desempregadas, maior o número de crimes

Materiais e métodos

O conjunto de dados foi adquirido a partir do repositório de aprendizado de máquina da Universidade da Califórnia, Irvine (UCI) e centraliza dados relacionados a 'Crime e Comunidades'. Ele foi preparado usando dados reais de dados socioeconômicos do Censo dos Estados Unidos da América (EUA) de 1990, dados de aplicação da lei da pesquisa LEMAS dos EUA de 1990 e dados criminais do Uniform Crime Reporting (UCR) do Federal Bureau of Investigation (FBI) de 1995. Este conjunto de dados contém um número total de 147 atributos e 2.216 instâncias.

Nossa análise se concentrou em 7 variáveis explicativas que incluem informações sobre população, índice de renda per capita, percentual de desemprego e escolaridade com o objetivo de avaliar se estão relacionadas ao número total de crimes violentos e não violentos em uma comunidade.

Variável resposta

- **CRIMES:** Total de crimes violentos e não violentos na comunidade: é a soma dos valores das informações de número de crimes violentos **(ViolentCrimesPerPop - VIOLENTOS)** e crimes não violentos **(nonViolPerPop - N_VIOLENTOS)**. Essa variável é um valor decimal, e tem como unidade um crime.

Variáveis explicativas

- **POPULACAO / population:** População de uma comunidade: é o número total de habitantes em uma comunidade. Essa variável é um valor inteiro, e tem como unidade uma pessoa.
- **RENDA / perCapInc:** Índice de renda per capita na comunidade
- **SEM_BASICO / PctLess9thGrade:** Percentual de pessoas na comunidade com mais de 24 anos que não completaram o nível básico de educação
- **SEM_MEDIO / PctNotHSGrad** Percentual de pessoas na comunidade com mais de 24 anos que não completaram o nível médio de educação
- **COM_SUPERIOR / PctBSorMore:** Percentual de pessoas na comunidade com mais de 24 anos que completaram o nível superior de educação, incluindo faculdade, mestrado, doutorado e outros
- **DESEMPREGADOS / PctUnemployed:** Percentual de pessoas na comunidade com mais de 15 anos que estão desempregadas

Para analisar esses dados a função Logarítmica foi aplicada nas variáveis **CRIMES, POPULACAO e RENDA** de maneira que os dados se distribuíssem melhor nos gráficos de dispersão e melhorando os índices de correlação.

O FBI observa que o uso desses dados para avaliar as comunidades é muito simplista, pois muitos fatores relevantes não estão incluídos. Por exemplo, comunidades com grande número de visitantes terão maior criminalidade per capita (medida pelos moradores) do que comunidades com menos visitantes, mantendo-se os demais fatores.

Para realizar a análise dos dados neste trabalho será utilizado o Jupyter Notebook que é uma ferramenta de trabalho para equipes de ciência de dados colaborarem e R, uma linguagem de programação multi-paradigma orientada a objetos, programação funcional, dinâmica, fracamente tipada, voltada à manipulação, análise e visualização de dados. Como repositório do projeto escolhi o GitHub, uma plataforma de hospedagem de código-fonte e

arquivos com controle de versão usando o Git. Ele permite que qualquer usuário cadastrado na plataforma contribua em projetos privados e/ou públicos de qualquer lugar do mundo.

Resultados e discussões

A primeira etapa da nossa análise visa extrair algumas informações que podem nos ajudar a entender o comportamento da nossa variável resposta.

Média: Em média acontecem 5370,9104 crimes por comunidade nos Estados Unidos da América por ano nessa série de dados.

Mediana: A mediana de crimes em uma comunidade nos EUA por ano desse conjunto de dados é de 4792,45.

Desvio Padrão: Com o desvio padrão de 3108,2336, essa série de dados representa alta dissonância nos dados, apresentando uma diferença palpável, o que indica uma taxa de crimes bem destoante entre as comunidades americanas analisadas.

Coeficiente de Assimetria: Apresentando uma assimetria de 1,7371, o conjunto de dados possui um viés à direita, ou seja, seus dados se distribuem à direita em um gráfico. Isso representa uma maior frequência de comunidades que sofrem com menor taxa de criminalidade.

Coeficiente de Kurtosis: Com o coeficiente de Kurtosis de 9.1857, esses dados apresentam uma distribuição platicúrtica, ou seja, sua distribuição é mais achatada, o que representa uma semelhança entre as comunidades analisadas.

Análise Multivariada

A segunda etapa da nossa análise visa extrair algumas informações que podem nos ajudar a entender o impacto de cada variável explicativa em nossa variável resposta.

Gráficos de Dispersão

Ao plotar os diagramas de dispersão com as variáveis percebemos que as variáveis CRIMES, POPULACAO e RENDA precisavam ser transformadas por meio da função Logarítmica para melhorar a linearidade dos dados. Depois disso construímos novos gráficos de dispersão. Há uma distribuição bem espaçada ao longo da reta de mínimos quadrados, tanto abaixo, quanto acima da reta. Isso nos mostra que a variabilidade de total de crimes violentos e não violentos em uma comunidade em relação às variáveis explicativas analisadas é maior do que esperávamos. É possível observar também que comportamento nos dados é um pouco diferente do que foi apresentado nas aulas, no diagrama de POPULACAO e SEM_MEDIO não conseguimos enxergar uma tendência linear. Podemos afirmar que existe em alguns casos uma tendência positiva (POPULACAO,

SEM_BASICO, SEM_MEDIO e DESEMPREGADOS) e outros negativa (RENDA e COM_SUPERIOR), esse fato condiz um pouco melhor com as hipóteses elaboradas.

O diagrama de dispersão de POPULACAO X CRIMES mostra um relacionamento positivo indicando que, à medida que a população de uma comunidade cresce, o número de crimes cresce. Quando os valores se aproximam de zero no eixo x, eles ficam mais espaçados, isso nos mostra que existem comunidades com pouca população e um número baixo de crimes e outras com pouca população, mas um grande número de crimes.

O diagrama de dispersão de RENDA X CRIMES mostra um relacionamento negativo indicando que, à medida que o índice de renda per capita da população cresce, o número de crimes diminui.

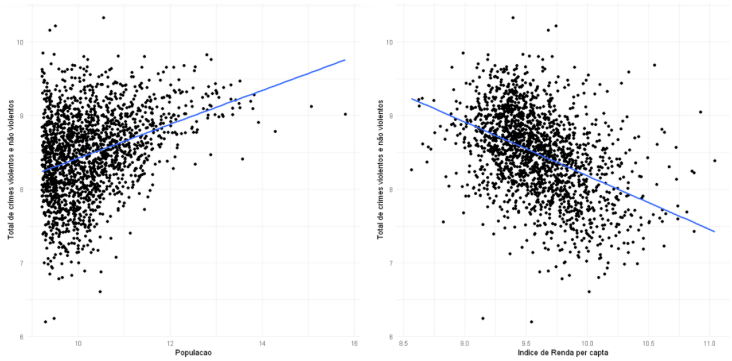

FIGURA n - Diagrama de dispersão de POPULACAO X CRIMES, Diagrama de dispersão de RENDA X CRIMES

O diagrama de dispersão de SEM_BASICO X CRIMES mostra um relacionamento positivo indicando que, à medida que a porcentagem de pessoas sem ensino básico cresce, o número de crimes cresce.

O diagrama de dispersão de SEM_MEDIO X CRIMES mostra um relacionamento positivo indicando que, à medida que a porcentagem de pessoas sem ensino médio cresce, o número de crimes cresce.

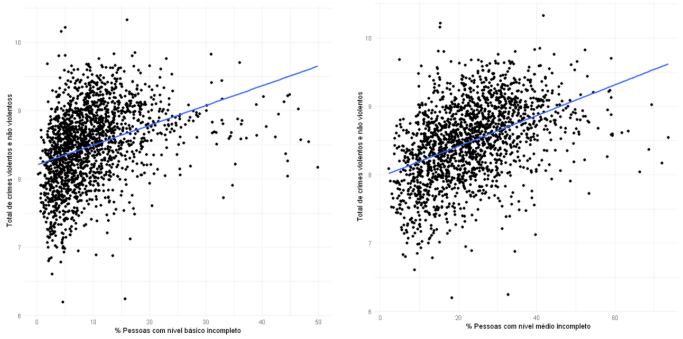

FIGURA n - Diagrama de dispersão de SEM_BASICO X CRIMES, Diagrama de dispersão de SEM_MEDIO X CRIMES

O diagrama de dispersão de COM_SUPERIOR X CRIMES mostra um relacionamento negativo indicando que, à medida que a porcentagem de pessoas com ensino superior cresce, o número de crimes diminui.

O diagrama de dispersão de DESEMPREGADOS X CRIMES mostra um relacionamento positivo indicando que, à medida que a porcentagem de pessoas desempregadas cresce, o número de crimes cresce.

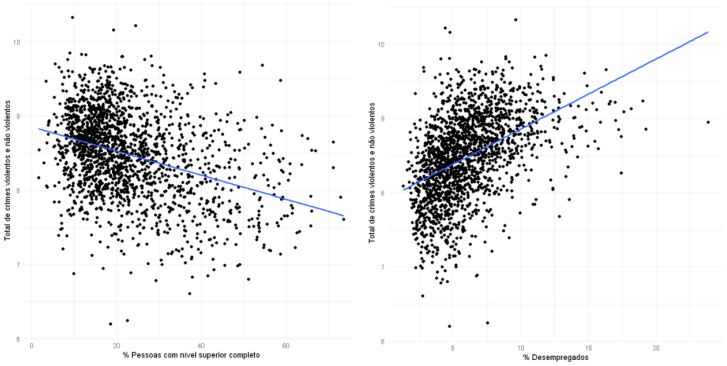

FIGURA n - Diagrama de dispersão de COM_SUPERIOR X CRIMES, Diagrama de dispersão de DESEMPREGADOS X CRIMES

Esperávamos exatamente esse comportamento conforme as 6 hipóteses elaboradas.

Coeficiente de correlação de Pearson

Ao utilizar a função *cor(x, y)* conseguimos calcular os coeficientes de correlação para cada variável explicativa em relação à variável resposta. Para analisá-los consideramos a seguinte escala:

Intervalos [-1, -0.7] e [1,-0.7] correlação linear forte; [-0.7, -0.3] e [0.7, 0.3] moderada; [-0.3, 0.3] fraca ou inexistente

As variáveis **RENDA (-0.4320) e COM_SUPERIOR (-0.3523)** apresentaram uma **correlação moderada negativa**. Já as variáveis **POPULACAO (0.3461), SEM_BASICO (0.3490), SEM_MEDIO (0.4338) e DESEMPREGADOS (0.4496)** apresentaram **correlação moderada positiva**. Esperávamos exatamente esse comportamento conforme as 6 hipóteses elaboradas.

Modelo

Ao utilizar a função *summary(modelo)* conseguimos calcular vários coeficientes importantes para a nossa análise.

```
Call:
lm(formula = CRIMES ~ POPULACAO + RENDA + SEM_BASICO + SEM_MEDIO +
    COM_SUPERIOR + DESEMPREGADOS, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.40609 -0.27563  0.00555  0.28023  2.08930

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.245600   0.579374  15.958  < 2e-16 ***
POPULACAO      0.207095   0.012726  16.274  < 2e-16 ***
RENDA         -0.385738   0.059172  -6.519 9.14e-11 ***
SEM_BASICO    -0.038676   0.005022  -7.702 2.19e-14 ***
SEM_MEDIO      0.037078   0.003959   9.365  < 2e-16 ***
COM_SUPERIOR   0.007274   0.001733   4.197 2.83e-05 ***
DESEMPREGADOS  0.027163   0.006571   4.134 3.73e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4607 on 1823 degrees of freedom
Multiple R-squared:  0.3576,    Adjusted R-squared:  0.3554
F-statistic: 169.1 on 6 and 1823 DF,  p-value: < 2.2e-16
```

FIGURA n - Resumo do modelo

Reta de mínimos quadrados

Y^ = 9.2456 + (0.2070*POPULACAO) + (-0.3857*RENDA) + (-0.0386*SEM_BASICO) + (0.03707*SEM_MEDIO) + (0.007274*COM_SUPERIOR) + (0.02716*DESEMPREGADOS)

POPULACAO: A cada 1 unidade de ln(POPULACAO) a mais eu espero que a variável resposta ln(CRIMES) seja acrescida 0.2070 unidades, considerando as outras variáveis fixas. Olhando o p-valor vemos que 2e-16 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise.

RENDA: A cada 1 unidade de ln(RENDA) a mais eu espero que a variável resposta CRIMES seja acrescida -0.3857 unidades, considerando as outras variáveis fixas. Olhando o p-valor vemos que 9.14e-11 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise.

SEM_BASICO: A cada 1 unidade de SEM_BASICO a mais eu espero que a variável resposta ln(CRIMES) seja acrescida -0.0386 unidades, considerando as outras variáveis fixas. **Esse valor não condiz com o esperado.** Conforme hipótese 3 imaginamos que quanto maior a porcentagem de pessoas sem nível básico, maior seria o número de crimes. Olhando o p-valor vemos que 2.19e-14 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise, o que nos deixa com suspeita de que precisamos descartar essa variável explicativa.

SEM_MEDIO: A cada 1 unidade de SEM_MEDIO a mais eu espero que a variável resposta **ln(CRIMES) seja acrescida 0.03707 unidades**, considerando as outras variáveis fixas. Olhando o p-valor vemos que 2e-16 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise.

COM_SUPERIOR: A cada 1 unidade de COM_SUPERIOR a mais eu espero que a variável resposta **ln(CRIMES) seja acrescida 0.007274 unidades**, considerando as outras variáveis fixas. **Esse valor não condiz com o esperado.** Conforme hipótese 5 imaginamos que quanto maior a porcentagem de pessoas com nível superior, menor seria o número de crimes. Olhando o p-valor vemos que 2.83e-05 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise o que nos deixa com suspeita de que precisamos descartar essa variável explicativa.

DESEMPREGADOS: A cada 1 unidade de DESEMPREGADOS a mais eu espero que a variável resposta **ln(CRIMES) seja acrescida 0.02716 unidades**, considerando as outras variáveis fixas. Olhando o p-valor vemos que 3.73e-05 é menor do que o nível de significância de 0.05 %, podemos rejeitar a hipótese nula, afirmando que esse valores são significativos para a análise.

Tendo o teste de utilidade do modelo com ***p-value: < 2.2e-16*** rejeitamos a hipótese nula, pois esse valor é menor do que 0.05 (nível de significância), então construímos um **modelo útil**.

Análise de Resíduos

Os resíduos indicam a variação natural dos dados, um fator aleatório (ou não) que o modelo não capturou. Se as pressuposições do modelo são violadas, a análise será levada a resultados duvidosos e não confiáveis para inferência. Para concluir com maior confiança vamos realizar as análises de resíduos para cada variável explicativa em relação à variável resposta.

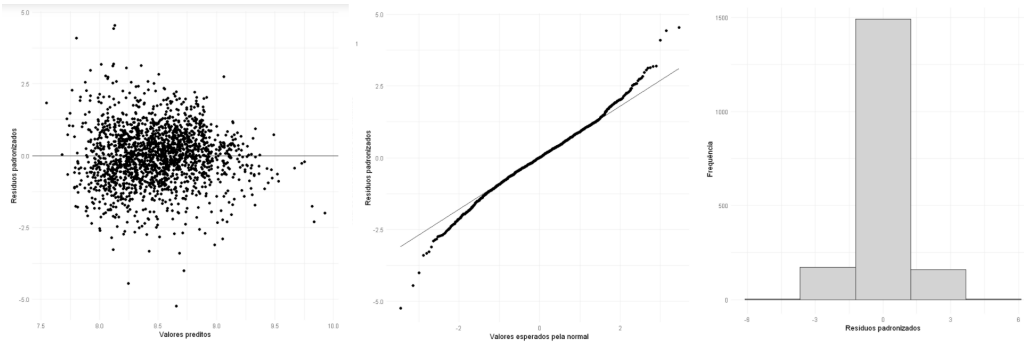

FIGURA N - ANÁLISE DE RESÍDUOS DO MODELO

Ao analisar o diagrama de **Resíduos Padrões X Valores preditos** do nosso modelo, percebemos que eles não apresentavam nenhum padrão específico. A quantidade era aproximadamente a mesma, mas a partir de 8.5 observamos mais resíduos negativos do que positivos. A magnitude dos resíduos positivos era aproximadamente a mesma dos resíduos negativos no diagrama de dispersão, apesar disso os valores **não estavam no**

intervalo [-3;3] e se concentram próximos à zero no eixo y (valores preditos)**.** Além disso, o gráfico de **Resíduos Padrões X Valores esperados pela normal não segue a reta de mínimos quadrados em seus extremos.** Já o histograma de **Frequência X Resíduos Padrões é simétrico**, mas não apresenta um desenho parecido com os desenhos vistos em aula pois em 0 apresenta um pico muito grande e isso explica a concentração de pontos próximos à reta no diagrama de dispersão dos resíduos.

O teste de Shapiro-Wilk apresentou **p-value = 8.082e-11 esse valor é menor do que 0.05 (nível de significância),** ou seja, precisamos **rejeitar a hipótese nula e podemos afirmar** que **esses resíduos não tem uma distribuição normal.**

Ao analisar os **Resíduos Padronizados em relação a cada variável explicativa** encontramos problemas parecidos aos encontrados no diagrama de **Resíduos Padrões X Valores preditos.** Em particular, **POPULACAO, SEM_BASICO E DESEMPREGADOS** temos uma concentração maior de pontos nos intervalos [0, 10] que se dilui ao longo do eixo y (variável explicativa).

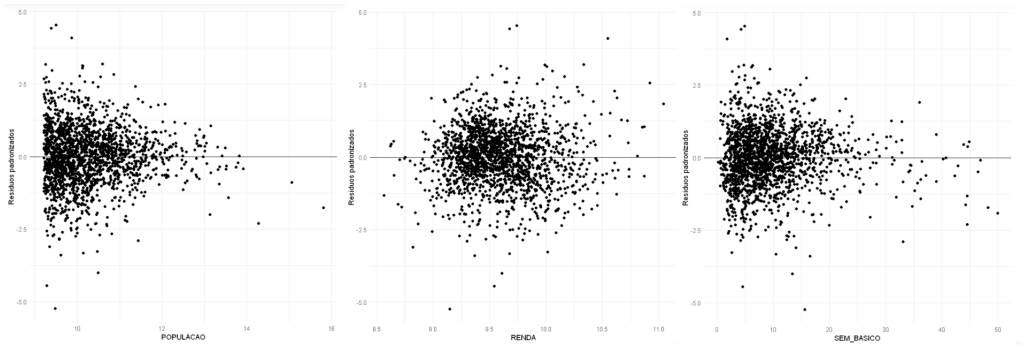

FIGURA N - Resíduos padronizados X POPULACAO, Resíduos padronizados X RENDA, Resíduos padronizados X SEM_BASICO

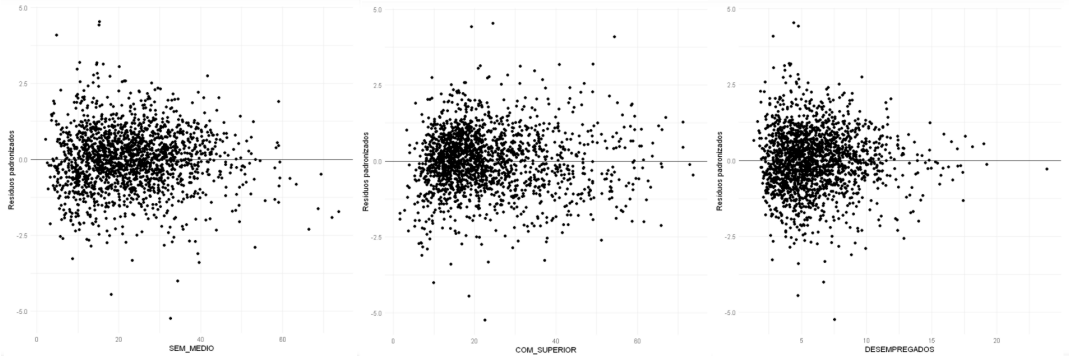

FIGURA N - Resíduos padronizados X SEM_MEDIO, Resíduos padronizados X COM_SUPERIOR, Resíduos padronizados X DESEMPREGADOS

Modelo Ajustado

Após uma análise do modelo e seus resultados, um ajuste foi feito, retirando as variáveis com valores não esperados. Esse modelo ajustado contempla as variáveis **RENDA**, **SEM_MEDIO** e **DESEMPREGADOS** apresentando os seguintes coeficientes:

```
Call:
lm(formula = CRIMES ~ RENDA + SEM_MEDIO + DESEMPREGADOS, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-2.49796 -0.31182  0.00974  0.33593  1.91288

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  10.611825   0.569915  18.620  < 2e-16 ***
RENDA        -0.270159   0.055809  -4.841 1.40e-06 ***
SEM_MEDIO     0.007621   0.001768   4.311 1.71e-05 ***
DESEMPREGADOS 0.047388   0.006890   6.878 8.29e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5025 on 1826 degrees of freedom
Multiple R-squared:  0.2346,    Adjusted R-squared:  0.2334
F-statistic: 186.6 on 3 and 1826 DF,  p-value: < 2.2e-16
```

Figura N - Resumo do modelo ajustado

No geral, o impacto das variáveis foi similar, apresentando somente um leve ajuste no impacto de cada uma:

RENDA: Apresentou uma diminuição no impacto, saindo de -0,3857 para -0,2701, ou seja, a cada unidade de RENDA, é esperado que a variável resposta ln(CRIMES) seja acrescida em -0,2701 pontos, considerando as outras variáveis fixas. Seu **p-value** também foi modificado de 9,14e-11 para 1,4e-06, porém ainda se manteve dentro dos 5% de confiança.

SEM_MEDIO: Essa variável apresentou uma diminuição de impacto considerável, saindo de 0.03707 para 0,0076, apresentando também um aumento em seu **p-value** de 2e-16 para 1,71e-5, mas se manteve dentro do intervalo de confiança.

DESEMPREGADOS: Essa variável demonstrou um aumento no impacto, saindo de 0,0271 para 0,0474, um aumento expressivo. Apresentou uma diminuição em seu **p-value**, saindo de 3,73e-05 para 8,29e-12, mantendo-se no intervalo de 5% de confiança.

Mesmo com essa diferença dos valores, o modelo ainda apresentou um **p-value** de 2e-16, fazendo com que ele esteja no intervalo de confiança, ou seja, a hipótese nula continua sendo rejeitada. Essa rejeição torna o modelo útil.

Análise de Resíduos

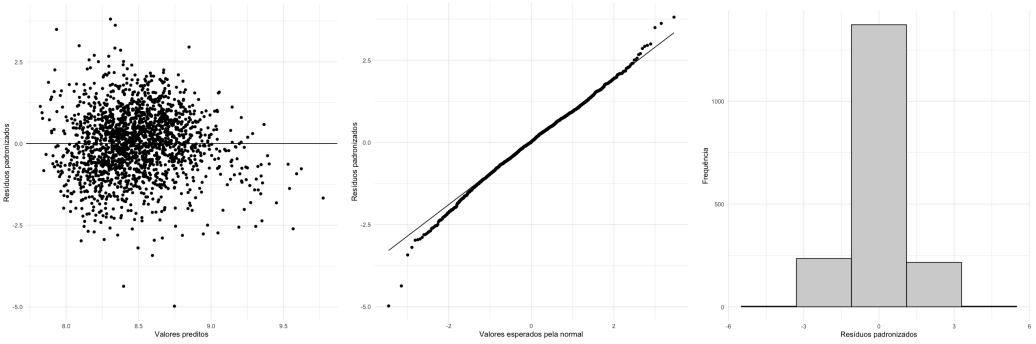

FIGURA N - ANÁLISE DE RESÍDUOS DO MODELO AJUSTADO

Comparando com o resultado posterior, não houve uma mudança significativa nos gráficos, a dispersão se manteve, porém dessa vez fica notável uma quantidade maior de dados negativos do que positivos para todos os segmentos do gráfico. O **p-value** do teste

de Shapiro-Wilk foi de 1,063e-05, um valor menor que o de 5% para que seja aceito a hipótese nula de que os dados estão em uma distribuição normal.

Ao analisar a distribuição de resíduos das variáveis explicativas, podemos verificar uma concentração entre os valores **[9, 10]** para a variável **RENDA**, entre **[0, 40]** para a **SEM_MEDIO** e entre **[2,5, 12,5]** para **DESEMPREGADOS**.

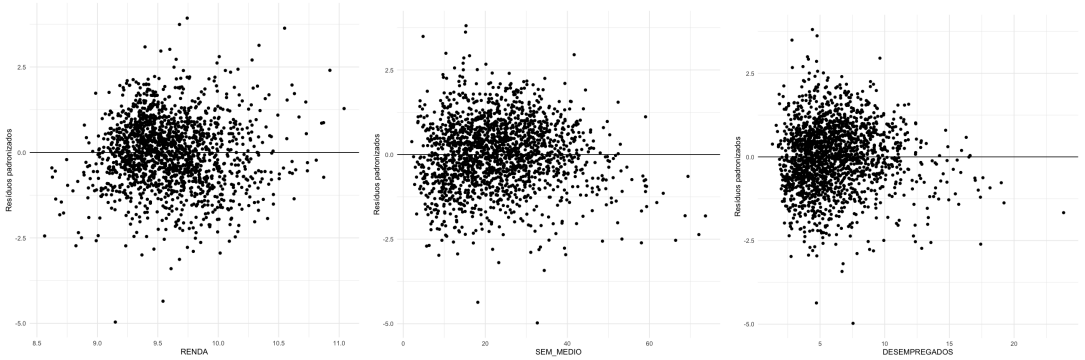

FIGURA N - Resíduos padronizados X RENDA, Resíduos padronizados X SEM_MEDIO, Resíduos padronizados X DESEMPREGADOS

Considerações finais

Em suma, nem todas as variáveis explicativas foram significativas para a análise, conforme seu **p-value** e a análise do modelo demonstrou. A presunção inicial de que o valor **SEM_BASICO** iria influenciar positivamente o número de crimes foi mostrada como falsa, já que esse demonstrou um impacto negativo no modelo, assim como a variável **COM_SUPERIOR**, que teve seu impacto suposto como negativo, ou seja, quanto maior presença de pessoas com ensino superior, menor seria a taxa de crimes, algo que foi demonstrado como falso pelo modelo. Outro comportamento estranho identificado no modelo foi a distribuição dos resíduos da **POPULACAO**, que apresentou um padrão de funil, o que pode ser interpretado como a presença de alguma tendência na amostra.

Após o ajuste do modelo, foi possível notar uma melhora na predição, porém a análise de resíduos apresentando um intervalo maior que o estipulado para se ter uma normal, faz com que a confiança nesse modelo seja abalada.

Referências bibliográficas

NAKAMURA, Luiz. Estudos de correlação. 2022. Disponível em: https://moodle.ufsc.br/pluginfile.php/4760566/mod_resource/content/3/01_Correlacao.pdf Acesso em: 06/03/2022.

NAKAMURA, Luiz. Análise de regressão linear. 2022. Disponível em: https://moodle.ufsc.br/pluginfile.php/4760577/mod_resource/content/4/02_IntroReg.pdf Acesso em: 06/03/2022.

NAKAMURA, Luiz. Inferência sobre o modelo de regressão linear simples. 2022. Disponível em: https://moodle.ufsc.br/pluginfile.php/4760584/mod_resource/content/1/03_InfReg.pdf Acesso em: 06/03/2022.

NAKAMURA, Luiz. Análise de resíduos e transformações de variáveis. 2022. Disponível em: https://moodle.ufsc.br/pluginfile.php/4760593/mod_resource/content/1/04_Residuos.pdf Acesso em: 06/03/2022.

NAKAMURA, Luiz. Análise de regressão linear múltipla. 2022. Disponível em: https://moodle.ufsc.br/pluginfile.php/4760599/mod_resource/content/1/05_Multipla.pdf Acesso em: 06/03/2022.

Communities and Crime Unnormalized Data Set, UCI Machine Learning Repository. Disponível em: http://archive.ics.uci.edu/ml/datasets/Communities%20and%20Crime%20Unnormalized Acesso em: 06/03/2022