Relatório Otimização Não-Linear

Eduardo Zaffari Monteiro - 12559490

Gustavo Silva de Oliveira - 12567231

Lucas Ivars Cadima Ciziks - 12559472

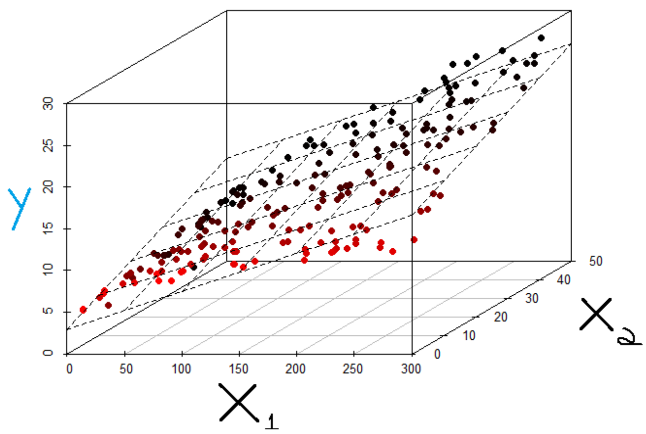
Pedro Henrique Maçonetto - 12675419

1. Introdução

Conforme definido por Montgomery (2021), *Regressão* é a metodologia responsável por estudar e compreender o **relacionamento estatístico** entre uma variável resposta (ou dependente) e suas variáveis preditoras (ou independentes). Esse método captura a correlação entre as variáveis em investigação e avalia se tal relação é estatisticamente significativa ou não. A técnica mais simples e comumente utilizada é a **Regressão Linear**, em que o relacionamento entre as variáveis é modelado de forma linear, procurando a reta ou plano que melhor se ajuste aos dados observados. O seguinte modelo com duas covariáveis ilustra a técnica:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- Y: Variável Resposta ou Dependente;
- X_1 e X_2: Variáveis Preditoras ou Independentes;
- β_0: Parâmetro que representa o intercepto;
- β_1 e β_2: Parâmetros cuja taxa de mudança quando as outras covariáveis são mantidas constantes;
- ε: Erro aleatório associado ao modelo.



Desse modo, a regressão busca estimar os parâmetros do modelo proposto. Isso pode ser realizado de diversos modos, mas o método mais disseminado é o **Método dos Mínimos Quadrados** ou *Least-Squares*. Seu procedimento consiste em minimizar a soma dos erros ao quadrado, encontrando estimadores **não-viesados** para os parâmetros do modelo e com a **menor variância** dentre todos os estimadores não-viesados (*Teorema de Gauss-Markov*). Para o modelo de duas covariáveis, o problema pode ser definido como:

$$min \sum_{i=1}^{n} \varepsilon_i^{\ 2} = \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2)^2$$

Essa otimização pode ser resolvida de forma analítica, todavia, o interesse aqui é em observar e avaliar o comportamento dos algoritmos de otimização não-linear vistos durante a disciplina no contexto da regressão linear. Assim, é possível comparar os resultados encontrados pelos métodos implementados com os resultados analíticos.

Como o problema se trata de uma **minimização irrestrita**, os algoritmos a serem utilizados funcionam gerando uma sequência de iterandos x_k a partir de um ponto inicial x_0 . Essa iteração se encerra quando a solução é encontrada com precisão suficiente ou não há mais como progredir. Nesse contexto, há dois tipos de estratégias para se decidir como movimentar do ponto x_k para x_{k+1}: **Busca Linear** e **Regiões de Confiança**.

Os algoritmos de busca linear calculam primeiramente a direção de procura, para assim encontrar, ao longo da direção, o tamanho do passo que minimiza a função de interesse e atinge o próximo x_{k+1}. Já os métodos de região de confiança invertem a ordem: determinam a maior distância permitida entre os iterandos (isto é, o raio da região de confiança) e depois encontram a direção de iteração. Para analisar o problema da regressão, implementamos um método de busca linear, o **Método de Newton**, e um método de região de confiança, o **Método dogleg**.

2. Modelagem do Problema

O objetivo da regressão é estimar os ***p*** parâmetros $\vec{\beta} = [\beta_0, \beta_1, ..., \beta_{p-1}]$ que melhor ajustam o modelo aos dados observados. Para realizar essa estimação, o método comumente utilizado é o **mínimos quadrados**, cujo propósito é minimizar a soma quadrática dos erros ou ruídos do modelo, denotado por $S(\vec{\beta})$. Portanto, o problema da regressão é equivalente a resolver o seguinte processo de otimização não-linear:

$$S(\vec{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_1 - ... - \beta_{p-1} X_{p-1})^2$$

$$Minimizar\ S(\vec{\beta})$$

com $\vec{\beta} \in R^p$ e $S \in R^p \to R$ é uma função suave.

Como é possível observar, a função de interesse é quadrática, assim caracterizando um problema de natureza **não-linear** e **irrestrita**. Nessa conjuntura, tanto os métodos de busca linear quanto de região de confiança podem ser utilizados para resolvê-lo. No presente projeto, implementamos o **Método de Newton**, como representante dos algoritmos de busca linear, e um método de região de confiança, o **Método dogleg**. Assim, será possível comparar se os métodos atingiram o resultado correto e o número de iterações necessários para chegar em tal conclusão.

3. Implementação

3.1 Método de Newton

O método de Newton é um método de busca linear que tem como objetivo encontrar a direção de descida ao fazer uma aproximação da função por uma parábola, e, dessa forma, cria uma sequência de gradientes que convergem para zero, desde que as condições de Wolfe sejam satisfeitas. Nesse contexto, encontramos a direção de descida utilizando a fórmula:

$$p_k = -(\nabla^2 f_k)^{-1} \nabla f_k$$

O método é bem definido e possui convergência quadrática contanto que a partir de certa iteração o tamanho do passo seja sempre 1, o que ocorre na otimização de funções quadráticas como o mínimos quadrados do problema. Ademais, o método tem garantia de resolver problemas de minimização de funções quadráticas em uma única iteração com passo de tamanho 1, o que o torna uma excelente escolha para o problema da regressão.

O método de Newton foi implementado na linguagem de programação *Python* em conjunto com o pacote numpy, que auxilia no cálculo e multiplicação de matrizes e vetores. Para realizar a implementação, a função do problema foi representada na **forma matricial**, definida da seguinte forma:

$$\sum_{i=1}^{n} (Y - \beta_0 - \beta_1 X_1 - ... - \beta_{p-1} X_{p-1})^2 => (\vec{Y} - \vec{X}\vec{\beta})^2$$

Expandindo a equação:

$$(\vec{Y} - \vec{X}\vec{\beta})^2 = (\vec{Y} - \vec{X}\vec{\beta})^T(\vec{Y} - \vec{X}\vec{\beta}) = \vec{Y}^T\vec{Y} - 2\vec{Y}^T\vec{X}\vec{\beta} + \vec{\beta}^T\vec{X}^T\vec{X}\vec{\beta}$$

- $\vec{Y}$: Vetor com os valores de resposta;
- $\vec{X}$: Matriz com as covariáveis e uma coluna de 1's para representar o intercepto;
- $\vec{\beta}$: Vetor de coeficientes estimados (Variável da minimização).
-

A partir da forma matricial, é possível definir o **gradiente** e a **hessiana**, elementos importantes no cálculo da direção p_k. Assim, com relação ao vetor $\vec{\beta}$ na forma:

$$\nabla S(\vec{\beta}) = -2\vec{Y}^T\vec{X} + \vec{X}^T\vec{X}\vec{\beta}$$

$$\nabla^2 S(\vec{\beta}) = 2\vec{X}^T\vec{X}$$

Essas relações matriciais facilitaram o cálculo do gradiente e da hessiana no código do projeto, sem necessidade de programação explícita ou funções de alta complexidade. O método de newton para o problema da regressão é baseado no seguinte pseudocódigo:

Pseudocódigo do Método de Newton

```
funcao metodo_de_newton(X_k):
  // X_k: X inicial, posteriormente será o iterando
  i = 0   // número de iterações
  a = 1   // tamanho do passo
  enquanto gradiente_f(X_k) != 0:
    // Produto matricial
    p = hessiana_f(X_k)^(-1) * gradiente_f(X_k)
    a = alpha(X_k,p)
    X_k = X_k + a * p
    i += 1
  return X_k, i // retorna X_min e o número de iterações
```

Em que *alpha* é uma função responsável por calcular a solução analítica para o tamanho do passo de funções quadráticas convexas, e as outras funções são o código das formas matemáticas já apresentadas.

3.2 Método dogleg

O método dogleg é um método de otimização não-linear que tem como objetivo encontrar um mínimo local por meio de regiões de confiança. Assim, o problema original da regressão envolvendo os mínimos quadrados pode ser representado por:

$$\text{Minimizar} \quad m_k(p) = f(x_k) + \nabla f(x_k)^T p + 1/2 p^T B_k p$$

$$\text{Sujeito a} \quad ||p|| \leq \Delta_k$$

em que $m_k(p)$ é um modelo quadrático da função original e Δ_k é o raio da região do modelo. Para usufruir do método em questão, é necessário que B_k seja **definida positiva**, nesta situação, a solução irrestrita do problema é dada pelo passo completo definido por:

$$p_k^B * = -B_k^{-1} \nabla f(x_k)$$

Entretanto, é preciso minimizar o problema levando em conta as restrições, trocando a curva correta por uma aproximação composta por dois segmentos de reta, em que o primeiro segmento é dado por:

$$p_k^U = \frac{-\nabla f(x_k)^T \nabla f(x_k)}{\nabla f(x_k)^T B_k \nabla f(x_k)} \nabla f(x_k)$$

Enquanto o segundo segmento encaminha de p_k^U até o passo completo p_k^B. Neste contexto, é possível definir por final a trajetória dogleg da seguinte forma:

$$\tilde{p}_k(\tau) = \tau p_k^U, \quad 0 \leq \tau \leq 1,$$

$$\tilde{p}_k(\tau) = p_k^U + (\tau - 1)(p_k^B - p_k^U), \quad 1 \leq \tau \leq 2.$$

O método busca então escolher a direção p_k que minimiza o modelo quadrático nesta trajetória sujeito a p_k permanecer na região de confiança.

Para implementar o método no projeto, utilizou-se a biblioteca SciPy que disponibiliza uma variedade de algoritmos e métodos científicos, incluindo métodos de minimização vistos no decorrer da disciplina. Assim, aplicando o algoritmo *scipy.optimize.minimize* foi necessário estabelecer a função original do mínimos quadrados, o ponto inicial, o gradiente e hessiana, além de especificar que o método dogleg seria utilizado. Assim, foi possível executá-lo nas instâncias do problema para realizar as comparações propostas pelo projeto.

4. Resultados Numéricos

Após implementar os métodos propostos, foram selecionadas duas instâncias do problema para aplicá-los e comparar os resultados obtidos da otimização não-linear com o resultado analítico da regressão linear.

4.1. Produção de Uma Fábrica

O primeiro problema escolhido trata-se de uma tentativa de explicar a produção de uma fábrica com base na temperatura e na concentração de seu produto. Ao estimar os parâmetros da reta de regressão, tanto o método de Newton quanto o Dogleg apresentaram os seguintes resultados:

$$\beta_0 = 88.14\,;\ \beta_1 = 0.8925\,;\ \beta_2 = 2.532$$

Assim como esperado, o método de Newton encontrou esses valores em uma **única iteração**, pelo fato do problema ser de natureza quadrática e o método se adequar bem nesse contexto. Em compensação, foram necessárias **7 iterações** para que o método Dogleg atingisse o mesmo resultado.

Ao resolver o problema pela solução analítica do modelo de regressão, o mesmo resultado foi encontrado. Desse modo, é possível concluir que ambas as soluções encontraram corretamente os mínimos locais do problema, ou seja, estimaram adequadamente os parâmetros da reta de regressão.

4.2. Study on the Efficacy of Nosocomial Infection Control (SENIC)

O segundo problema designado trata-se de um estudo sobre a eficácia no controle de infecção nosocomial, isto é, infecções adquiridas durante a internação ou em procedimentos hospitalares. O objetivo da regressão é explicar o tempo que o paciente passa dentro do hospital em função das seguintes covariáveis: idade do paciente, probabilidade do risco de infecção, número de enfermeiras, cultura e número de camas oferecidas.

Esse problema apresenta um grau de dificuldade maior que o inicial, uma vez que mais covariáveis e parâmetros estão envolvidos. Novamente, tanto o método de Newton quanto o Dogleg obtiveram os mesmo estimadores:

$$\beta_0 = 1.006\,;\ \beta_1 = 0.098\,;\ \beta_2 = 0.532\,;\ \beta_3 = 0.023\,;\ \beta_4 = 0.006\,;\ \beta_5 = -\ 0.005$$

Por se tratar de uma função quadrática, no método de Newton, a solução foi encontrada em apenas **uma iteração**, enquanto pelo Dogleg foram necessárias **2 iterações** para atingir o resultado. Embora esse problema apresente grau maior de complexidade, por envolver mais dados, covariáveis e estimadores, os parâmetros encontrados são os mesmos da solução analítica, fortificando a eficácia dos métodos em teste.

5. Conclusão

Como previsto, para ambos os problemas, o método de Newton encontrou a solução em uma única iteração, enquanto o Dogleg precisou de mais iterações para atingir a resposta. Entretanto, mesmo havendo uma diferença no número de interações, e consequentemente no tempo de execução, os dois métodos convergiram para uma mesma solução em todos os casos.

Assim, é possível concluir que, para o problema de minimização do erro quadrático de uma regressão linear, o método de Newton é mais eficiente do que o método Dogleg, uma vez que ele encontra a solução em uma única iteração, o que era esperado já que se trata de um problema quadrático. Portanto, ambos os métodos foram eficazes e suficientes para estimar os parâmetros da regressão.

6. Referências

As seguintes referências foram utilizadas para o desenvolvimento do projeto:

- Material preparado e disponibilizado pela professora Marina Andretta.
- Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear statistical models*.
- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
- Chan, H. S. "*Introduction to Probability for Data Science.*" (2021).