Trabalho Prático de Cálculo Numérico

Universidade Federal de São João Del Rei – 2019/2
Curso: Ciências da Computação | Professor: Vinícius da Fonseca Vieira
Aluno: Felipe Henrique Faria | Matrícula: 152050049

Introdução

Este trabalho tem como objetivo implementar um algoritmo para o Métodos do Mínimos Quadrados estudado em sala de aula. O algoritmo será aplicado a um problema de regressão linear envolvendo múltiplas variáveis, o que envolve também implementação de algoritmos para operações algébricas, como: multiplicação de matrizes, matrizes transpostas e resolução de sistemas de equações lineares.

Descrição da base de dados

A base de dados escolhida para o trabalho foi "Auto MPG Data Set (University, 2019)", disponível no UCI (Center for Machine Learning and Intelligent Systems). Esta base consiste em informações sobre o consumo de combustível de diferentes carros.

Variáveis de entrada

- Cylinders: Número de cilindros do motor (discreta);
- Displacement: Distância percorrida pelo veículo (contínua);
- Horsepower: Potência do motor em cavalos (contínua);
- Weight: Peso do veículo (contínua);
- Acceleration: Aceleração média do veículo (contínua);
- Model year: Ano de fabricação (discreta).

Existem outras variáveis como origem de fabricação e modelo do veículo, no entanto elas não foram consideradas para esta aplicação por não apresentarem um formato adequado.

Variável de saída

A variável de saída é o consumo de combustível dos veículos medido em MPG (milhas por galão). O objetivo é estimar um valor aproximado desta variável considerando as variáveis de entrada descritas acima.

Descrição da solução

Operações algébricas

O primeiro passo foi criar uma biblioteca para as operações algébricas mais usuais, como:

- Multiplicação de matrizes;
- Matriz transposta;
- Matriz aumentada do sistema.

Estas funções já estão disponíveis na biblioteca *Numpy* porém para melhor aprendizado foram recriadas neste trabalho.

Solução de sistemas de equações lineares

Para solução dos sistemas de equações lineares foi escolhido o método de eliminação de Gauss que consiste dos seguintes passos:

1. Obter uma matriz aumentada do sistema na forma *[A|b]*;
2. Transformar a matriz aumentada em uma matriz triangular superior;
3. Resolver o sistema por substituição regressiva.

Separação da base de dados

A base de dados foi dividida em duas partes, uma para treino de 70% e uma para teste com os 30% dos dados restantes.

Além disso o conteúdo da base de dados foi embaralhado, isto porque foi observado que a base estava ordenada em relação ao ano de fabricação dos veículos, o que prejudicaria os resultados do algoritmo.

Método dos Mínimos Quadrados

Com as funções para operações algébricas e solução de sistemas lineares devidamente implementadas, agora é possível aplicar o método dos mínimos quadrados, esta será a etapa de treinamento do algoritmo que consiste das seguintes etapas:

1. Obter uma matriz *G* contendo os valores das variáveis de entrada;
2. Obter uma matriz Y contendo os valores de saída;
3. Gerar uma matriz transposta de G, apelidada *Gt*;
4. Multiplicar *Gt* por G e obter a matriz *GtG*;
5. Multiplicar *Gt* por Y e obter a matriz *GtY*;
6. Gerar a matriz aumentada do sistema *E* juntando *GtG* e *GtY*;
7. Resolver o sistema linear referente a matriz aumentada do sistema.

O resultado do sistema linear é uma equação que se aplicada a novas variáveis de entrada retorna um valor estimado para a variável de saída.

Testes

Com a equação gerada pela etapa de treinamento é possível estimar os valores para novas entradas da base de dados. Os valores estimados são comparados aos da base afim de medir a eficiência do algoritmo pelo cálculo de erros, sendo eles o erro médio absoluto e o erro médio quadrático.

Análise dos resultados

Antes de analisar a regressão para múltiplas variáveis, foram realizadas regressões simples para cada variável de entrada afim de comparar os dois métodos e perceber a diferença no impacto do resultado entre uma variável e outra.

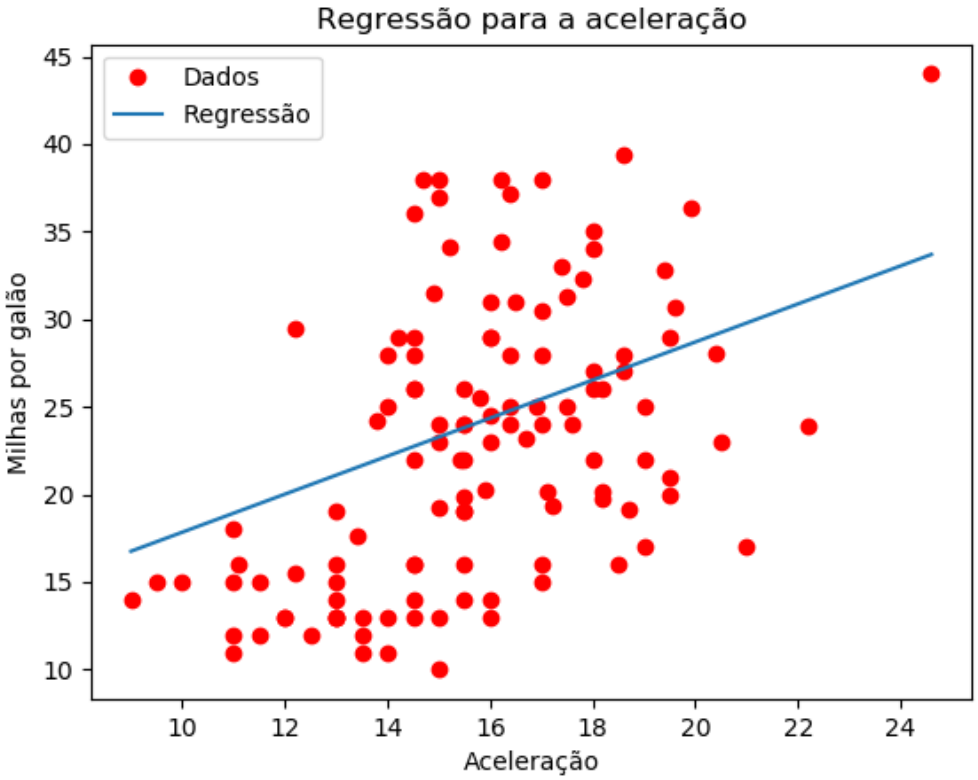

Figura 1 - Regressão para a aceleração

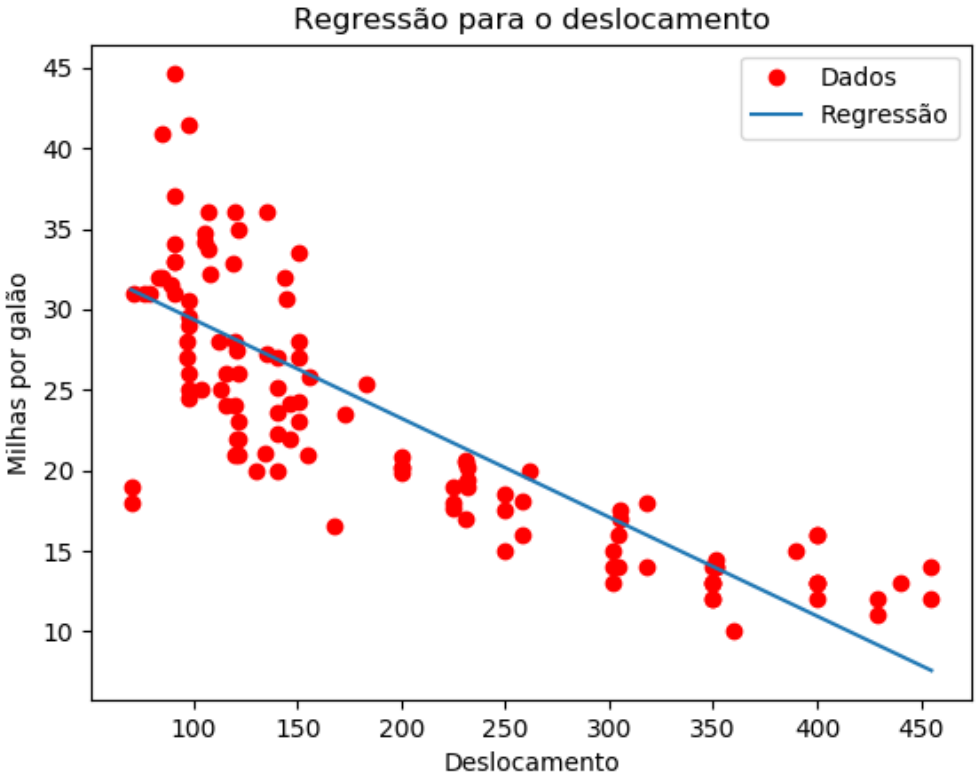

Figura 2 - Regressão para o deslocamento

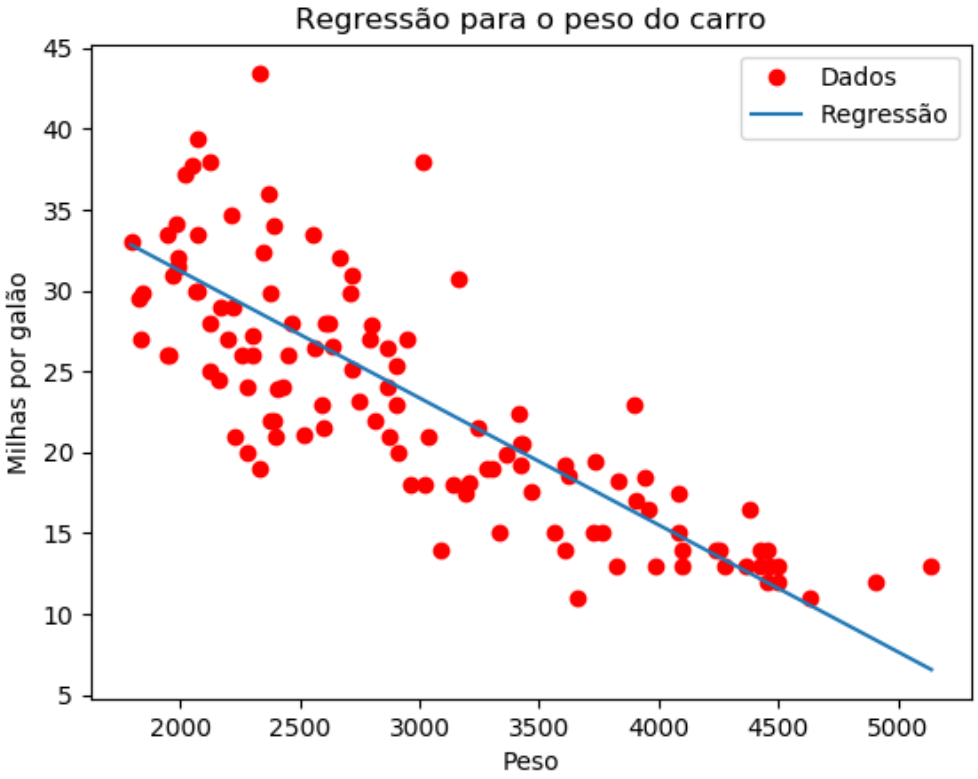

Figura 3 - Regressão para o peso do carro

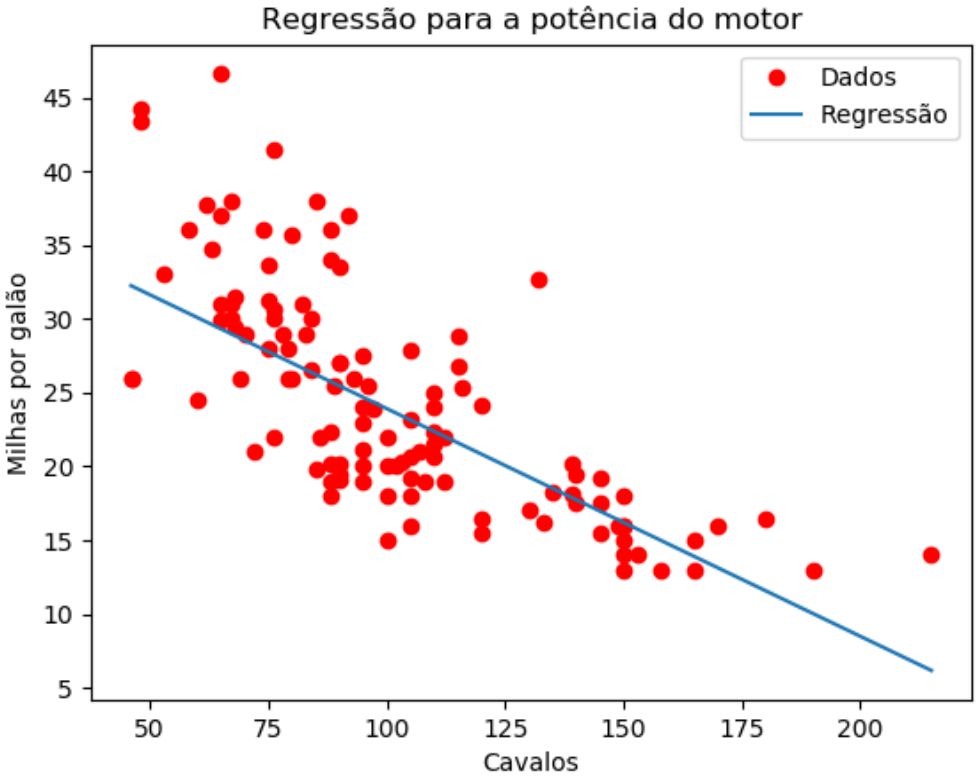

Figura 4 - Regressão para a potência do motor

Os erros também variam entre as variáveis de entrada conforme exibido na Tabela 1. É possível observar que a variável *peso* gera menor erro e a variável *aceleração* o maior.

Tabela 1 - Erros para cada variável de entrada

Variável	Erro médio absoluto	Erro médio quadrático
Cilindros	3.5940	19.8554
Deslocamento	3.4270	20.0698
Potência	3.9792	27.6361
Peso	3.3249	19.3185
Aceleração	5.9159	49.9799

Após executar o algoritmo separadamente para cada variável o mesmo é aplicado novamente em modelo de regressão múltipla, considerando todas as variáveis. Com isto foi possível diminuir o erro e melhorar a precisão das estimativas, conforme analisado na Tabela 2.

Tabela 2 - Erro para múltiplas variáveis

Variável	Erro médio absoluto	Erro médio quadrático
Todas	2.5713	11.0049

Os resultados podem ser observados no gráfico a seguir onde foram estimados os valores para 20 veículos aleatórios, as colunas em azul representam os dados originais e as colunas em laranja os dados correspondentes estimados pelo algoritmo.

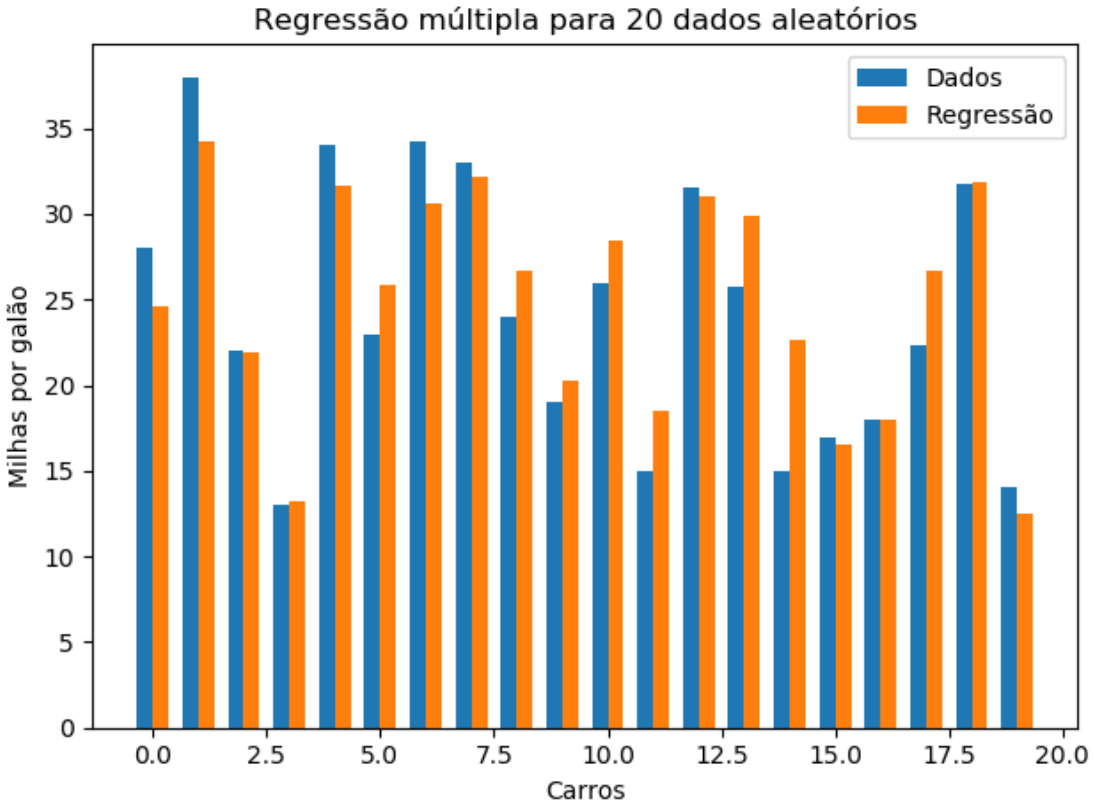

Figura 5 - Regressão múltipla para 20 dados aleatórios

A equação gerara pelo método dos mínimos quadrados foi:

$P(x) = -12.4535 - x_1*0.7132 + x_2*0.0184 - x_3*0.0084 – x_4*0.0069 + x_5*0.1591 + x_6*0.7278$

Tecnologias utilizadas

O trabalho foi inteiro desenvolvido na linguagem Python 3.7. Para plotagem de gráficos e visualização dos resultados foram utilizadas duas bibliotecas adicionais: *Numpy* e *Matplotlib*.

Limitações

A principal limitação encontrada foi tratar os dados que continham variáveis faltando ou em formatos inadequados. Como eram poucos estes dados foram desconsiderados.

Além disso algumas varáveis como modelo do carro não foram consideradas, porém poderiam ter impacto no resultado final e melhorar a precisão dos resultados estimados.

Conclusão

Com este trabalho foi possível compreender e aplicar o conteúdo visto em sala de aula. Além disso foi possível exercitar de forma computacional conceitos de álgebra para matrizes e solução de sistemas, e principalmente o trabalho proporcionou uma ótima introdução à ciência de dados e aprendizado de máquina.

Bibliografia

University, C. M. (21 de 11 de 2019). *Auto MPG*. Fonte: UCI: https://archive.ics.uci.edu/ml/datasets/Auto+MPG