10 open datasets for linear regression

AI DATA POSTED JANUARY 1, 2021

Every data scientist will likely have to perform linear regression tasks and predictive modeling processes at some point in their studies or career. For those of you looking to learn more about the topic or complete some sample assignments, this article will introduce open linear regression datasets you can download today. Additionally, some of the datasets on this list include sample regression tasks for you to complete with the data.

Linear regression datasets for machine learning

1. Cancer linear regression

This dataset includes data taken from cancer.gov about deaths due to cancer in the United States. Along with the dataset, the author includes a full walkthrough on how they sourced and prepared the data, their exploratory analysis, model selection, diagnostics and interpretation.

2. CDC data: nutrition, physical activity, obesity

From the Behavioral Risk Factor Surveillance System at the CDC, this dataset includes information about physical activity, weight and average adult diet.

3. Fish market dataset for regression

Built for multiple linear regression and multivariate analysis, the Fish Market Dataset contains information about common fish species in market sales. The dataset includes the fish species, weight, length, height and width.

4. Medical insurance costs

This dataset was inspired by the book *Machine Learning with R* by Brett Lantz. The data contains medical information and costs billed by health insurance companies. It contains 1338 rows of data and the following columns: age, gender, BMI, children, smoker, region and insurance charges.

5. New York Stock Exchange dataset

Created as a resource for technical analysis, this dataset contains historical data from the New York stock market. The dataset comes in four CSV files: prices, prices-split-adjusted, securities and fundamentals. Using this data, you can experiment with predictive modeling, rolling linear regression and more.

6. OLS regression challenge

The OLS regression challenge tasks you with predicting cancer mortality rates for US counties. The dataset contains data from cancer.gov, clinicaltrials.gov, and the American Community Survey. It is in CSV format and includes the following information about cancer in the US: death rates, reported cases, US county name, income per county, population, demographics and more.

7. Real estate price prediction

This real estate dataset was built for regression analysis, linear regression, multiple regression, and prediction models. It includes the date of purchase, house age, location, distance to nearest MRT station, and house price of unit area.

8. Red wine quality

From the UCI Machine Learning Repository, this dataset can be used for regression modeling and classification tasks. The dataset includes info about the chemical

properties of different types of wine and how they relate to overall quality.

9. Vehicle dataset from CarDekho

A useful dataset for price prediction, this vehicle dataset includes information about cars and motorcycles listed on CarDekho.com. The data is in a CSV file which includes the following columns: model, year, selling price, showroom price, kilometers driven, fuel type, seller type, transmission and number of previous owners.

10. WHO statistics on life expectancy

This dataset contains information compiled by the World Health Organization and the United Nations to track factors that affect life expectancy. The data contains 2938 rows and 22 columns. The columns include: country, year, developing status, adult mortality, life expectancy, infant deaths, alcohol consumption per capita, country's expenditure on health, immunization coverage, BMI, deaths under 5-years-old, deaths due to HIV/AIDS, GDP, population, body condition, income information and education.

Machines can't learn without data. But don't fear; if you're looking for more datasets, we've got you covered. Check out this compilation of the 50 best free datasets for machine learning.

Learn more

Related insights

Brochure

AI DATA AUTOMOTIVE

High-quality driving dataset for training and validating ADAS and AV models

AI DATA

Fraud prevention best practices when crowdsourcing AI data

Brochure

AI DATA

Fraud prevention in crowdsourcing AI training data

Solutions

Industries

About Us

Humanity-in-the-Loop

Insights

Careers

Contact

Subscribe to Newsletter

WillowTree, a TELUS Digital Company

Cookie Preferences Do Not Sell my Personal Information

Website User Terms

Privacy Policy

TELUS Digital is a subsidiary of:
TELUS Communications Inc.