Linear Regression – Practice Areas

Practice Problems:

1. Boston Housing Dataset

Boston Housing Dataset contains information collected by U.S Census Service about houses in the Boston area. It includes various attributes such as the crime rate, average number of rooms, proportion of non-retail business, etc.

- **Dataset Source:** Boston Housing Dataset
- **Labels**: Continuous values representing median value of owner-occupied homes (in $1000s).
- **Size**: 506 samples each with 14 attributes.
- **Data**: Numerical data

2. Advertising Dataset

This dataset contains data about the sales of a product in relation to the advertising budgets spent on TV, radio and newspaper. It's commonly used to explore the relationship between advertising efforts and sales.

- **Dataset Source:** Advertising Dataset
- **Labels**: Continuous values representing sales of the product (in thousands of units).
- **Size**: 200 samples each with 4 attributes.
- **Data**: Numerical data

3. California Housing Dataset

It is collected by U.S. Census in 1990 and this dataset includes various attributes for California districts such as median house value, median income, housing median age, total rooms, total bedrooms, population, households, latitude and longitude.

- **Dataset Source**: California Housing Dataset
- **Labels**: Continuous values representing the median house value (in $1000s).
- **Size**: 20,640 samples each with 9 attributes.
- **Data**: Numerical data

4. Auto MPG Dataset

This dataset contains data on the fuel consumption (miles per gallon) of various car models along with other attributes like engine displacement, horsepower, weight, acceleration and model year.

- **Dataset Source**: Auto MPG Dataset
- **Labels**: Continuous values representing miles per gallon (mpg).
- **Size**: 398 samples, each with 8 attributes.
- **Data**: Numerical data

5. Diabetes Dataset

This dataset includes medical predictor variables and one target variable that is quantitative measure of disease progression one year after baseline. It is used to predict the progression of diabetes based on factors such as age, sex, BMI, blood pressure and six blood serum measurements.

- **Dataset Source**: Diabetes Dataset
- **Labels**: Continuous values representing disease progression after one year.
- **Size**: 442 samples each with 10 attributes.
- **Data**: Numerical data

6. Fish Market Dataset

This dataset includes data on the common fish species in fish market sales. Attributes include weight, length, height and width of fish used to predict fish weight based on these physical characteristics.

- **Dataset Source**: Fish Market Dataset
- **Labels**: Continuous values representing the weight of the fish (in grams).
- **Size**: 159 samples each with 7 attributes.
- **Data**: Numerical data

7. Wine Quality Dataset

This dataset contains various chemical properties of wine such as acidity, residual sugar, chlorides and sulfur dioxide levels and quality ratings. It is often used to predict wine quality based on these chemical properties.

- **Dataset Source**: Wine Quality Red Dataset, Wine Quality White dataset
- **Labels**: Continuous values representing the quality score of wine.
- **Size**: Red wine: 1,599 samples; White wine: 4,898 samples. Each with 12 attributes.
- **Data**: Numerical data

8. Insurance Charges Dataset

This dataset includes information about medical charges billed by health insurance companies with features like age, sex, BMI, children, smoker status, region and the charges billed.

- **Dataset Source**: Insurance Charges Dataset
- **Labels**: Continuous values representing individual medical costs.
- **Size**: 1,338 samples each with 7 attributes.
- **Data**: Numerical data

9. Salary Dataset

This dataset contains information on years of experience and the corresponding salary which is useful for predicting salary based on experience.

- **Dataset Source**: Salary Dataset
- **Labels**: Continuous values representing salary.
- **Size**: 30 samples each with 2 attributes.
- **Data**: Numerical data

10. Energy Efficiency Dataset

This dataset provides data on the energy efficiency of buildings including features such as relative compactness, surface area, wall area, roof area, overall height, orientation, glazing area and more. It is used to predict the heating and cooling load requirements of buildings.

- **Dataset Source**: Energy Efficiency Dataset
- **Labels**: Continuous values representing heating and cooling loads (energy efficiency measures).
- **Size**: 768 samples each with 8 attributes.
- **Data**: Numerical data

11. Stock Market Dataset

This dataset contains historical stock market data for various companies including attributes such as opening price, closing price, high, low, trading volume and other financial indicators. It is useful for financial market predictions and trend analysis.

- **Dataset Source**: Stock Market Dataset
- **Labels**: Continuous values for attributes like closing price (could be a target variable).
- **Size**: Typically contains thousands to millions of records depending on the dataset.
- **Data**: Numerical data

12. Customer Churn Dataset

This dataset includes information about customers of a business with labels indicating whether the customer churned (stopped using the service) or not. Attributes include demographics, usage patterns and customer satisfaction scores which can be used to predict churn.

- **Dataset Source**: Customer Churn Dataset: Training, Customer Churn Dataset: Testing
- **Labels**: Binary values indicating whether a customer has churned (yes/no).
- **Size**: Varies but typically contains thousands of records.
- **Data**: Numerical and Categorical data.

13. Student Performance Dataset

This dataset contains information about students' academic performance including attributes such as study time, previous grades, socioeconomic status and demographic factors. It is useful for predicting student performance based on these factors.

- **Dataset Source**: Student Performance Dataset
- **Labels**: Continuous values representing grades or binary labels for passing/failing.
- **Size**: 1,000 samples each with 33 attributes.
- **Data**: Numerical and Categorical data.

14. 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.

15. 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.

16. 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.

17. 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.

18**.** 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.

19. 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.

20. 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.

21. 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.

22. 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.

23. 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.

24. Other dataset for practice for linear regression:

https://www.kaggle.com/datasets?tags=13405-Linear+Regression

(Unlimited practice for linear regression)

All the best.

Let's refine and refine.

Thanks.