

ASSIGNMENT # 4 ADITYA SHROFF

BUS 4023 - FALL 2024

5POINTS

CHAPTER 6 – MULTIPLE LINEAR REGRESSION

ASSIGNMENT QUESTIONS

PROBLEM #1, PARTS A TO D

Requires both written response and Python code saved as Jupyter notebook (.ipynb). Please save a copy of your python output and code in pdf format as well.

(a) Why should the data be partitioned into training and validation sets?

To guarantee that the model performs properly when applied to new data, the data must be divided into training and validation sets. The model learns the associations between predictors and the target variable by fitting it to the training set. By evaluating the model's performance on unseen data, the validation set helps identify overfitting and makes sure the model is capable of making accurate predictions on fresh data rather than merely remembering patterns.

The model might function well on training data but not generalize to real-world situations if it is not properly partitioned.

(b) Multiple Linear Regression Model

The estimated regression equation is:

$$\text{MEDV} = -30.52 - 0.26 \times \text{CRIM} + 3.88 \times \text{CHAS} + 8.55 \times \text{RM}$$

Where:

DELIVERABLES FOR ASSIGNMENTS INCLUDE

...

#1) Python code and

output saved as notebook [in Jupyter notebook click on *File* >

Download as > Notebook (.ipynb)]

SAVE THE NOTEBOOK/WORD/PDF FILES IN THE FOLLOWING FORMAT:

WEEKLY_ASSIGNMENT_
X_QY_LL_F

Where,

X=weekly assignment number

Y=assignment problem number

LL=your full last name

F=Initial of your first name

Please remember to add the following on the first line of the notebook as a comment:

- your full last name and first name
- the assignment problem number the notebook code and output pertains to

Weekly Assignment

- **Intercept** = -30.52 (Baseline house price when all predictors are zero)
- **CRIM Coefficient** = -0.26 (Higher crime rates lead to lower house prices)
- **CHAS Coefficient** = 3.88 (Homes near the Charles River tend to be more expensive)
- **RM Coefficient** = 8.55 (More rooms per dwelling increase house prices)

(c) Predict the Median House Price

$$\text{MEDV} = -30.52 - (0.26 \times 0.1) + (3.88 \times 0) + (8.55 \times 6)$$

$$\text{MEDV} = 20.72$$

Thus, the **predicted median house price** is \$20,720.

(d) Reduce

Among the 13 predictors in the Boston Housing dataset, some are highly correlated and likely measure similar aspects of urbanization, pollution, and socioeconomic conditions. Based on the correlation matrix, the following groups of variables are likely measuring the same underlying factors:

1. Urbanization and Industrialization:

- **INDUS** (Proportion of non-retail business acres per town)
- **NOX** (Nitric oxide concentration - parts per 10 million)
- **TAX** (Full-value property tax rate per \$10,000)
- **RAD** (Index of accessibility to radial highways)
- **AGE** (Proportion of owner-occupied units built before 1940)

These variables tend to be correlated because more industrialized or urban areas often have higher pollution (NOX), higher property taxes (TAX), greater accessibility to highways (RAD), and older housing stock (AGE).

2. Housing Quality and Socioeconomic Status:

- **RM** (Average number of rooms per dwelling)
- **LSTAT** (Percentage of lower-status population)
- **PTRATIO** (Pupil-teacher ratio by town)

These variables capture housing quality and socioeconomic conditions, as neighborhoods with larger homes (higher RM) tend to have wealthier residents (low LSTAT) and better school systems (low PTRATIO).

Relationships Among INDUS, NOX, and TAX

1. INDUS and NOX (Correlation = 0.76, Strong Positive Relationship)

- Higher INDUS values indicate more industrial and commercial areas.
- Such areas tend to have higher pollution levels (NOX) due to factories, traffic, and emissions.
- This strong correlation suggests that INDUS can be a proxy for air pollution levels, meaning keeping both variables may introduce redundancy in the model.

2. INDUS and TAX (Correlation = 0.72, Strong Positive Relationship)

- Industrial and urban areas typically have higher property tax rates (TAX) because of increased government services, infrastructure, and public expenditures.
- This means INDUS and TAX both measure urban density and industrial influence to some extent.

3. NOX and TAX (Correlation = 0.67, Moderate-Strong Positive Relationship)

- Higher pollution levels (NOX) tend to be found in urbanized areas with higher property taxes (TAX).
- This is because cities often have both higher emissions and higher government costs, leading to higher taxation rates.

Conclusion: Which Predictor to Remove?

Since INDUS, NOX, and TAX measure similar urbanization-related characteristics, we may consider removing one of them to reduce multicollinearity.

- If we are interested in pollution effects → Keep NOX, remove INDUS or TAX.
- If we are focusing on economic factors → Keep TAX, remove NOX or INDUS.
- If we want general urbanization measures → Keep INDUS, remove NOX or TAX.

✿ INDUS and NOX: $r=0.76$ (strong correlation)

✿ INDUS and TAX: $r=0.72$ (strong correlation)

✿ NOX and TAX: $r=0.67$ (moderate correlation)

(e)

- ⊙ **Neighborhood Factors:** Crime rates, school quality, and accessibility to public transportation.
- ⊙ **Market Conditions:** Mortgage rates, interest trends, and overall market dynamics.
- ⊙ **Property Features:** Number of bathrooms, available parking spaces, and garden size.

PROBLEM #4, PARTS A & B

*Requires both written response and Python code saved as Jupyter notebook (.ipynb).
Please save a copy of your python output and code in pdf format as well.*

(a)

- Training Set (50%): For calibrating the model.
- Validation Set (30%): For fine tuning parameters and avoiding overfitting.
- Test Set (20%): Presents a start-to-finish objective test of the model.

(b)

The following predictors have been used:

- Age_08_04 (Car Age)
- KM (Mileage)
- Fuel_Type (Petrol, Diesel, CNG)
- HP (Horsepower)
- Automatic Transmission
- Doors, Quarterly taxation, Manufacturer guarantee
- Features like cassette player, sport, and air conditioner, etc.

Since "Fuel_Type" is a categorical, we have encoded it in numbers using One-Hot Encoding.

Model Performance:

- Mean Absolute Error (MAE): 911.88

Weekly Assignment

- Root Mean Squared Error (RMSE): 1271.21
- R-squared Score (R^2): 0.89