DSCI531: Fairness in Artificial Intelligence

Homework 1: Linear Models

Due Date: 27 January 2025 at 4 PM
Submit your solutions to Brightspace in a Python Jupyter Notebook. Make sure the outputs are visible.

1. Exploratory Data Analysis [40 points]

In this homework, you are given a dataset `data_la_happiness.csv` of happiness (`meanvalence`) in different census tracts in Los Angeles county. This dataset assesses the influence of socio-demographic features on happiness. You are given four socio-demographic features:

- `meanHSize` - Mean household size
- `percent_bachelorPlus` - Percent of census tract population with a Bachelor's degree or higher
- `totalRace1` - Number of people from race 1
- `totalRace2` - Number of people from race 2

1.1 Data Sanctity [8 points]

Load the file `data_la_happiness.csv`. Check for missing values. Note which columns have missing values and how many?If there are no missing values, proceed to 1.2. Otherwise, address how you can handle missing values and implement it. Report mean and medians of all variable columns - meanvalence, totalRace1, totalRace2, percent_bachelorPlus.

1.2 Outlier Detection [8 points]

Generate a boxplot to visualize outliers in the outcome variable `meanvalence`. Use the inter-quartile range (IQR) method to remove census tracts with values less than $1.5 \times$ IQR below Q1 or greater than $1.5 \times$ IQR above Q3.

1.3 Variable Relationships [8 points]

Use `seaborn`'s `pairplot` to analyze relationships between the variables. Report correlations between variables. Discuss the distributions of `totalRace1` and `totalRace2`.

1.4 Simple Models and Residual Analysis [8 points]

Build two simple models using `statsmodels` and generate residual plots for each:

$$\texttt{meanvalence} \sim \texttt{totalRace1}$$
$$\texttt{meanvalence} \sim \texttt{totalRace2}$$

What do the residual plots reveal? Are any linear regression assumptions violated?

1.5 Log Transformation [8 points]

Apply the log transformation to `totalRace1` and `totalRace2`. Explain briefly why this transformation is necessary.

2. Multivariate Regression [40 points]

2.1 Model Building [15 points]

After applying the log transformation, build the following multivariate regression model using `statsmodels`:

$$\texttt{meanvalence} \sim \texttt{percent\_bachelorPlus} + \log(\texttt{totalRace1}) + \log(\texttt{totalRace2})$$

Report your findings and discuss interpretations for each independent variable.

2.2 Scatterplot Analysis [10 points]

Generate a scatterplot between `meanvalence` and `percent_bachelorPlot`, colored by log(`totalRace2`). Repeat this with predicted values (`predicted_meanvalence`) from the model. Discuss any differences observed.

2.3 Correlation Heatmap [15 points]

Generate a correlation heatmap using `seaborn`. Report correlations between:

- log(`totalRace1`) and `meanvalence`
- log(`totalRace2`) and `meanvalence`
- log(`totalRace1`) and `predicted_meanvalence`
- log(`totalRace2`) and `predicted_meanvalence`

Discuss whether the model appears biased.

3. Analyzing Bias in the Model [20 points]

3.1 Protected Variable Analysis [15 points]

Run the following model, considering log(`totalRace2`) as a protected variable:

$$\texttt{meanvalence} \sim \texttt{percent\_bachelorPlus} + \log(\texttt{totalRace1})$$

Report regression results and correlations as in Section 2.3. Compare these correlations to those in 2.3.

Discuss whether bias was reduced? [5 points].