Linear Regression - Revisit

Agenda

1. Machine Learning and Supervised Learning
2. Simple and Multiple Linear Regression
3. Evaluation Metrics for a Regression Model
4. Assumptions of Linear Regression
5. Statistical Inferences from a Linear Regression Model
6. ReCell Project
   a. Business Context and Objectives
   b. Dataset
   c. Solution Approach
   d. Q/A

Machine Learning and Supervised Learning

- Machine Learning (ML) is the ability of a computer to do some task without being explicitly programmed using an underlying model which is the result of the learning process

- The model is generated by learning from huge volumes (both in breadth and depth) of historical data reflecting the real world in which the processes are performed

- Supervised learning refers to the class of MLS models which are built using data that contains both the inputs and the desired output (labels or ground truth)

- There are basically two types of supervised learning:
  - Regression - where the desired output is in the form of continuous values
    - **e.g.** predicting the house prices based on some features like area, the number of rooms, etc.
  - Classification - where the desired output is in the form of categorical values
    - **e.g.** predicting if the person is likely to default on a loan based on the features like age, past transactions, etc.

- The model learns from the training data using the target variable(s) as reference and the model thus generated is used to make predictions about data not seen by the model before

Simple and Multiple Linear Regression

- Linear regression is a way to identify a relationship between the independent variable(s) and dependent variable
- We can use these relationships to predict values for one variable for the given value(s) of the other variable(s)
- The variable which is used in prediction is termed as independent/explanatory/regressor, and the predicted variable is termed as dependent/target/response variable.
- In the case of linear regression with a single explanatory variable, the linear combination can be expressed as

  target = intercept + constant * explanatory variable

- Multiple linear regression is just the extension of the concept of simple linear regression with one variable
- It can be represented by

  target = intercept + constant1*feature1 + constant2*feature2 + constant3*feature3 + ...

- The model aims to find the constants and intercept such that the hyperplane is the best fit

Evaluation Metrics for a Regression Model

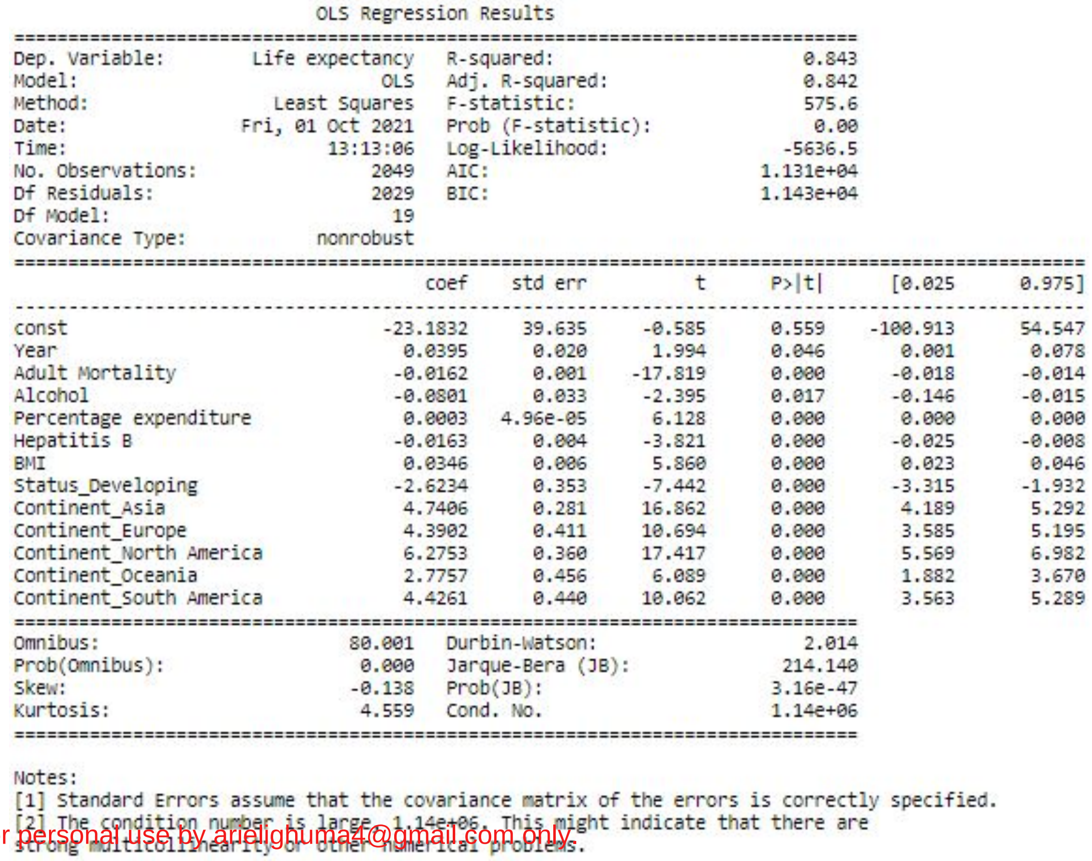
R-squared	Adjusted R-squared	Mean Absolute Error	Root Mean Square Error
● Measure of the % of variance in the target variable explained by the model ● Generally the first metric to look at for linear regression model performance ● Higher the better	● Conceptually, very similar to R-squared but penalizes for the addition of too many variables ● Generally used when you have too many variables as adding more variables always increases R^2 but not Adjusted R^2 ● Higher the better	● Simplest metric to check prediction accuracy ● Same unit as the dependent variable ● Not sensitive to outliers i.e. errors doesn't increase too much if there are outliers ● Difficult to optimize from a mathematical point of view (pure maths logic) ● Lower the better	● Another metric to measure the accuracy of prediction ● Same unit as the dependent variable ● Sensitive to outliers - errors will be magnified due to the square function ● But has other mathematical advantages that will be covered later ● Lower the better

Assumptions of Linear Regression

Assumption	How to test	How to fix
No multicollinearity in independent variables	Heatmaps of correlations or VIF (Variance inflation factor)	Remove correlated variables
There should be a linear relationship between dependent and independent variables	Plot residuals vs. fitted values and check the plot	Transform variables that appear non-linear (log, square root, etc.)
The residuals should be independent of each other	Plot residuals vs. fitted values and check the plot	Transform variables (log, square root, etc.)
Residuals must be normally distributed	Plot residuals or use Q-Q plot	Non-linear transformation of the independent or dependent variable
No heteroscedasticity, i.e., residuals should have constant variance	Use statistical test (like goldfeldquandt test)	Non-linear transformation of the dependent variable or add other important variables

Statistical Inferences from a Linear Regression Model

- The 95% confidence interval for Alcohol is [-0.146, -0.015].
- The p-value for Alcohol is 0.017, indicating that it is statistically significant in predicting life expectancy
- The coefficient for Alcohol (a numerical variable) is -0.0801
    - One unit increase in Alcohol results in a 0.0801 unit decrease in life expectancy, all other variables held constant
- The coefficient for Status_Developing (a categorical variable) is -2.6234
    - The life expectancy for developing countries will be 2.6234 years lesser than a developed country, all other variables held constant



```
                            OLS Regression Results
==============================================================================
Dep. Variable:        Life expectancy   R-squared:                       0.843
Model:                            OLS   Adj. R-squared:                  0.842
Method:                 Least Squares   F-statistic:                     575.6
Date:                Fri, 01 Oct 2021   Prob (F-statistic):               0.00
Time:                        13:13:06   Log-Likelihood:                -5636.5
No. Observations:                2049   AIC:                         1.131e+04
Df Residuals:                    2029   BIC:                         1.143e+04
Df Model:                          19
Covariance Type:            nonrobust
============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                      -23.1832     39.635     -0.585      0.559    -100.913      54.547
Year                         0.0395      0.020      1.994      0.046       0.001       0.078
Adult Mortality             -0.0162      0.001    -17.819      0.000      -0.018      -0.014
Alcohol                     -0.0801      0.033     -2.395      0.017      -0.146      -0.015
Percentage expenditure       0.0003   4.96e-05      6.128      0.000       0.000       0.000
Hepatitis B                 -0.0163      0.004     -3.821      0.000      -0.025      -0.008
BMI                          0.0346      0.006      5.860      0.000       0.023       0.046
Status_Developing           -2.6234      0.353     -7.442      0.000      -3.315      -1.932
Continent_Asia               4.7406      0.281     16.862      0.000       4.189       5.292
Continent_Europe             4.3902      0.411     10.694      0.000       3.585       5.195
Continent_North America      6.2753      0.360     17.417      0.000       5.569       6.982
Continent_Oceania            2.7757      0.456      6.089      0.000       1.882       3.670
Continent_South America      4.4261      0.440     10.062      0.000       3.563       5.289
==============================================================================
Omnibus:                       80.001   Durbin-Watson:                   2.014
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              214.140
Skew:                          -0.138   Prob(JB):                     3.16e-47
Kurtosis:                       4.559   Cond. No.                     1.14e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.14e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

ReCell - Business Context and Objectives

- Buying and selling used phones and tablets used to be something that happened on a handful of online marketplace sites. But the used and refurbished device market has grown considerably over the past decade, and a new IDC (International Data Corporation) forecast predicts that the used phone market would be worth $52.7bn by 2023 with a compound annual growth rate (CAGR) of 13.6% from 2018 to 2023. This growth can be attributed to an uptick in demand for used phones and tablets that offer considerable savings compared with new models.

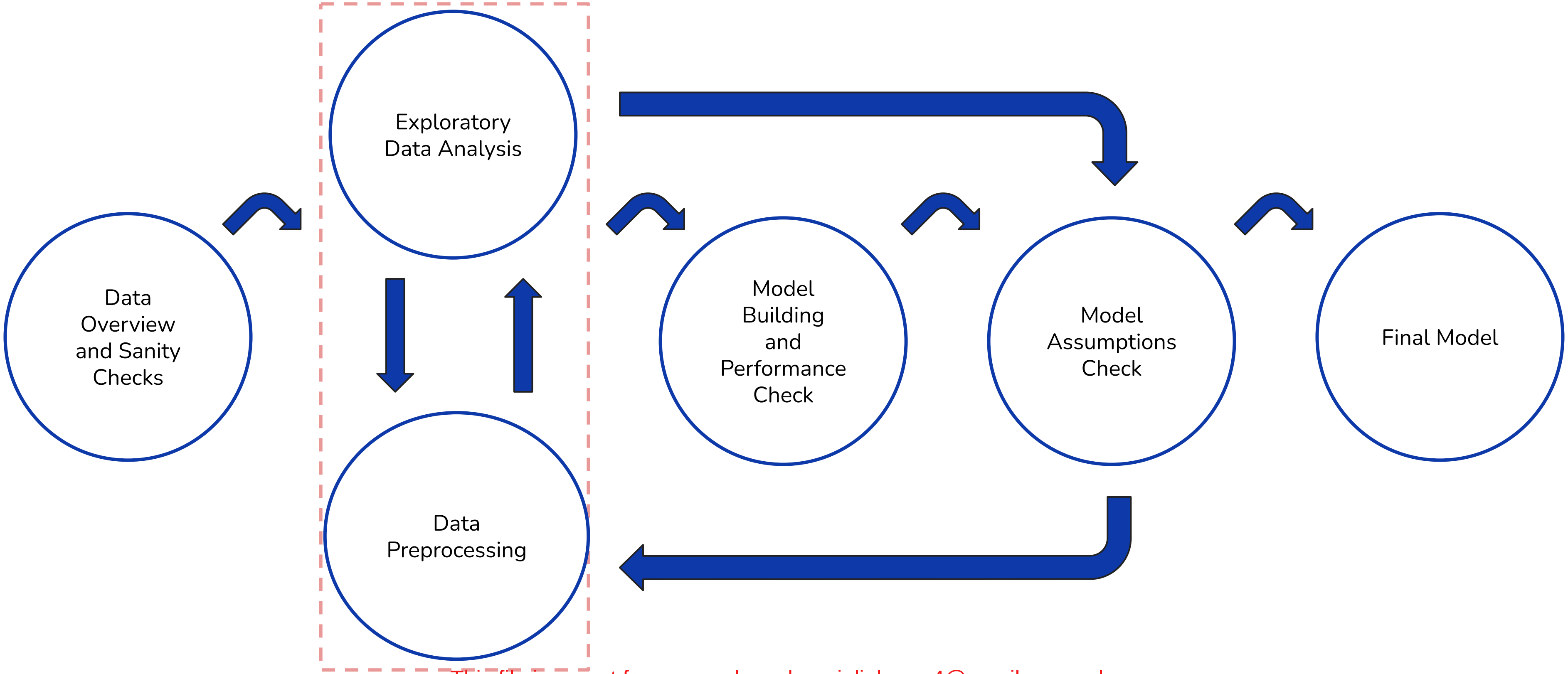
- Refurbished and used devices continue to provide cost-effective alternatives to both consumers and businesses that are looking to save money when purchasing one. There are plenty of other benefits associated with the used device market. Used and refurbished devices can be sold with warranties and can also be insured with proof of purchase. Third-party vendors/platforms, such as Verizon, Amazon, etc., provide attractive offers to customers for refurbished devices. Maximizing the longevity of devices through second-hand trade also reduces their environmental impact and helps in recycling and reducing waste. The impact of the COVID-19 outbreak may further boost this segment as consumers cut back on discretionary spending and buy phones and tablets only for immediate needs.

- The rising potential of this comparatively under-the-radar market fuels the need for an ML-based solution to develop a dynamic pricing strategy for used and refurbished devices. ReCell, a startup aiming to tap the potential in this market, has hired you as a data scientist. They want you to analyze the data provided and build a linear regression model to predict the price of a used phone/tablet and identify factors that significantly influence it.

ReCell - Dataset

- The data contains the different attributes of used/refurbished phones and tablets. The data was collected in the year 2021. The detailed data dictionary is given below.
    - brand_name: Name of manufacturing brand
    - os: OS on which the device runs
    - screen_size: Size of the screen in cm
    - 4g: Whether 4G is available or not
    - 5g: Whether 5G is available or not
    - main_camera_mp: Resolution of the rear camera in megapixels
    - selfie_camera_mp: Resolution of the front camera in megapixels
    - int_memory: Amount of internal memory (ROM) in GB
    - ram: Amount of RAM in GB
    - battery: Energy capacity of the device battery in mAh
    - weight: Weight of the device in grams
    - release_year: Year when the device model was released
    - days_used: Number of days the used/refurbished device has been used
    - normalized_new_price: Normalized price of a new device of the same model in euros
    - normalized_used_price: Normalized price of the used/refurbished device in euros

ReCell - Solution Approach



Data Overview and Sanity Checks

Exploratory Data Analysis

Data Preprocessing

Model Building and Performance Check

Model Assumptions Check

Final Model

ReCell - Q/A

ReCell - Project FAQs

Since we have missing values in the dataset, what is the best way to handle or treat those missing values?

The strategy to deal with missing values varies with the problem at hand, the data provided, and other factors too. Some of the common strategies are listed below.

- Drop the missing values
- Impute the missing values
  - Using central tendency measures (mean, median, mode) of a column
    - With mean: Missing values are imputed with the mean of the column. Preferred for continuous data with no outliers
    - With median: Missing values are imputed with the median of the column. Preferred for continuous data with outliers
    - With mode: Missing values are imputed with the mode of the column. Preferred for categorical data
  - Using central tendency measures (mean, median, mode) of a column grouped by categories of a categorical column: Preferred for cases where the data under similar categories of a categorical column are likely to have similar properties

ReCell - Project FAQs

What one should do if the VIF is high (> 5) for some dummies and not for the other dummies of a categorical variable?

The VIF values for dummy variables can be ignored.

If, however, the VIF value is inf or NaN, then one should check if one of the dummy variables was dropped during one-hot encoding. If the VIF value is still inf or NaN, a different dummy variable than the one dropped by using drop_first=True should be dropped and VIF values should be checked again.

- For example, if a categorical variable 'Season' has four levels 'Spring', 'Summer', 'Fall' and 'Winter', and using drop_first=True drops the dummy variable for 'Fall', then one can keep the dummy variable for 'Fall' and drop the dummy variable for 'Summer', and then check the VIF values.

ReCell - Project FAQs

What one should do if the p-values are high (> 0.05) for some dummies and not for the other dummies of a categorical variable?

The dummy variables with p-value > 0.05 should be dropped one by one until there are no such variables. After removing each high p-value variable, the regression should be run again, and the p-values of all the variables should be checked.

If all the dummy variables of a categorical column have a p-value > 0.05, then all the dummy variables for that column can be dropped at once.

ReCell - Project FAQs

X=sm.add_constant(X) is not working for my project as a new column is not created. Why is this happening? How can this be resolved?

add_constant() does not add a constant column to the data if a constant column already exists in it. Please check if the data has a constant column before using add_constant().

As none of the independent variables should ideally be constants as they have variability in them initially, the step where the variable(s) became constant has to be identified. The outlier treatment step is a good place to start.

greatlearning

Power Ahead

Happy Learning !