Project 6 Report

Jake Epperson | Yagnashree Velanki

Introduction

In this project, we will apply multiple linear regression to a hydrologic dataset to predict the baseflow. The dataset contains information about various parameters such as date, segment ID, spatial location, evapotranspiration, precipitation, irrigation pumping, and observed baseflow. The target variable is observed baseflow, and we want to use linear regression to predict what the baseflow will be.

Repository: https://github.com/jakepper/CS5830-Group15-Project6.git
Presentation: ▭ CS5830-Group15-Project6-Presentation

Dataset

For this project we have used the hydraulic dataset that contains information about various factors that affect the baseflow of a river, including Segment_id, x, y which are spatial locations and evapotranspiration, precipitation, irrigation pumping, which will affect the ‘Observed’ baseflow

Preprocessing:

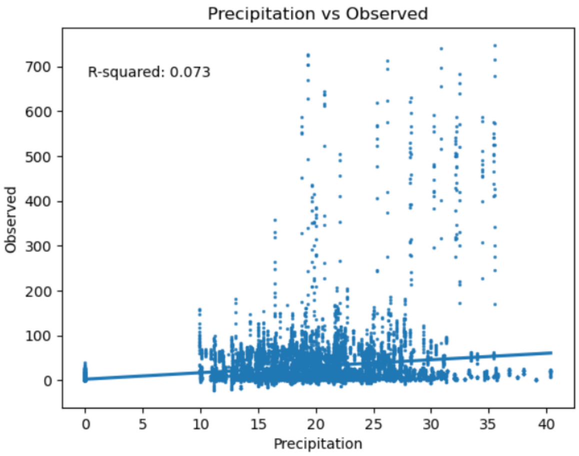
As part of pre-processing, We have converted the date column to date format by subtracting 693963 from each value and then converted them into DateTime format with the date starting from 1900-1-1. Additionally, we created a new column to extract the year from the date for further analysis. After we process the Date column we get the range of years present in the Date column will be 1939 - 2000.

Analysis Technique

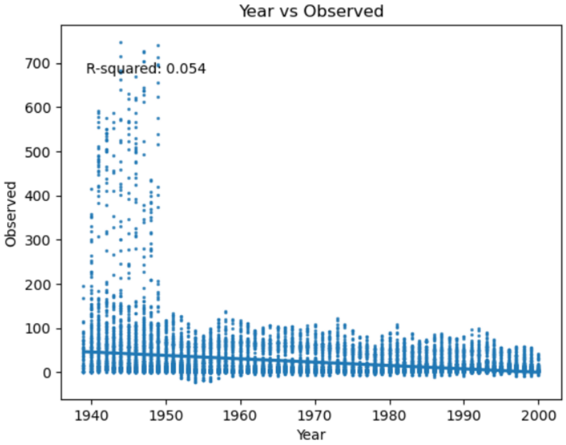
We performed several linear and multiple linear regression analyses using various types of variables from the dataset to evaluate the ability of independent variables to accurately predict the target variable. The aim of these analyses was to determine the strength of the relationship between the predictors and the target variable. Additionally, we plotted a line reg plot to visualize how the "Observed" value changes over the years.

Results

Initially, we analyzed the relationship between the target variable 'Observed' and several independent variables, including 'x', 'y', 'Evapotranspiration', 'Precipitation', 'Irrigation_pumping', and 'Year', using linear regression plots. We generated separate regression plots for each independent variable to visualize their association with the target variable. Additionally, we calculated the R-squared value and reported the mean and standard deviation of the Pearson correlation coefficient for each independent variable. Our analysis revealed a moderate positive correlation between 'Precipitation' and 'Observed' variables, with a R-squared value of 0.073 Pearson correlation coefficient of 0.270 and a significant p-value of less than 0.05. This implies that when the precipitation level increases, the observed variable is likely to increase as well. The average value of precipitation is 14.92, with a standard deviation of 10.51 as shown in fig(1) and a negative correlation was seen between Year and Observed variables with a Pearson correlation coefficient of -0.23 and significant p-value of 9.309. The average value of 1968.58, with a standard deviation of 17.21 and R-squared value of 0.054 as shown in fig(2). Remaining visualizations of other variables are in the git due to the space constraint we are not including every graph in the report.

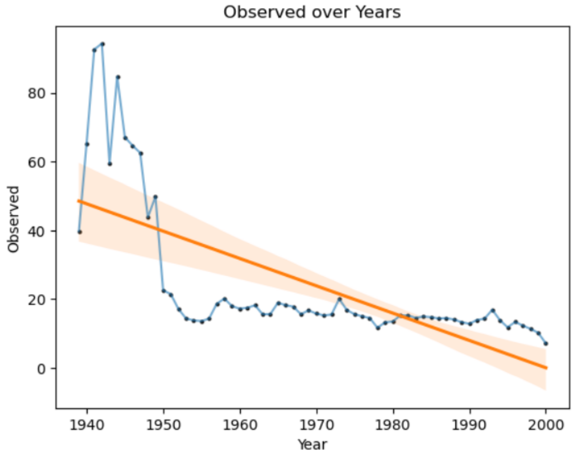


fig(1)

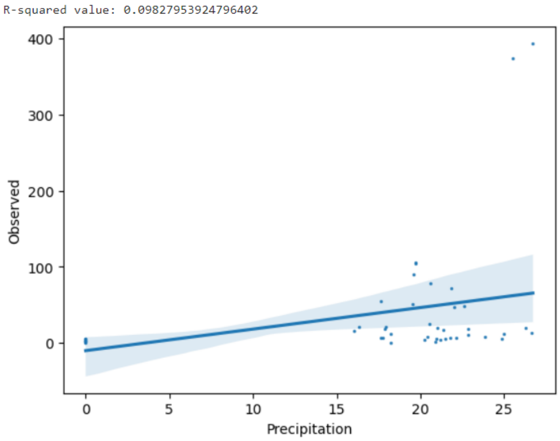


fig(2)

We generated a line plot with a regression line to analyze the changes in the data over time (represented by the Year variable), with a focus on the Observed variable that represents the observed baseflow of water. Our assumption was that pollution has led to a decrease in base flow over the years. As expected, our analysis in fig(3) shows a decreasing trend in base flow over time.



fig(3)

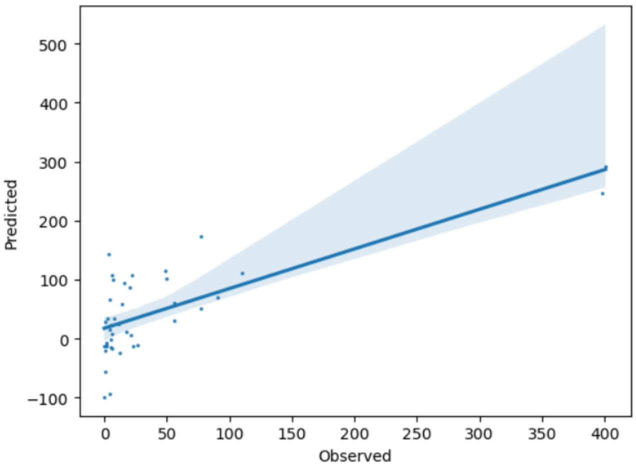


```
R-squared value: 0.09827953924796402
```

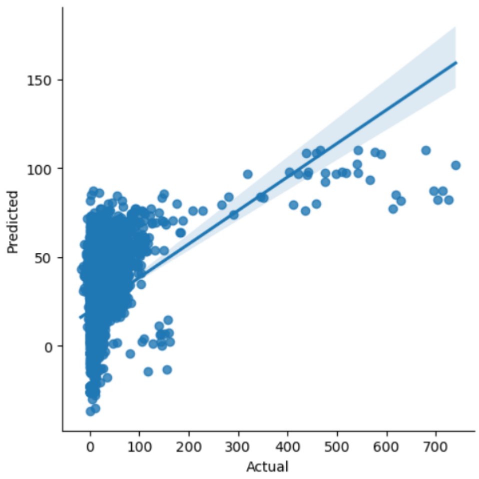
fig(4)

We used a linear regression plot to analyze the relationship between precipitation and the observed base flow, grouping the data by their respective x and y spatial locations and got the R-squared value as 0.0502 as shown in fig(4). We also conducted a multiple linear regression analysis on a dataset where we grouped the segment id and considered the independent variables such as 'Segment_id','Evapotranspiration', 'Precipitation', 'Irrigation_pumping', and dependent variable as 'Observed'. However, the results were not as expected as we obtained a negative R-squared value of -6.523649895444377, which indicates that the model does not fit the data well. Moreover, we got a high mean squared error (MSE) value of 4759.478659457904 and root mean squared error (RMSE) value of 68.98897491235758. The coefficients for the independent variables were 'Segment_id': 0.9184690903972462, 'Evapotranspiration':

-39.22994041540714, 'Precipitation': 9.89859985649145, and 'Irrigation_pumping': 4520.643317056029. We have created a column in dataframe to get the predicted values and plotted a regression plot for the predicted values and 'Observed' which is a target column fig(5). Similar to the mentioned multiple regression we have performed another Multiple regression plot using the predictor variables as 'Segment_id', 'x', 'y', 'Evapotranspiration', 'Precipitation', 'Irrigation_pumping' and target variable as 'Observed' and plotted a regression plot for actual and predicted values which is shown in fig(6).The scatter plot displays the relationship between the actual and predicted values of a dependent variable for a testing dataset. The x-axis shows the actual values, the y-axis shows the predicted values, and each point on the plot represents an observation in the dataset with its actual and predicted values.

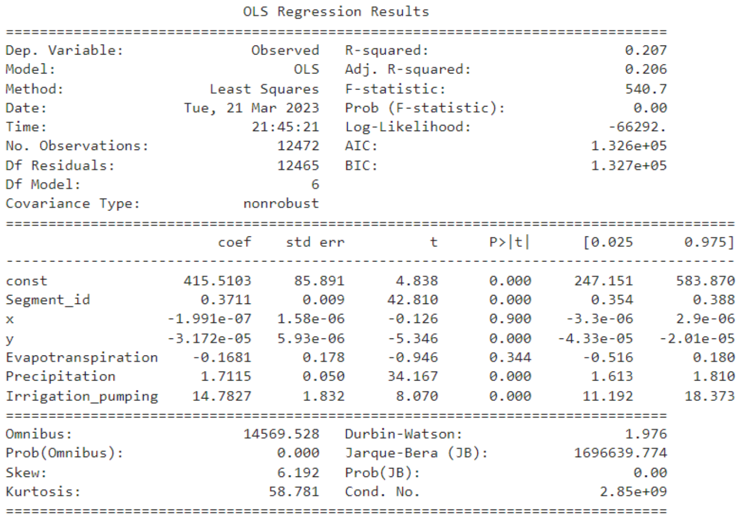


fig(5)



fig(6)

We have performed an OLS method for multi-linear regression and extracted the summary of the regression results which includes various statistics such as R-squared, the coefficient estimates, standard errors, t-values, and p-values for each predictor in the model. From which we received the R-squared value as 0.207, which means that the model explains 20.7% of the variance in the observed data. We also obtained the other data such as coef, std err, t, P>|t|, [0.025, 0.975] etc., which can be seen in fig(7).



```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Observed   R-squared:                       0.207
Model:                            OLS   Adj. R-squared:                  0.206
Method:                 Least Squares   F-statistic:                     540.7
Date:                Tue, 21 Mar 2023   Prob (F-statistic):               0.00
Time:                        21:45:21   Log-Likelihood:                -66292.
No. Observations:               12472   AIC:                         1.326e+05
Df Residuals:                   12465   BIC:                         1.327e+05
Df Model:                           6
Covariance Type:            nonrobust
======================================================================================
                         coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------
const                415.5103     85.891      4.838      0.000     247.151     583.870
Segment_id             0.3711      0.009     42.810      0.000       0.354       0.388
x                  -1.991e-07   1.58e-06     -0.126      0.900    -3.3e-06     2.9e-06
y                  -3.172e-05   5.93e-06     -5.346      0.000   -4.33e-05   -2.01e-05
Evapotranspiration    -0.1681      0.178     -0.946      0.344      -0.516       0.180
Precipitation          1.7115      0.050     34.167      0.000       1.613       1.810
Irrigation_pumping    14.7827      1.832      8.070      0.000      11.192      18.373
==============================================================================
Omnibus:                    14569.528   Durbin-Watson:                   1.976
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          1696639.774
Skew:                           6.192   Prob(JB):                         0.00
Kurtosis:                      58.781   Cond. No.                     2.85e+09
==============================================================================
```

fig(7)

Technical

We have imported several python libraries to do the analysis for the data. We have performed our data preprocessing using Pandas. As a part of pre processing we need to convert the Date column into proper date format. We were able to accomplish this by adding a time delta to the date time value of January 1st, 1900. Our analysis primarily focused on exploring the relationships between different variables in the dataset. We used Pearson correlation tests, and calculated R-squared values to measure the strength and direction of the linear association between two variables. In addition, we created scatter plots to visualize the relationship between the target variable (Observed) and other variables in the dataset, as well as line plots to observe how the target variable changed over time. Furthermore, we developed a linear regression model to predict the base flow using scikit-learn's LinearRegression platform. This model can help us understand how changes in the predictor variables affect the target variable, and can be used to make predictions for new data points.