Project 6 - Linear Regression Chris Brown and Anny Chen

Presentation Slides

Introduction

In this analysis, we investigated how Linear Regression performs in predicting the observed baseflow of a river network in Nebraska from 2001 to 2004. Linear Regressions were chosen as they are simple predictive models that perform well in predicting continuous variables. Using a simple method allows us to spend time with feature exploration instead of fine-tuning hyperparameters. Our findings provide Natural Resource Specialists with the leading causes of increases and decreases in river baseflow.

Dataset

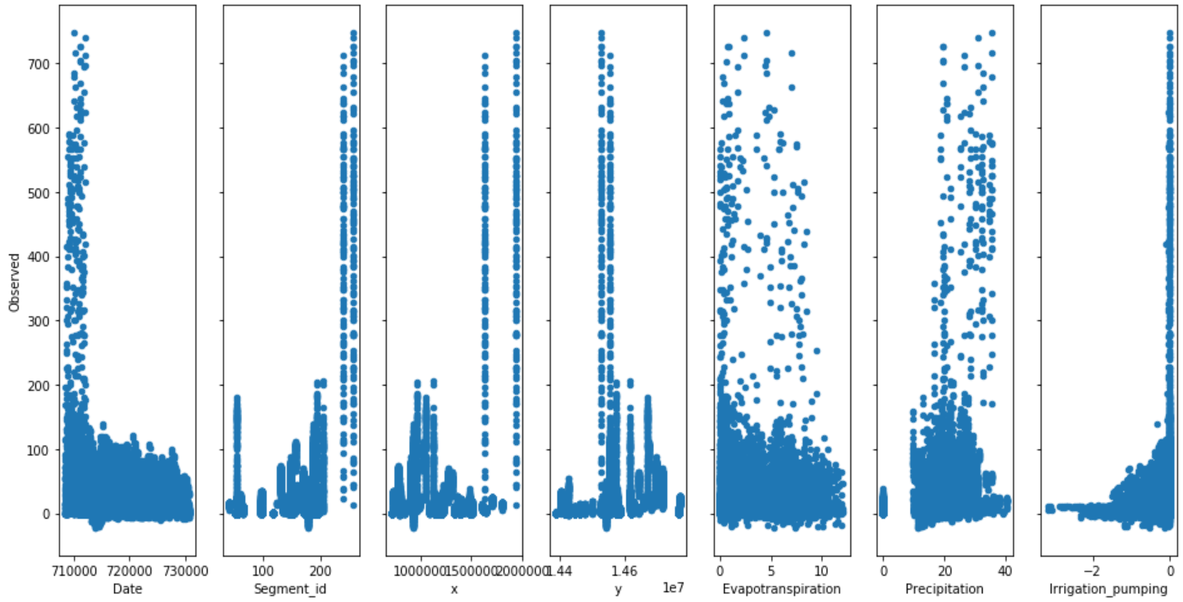
McDonald Morrissy published a “Review of RRCA Model For the Period 2001 to 2004” for the Nebraska Department of Natural Resources in 2006. The Republican River Compact Administration (RRCA) ground-water model was developed jointly by several States to estimate streamflow depletions and accretions caused by pumping and recharge from imported water. All results from the study were checked and calibrated against other independently published models. There were118 measured wells in Nebraska’s water network, but only 42 reported segments as they are the only ones that matched other sources’ models. The study also reported that the model has imprecisions in some areas as the calculated water levels are consistently too high. The study provides a comprehensive dataset for the 42 reported segments from 2001 to 2004 and includes the attributes: Date, Segment ID, X & Y coordinates, Evapotranspiration, Precipitation, Irrigation Pumping, and the Observed Baseflow. The baseflow is the groundwater elevation minus ground elevation. Thus, a negative baseflow indicates the groundwater is below ground level and a positive baseflow indicates the groundwater is above ground level. There are no missing values or corrupted data values in the dataset.

In order to avoid overfitting we began our analysis techniques by randomly partitioned the data into training (80%) and testing (20%) datasets. We further split the data into training (60%) and validation (20%) datasets. We used the validation during feature selection and testing for final evaluation. No further data cleanup was necessary. To analyze individual river segments we grouped the data by segment ids using Panda’s “groupby” and “apply” functions.

Analysis Technique

To analyze the hydrologic dataset -- RRCA Model for the Period 2001 to 2004, we used a linear regression model to fit the data. To get the best model, we selected features using mean-square-error (MSE), p-values, and R-squared values. To determine the goodness of our model, we use MSE as the standard to judge our model as our target “Observed” is continuous.

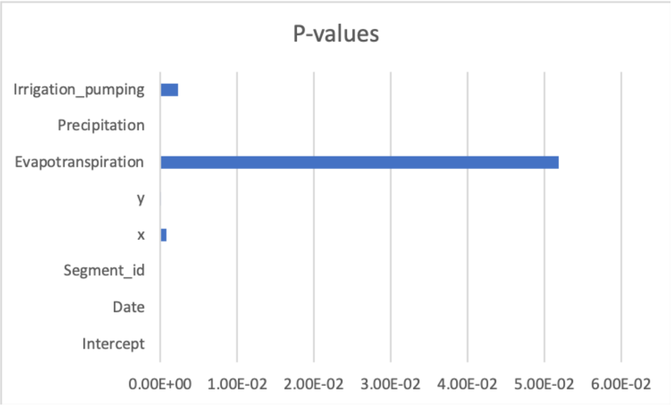
As preliminary work to determine feature importance we first describe all the features by finding their std, avg, max, and min. We then visualize each **x** feature and **y**, for a better understanding of the underlying data distribution. In the following plot, we find a large value distribution in the features. From this alone, we believe that there are not enough strong correlations that will allow us to make a good Linear Regression model for the entire river network. In other words, considering the hydrologic dataset of each segment separately is necessary.



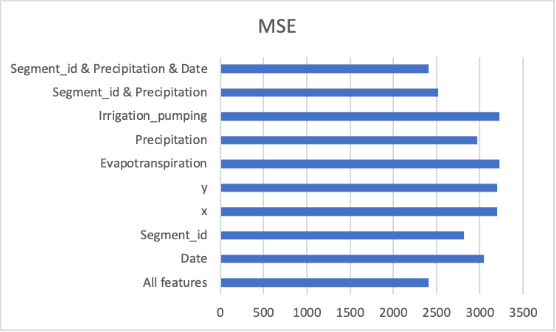
After doing some preliminary data exploration, we built a Linear Regression model for both the entire river system and for individual river segments. To select features for the whole river system, we calculate MSE, p-values, and R-squared values. MSE and R-squared values represent whether the model fits the data or not. Once we got features we think predict the observed river baseflow the best, we used them to build a final model and used our test dataset to evaluate the model. P-values represent whether we or confident our model is a 'lucky' model or not. For the individual river segments, we grouped the data by their river segment and performed a similar analysis with the clusters of segments.

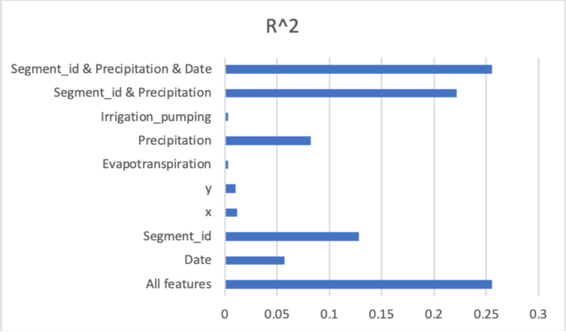
Results

We found the P-value of Evapotranspiration is bigger than 0.05, while others are smaller. Therefore all features except Evapotranspiration are important.

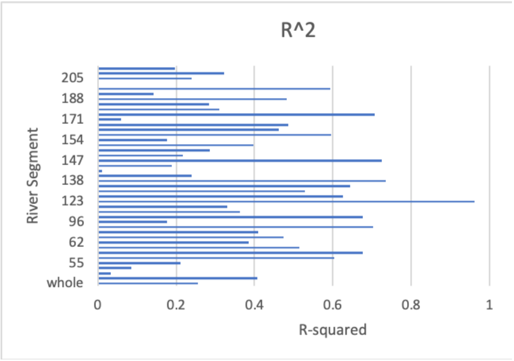


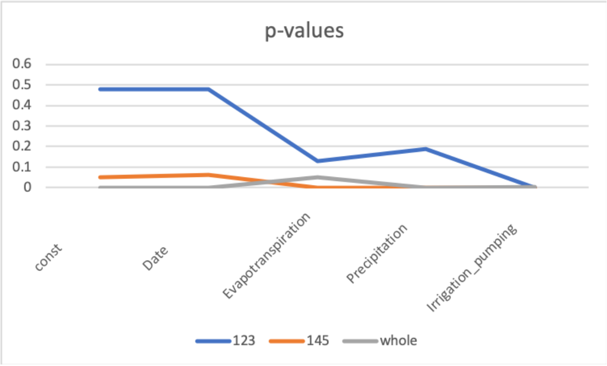
From the R-squared picture, we figure out that the number of features influence R-squared. But if we only consider one feature, the Segment_id, Precipitation, and Date perform best. But MSE will not be influenced by the number of features. One interesting thing is MSE of Segment_id & Precipitation & Date is the same as MSE of all features. From MSE values, we get the conclusion that for the whole river, Segment_id & Precipitation & Date are the most predictive explanatory variables.



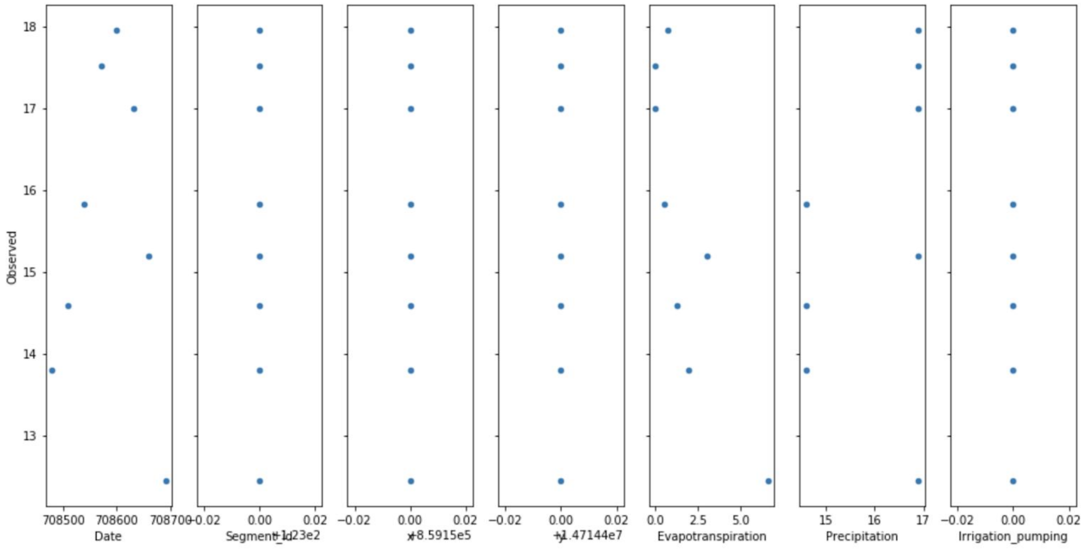


For each river segment, we calculate R-squared and p-values. Here are all the R-squared for each segment. Since each river segment has several p-values, we just pick two segments(123 and 145) for visualization.

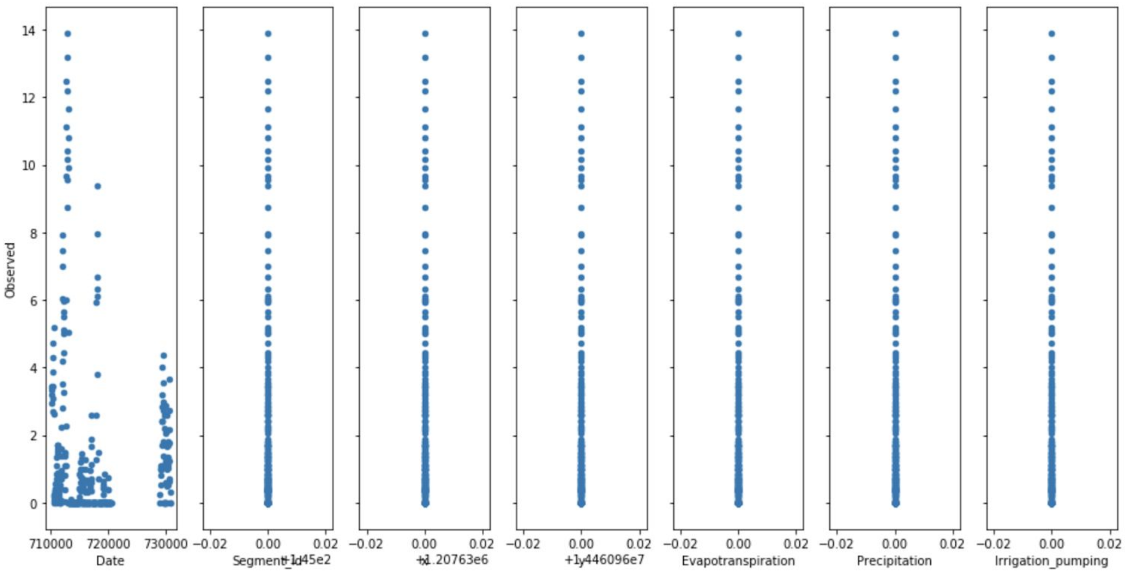




Segment 123 has the highest R-squared value of 0.9604597, and p-values are all bigger than 0.05 or NaN. Segment 145 has a low R-squared value of 0.0109074, and p-values are all bigger than 0.05 or NaN. For deeper analyzation, we visualize 123 and 145 of all features.



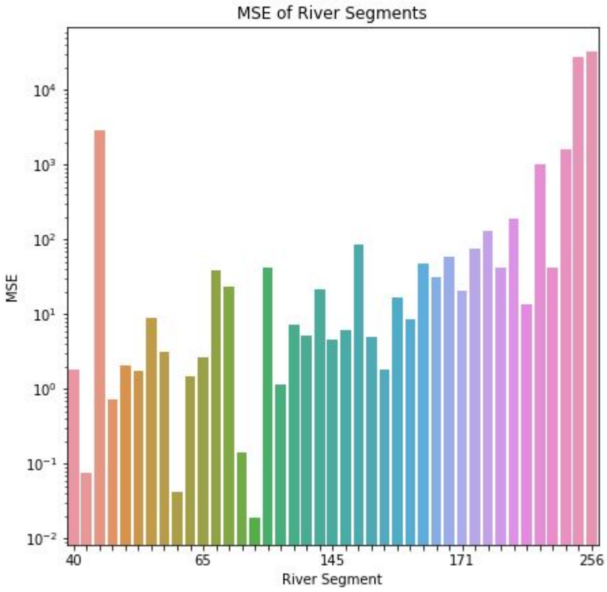
Segment 123



Segment 145

From picture segment 123, we figure out data is too little but built a linear distribution. Too little data may result in high p-values, and if feature values are the same, p-values will get NaN. And linear distribution results in high R-squared. For picture segment 145, although there is enough data, most features are zero, which results in NaN of p-values and low R-squared. In general, the R-squared of each segment is different from the whole river, since the distribution of each segment data is linear (R-squared is better) or most belongs to zero (R-squared is worse) or not linear(R-squared is similar). Besides, p-values of each segmentation are different from the whole, and the number of data may be the reason(if data is little, p-values of features might be much higher).

Visualizing the mean-squared-errors of every model for the river segments shows that some segments cannot be modeled well while others are modeled very well. On average, we have an MSE of about 10 when predicting the river baseflow.



From our analysis, Natural Resource Specialists can see that the best indicators of river baseflow are the current date and evapotranspiration at the time. They can also see that the baseflow is highly dependent on the segment being monitored.