Gabriel Daiess

Multiple Linear Regression Analysis of Stream Runoff Predicted by Precipitation

5/9/2023

Submitted as Final Paper for STAT 632, Spring 2023

Dr. Joshua Kerr

California State University, East Bay

Multiple linear regression is a widely employed statistical method for forecasting environmental phenomena. Multiple linear regression is particularly well-suited for environmental data, especially time-series, in which environmental data is recorded with lags that follow many of Earth's temporal processes. One such process is the hydrological cycle, in which seasonal snowfall contributes to streamflow, or the water movement in a river, after a progression of melting and laminar flow across surfaces as runoff, eventually depositing as water in rivers.

This study is a multiple linear regression analysis of hydrologic data that aims to forecast stream runoff as a response using snowfall data at 6 precipitation stations as explanatory variables. The focus of this study is to answer the following interests: (1) the relationship between stream runoff and precipitation, (2) the type (linear or otherwise) of relationship between response and predictor, (3) the goodness of fit and statistical significance of different multiple linear regression models, in the context of the data and accompanying research questions

The data for this study was compiled by the UCLA statistics department pooling from publicly available water data in California. It is worth noting the process by which it was compiled and the sources it was compiled from remain unclear, and as a result, the data is limited in terms of further studies as well as reproducibility. The dataset itself is from "Applied Linear Regression" (4th ed.) by Sanford Weisburg of the University of Minnesota (Weisburg, 2013). The dataset was installed directly in RStudio via CRAN using the "ALR4" package, where it is titled "water."

The dataset dimension is 43 rows by 8 columns. The data is a time series with each observation from a unique year. The dataset begins from 1948 and ends in 1990 (yielding the 43 years of observations). The first column "Year" was not used as a variable in the multiple linear regression model, though it was useful for exploratory data analysis and visualizations. The other columns (as coded in the dataset) are 6 precipitation stations: "APMAM", "APSAB", "APSLAKE", "OPBPC," "OPRC," and "OPSLAKE." Figure 1 is a visualization of precipitation at each station over time. The last column in the dataset is "BSAAM" which according to the author is "stream runoff near Bishop, CA, in acre-feet."

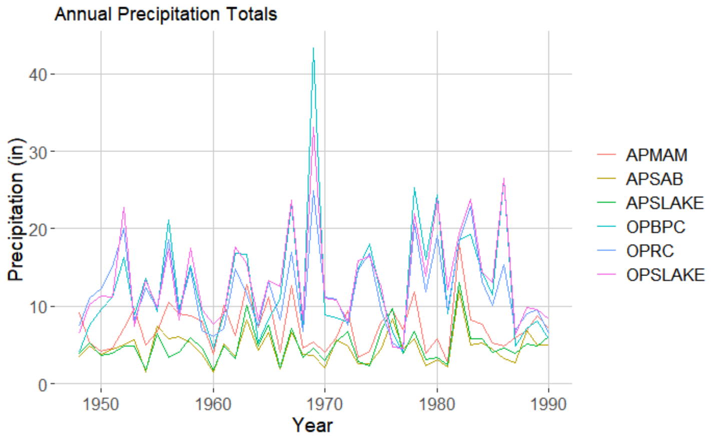


Figure 1: Time Series Precipitation by Station

The response variable chosen in this study is stream runoff in acre-feet, collected each year, at the "BSAAM" site. The predictor variables for the full-model are precipitation, specifically snow in inches, recorded each year at each station "APMAM", "APSAB", "APSLAKE", "OPBPC," "OPRC," and "OPSLAKE."

The values of precipitation at each station are relatively similar to one another, and notably similar to one another if viewing the data at stations APMAM, APSAB and APSLAKE as group 1 and OPBPC, OPRC and OPSLAKE as group 2 (see figure 2). The summary statistics of the data is not particularly noteworthy, as the bulk of the observations follow a normal distribution. There are a few outliers in the data, none of which are also leverage points, and as a result did not warrant any further investigation. The data for the response variable is in the order of tens of thousands while the predictor variables are in terms of single digits.

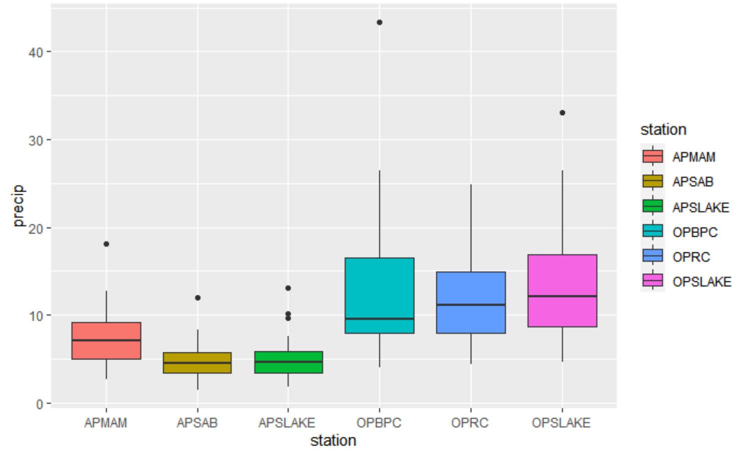
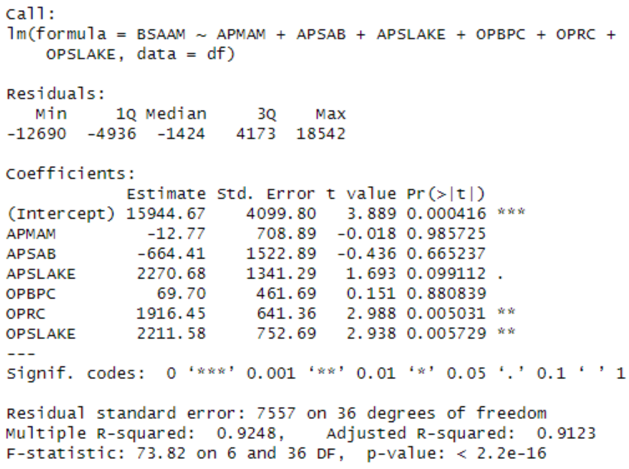


Figure 2: Boxplot of Precipitation (in) by Station

After the exploratory data analysis, a full multiple linear regression model employing all of the precipitation stations as predictors for stream runoff is tested. Expressed mathematically, the full model is:

$$\widehat{StreamRunoff} = \beta_0+\beta_1(APMAM)+\beta_2(APSAB)+\beta_3(APSLAKE)+ \beta_4(OPBPC) + \beta_5(OPRC) + \beta_6(OPSLAKE) + \epsilon$$

The summary of the full model (shown in Figure 3) displays some very obvious issues that would render it ineffective at answering the research questions in an accurate and reliable manner. First, the estimated values of precipitation at stations APMAM and APSAB are negative. In the context of the study this does not make sense, as it would suggest snowmelt leaving the Owens River Valley region would result in an increase in stream runoff at the BSAAM site near Bishop. From the hydrologic cycle in the region we know that mathematically, the relationship between the predictor and response variables is positively linear. Second, aside from the last two precipitation stations (OPRC and OPSLAKE), the other four predictors individual t-tests yield insignificant p-values (> 0.05) in the presence of one another. This (falsely) suggest that 4 of the 6 predictors are not contributing to the model in a meaningful way. Of great importance, despite the issues with non-significant predictors and problematic negative estimates, the coefficient of determination for the model is 0.9123, meaning that about 91.23% of the variability in the response can be explained by the predictors. More explicitly, the model does indeed explain the linear relationship between stream runoff and precipitation, though given the aforementioned issues is not appropriate as a final model.



```
Call:
lm(formula = BSAAM ~ APMAM + APSAB + APSLAKE + OPBPC + OPRC +
    OPSLAKE, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-12690  -4936  -1424   4173  18542 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 15944.67    4099.80   3.889 0.000416 ***
APMAM         -12.77     708.89  -0.018 0.985725    
APSAB        -664.41    1522.89  -0.436 0.665237    
APSLAKE      2270.68    1341.29   1.693 0.099112 .  
OPBPC          69.70     461.69   0.151 0.880839    
OPRC         1916.45     641.36   2.988 0.005031 ** 
OPSLAKE      2211.58     752.69   2.938 0.005729 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7557 on 36 degrees of freedom
Multiple R-squared:  0.9248,	Adjusted R-squared:  0.9123 
F-statistic: 73.82 on 6 and 36 DF,  p-value: < 2.2e-16
```

Figure 3: Summary of Full Model

Following the creation of the first model and my interpretation of it, the following assumptions of multiple linear regression were assessed and evaluated: (1) linearity between predictors and response variable (2) normally distributed residuals, (3) homoscedasticity/constant variance of residuals, (4) independence of observations and no multicollinearity. These assumptions were assessed both graphically and mathematically. The diagnostic plots (“Normal QQ” plot and “Residuals vs. Fitted” values plot) are shown below in Figure 4.

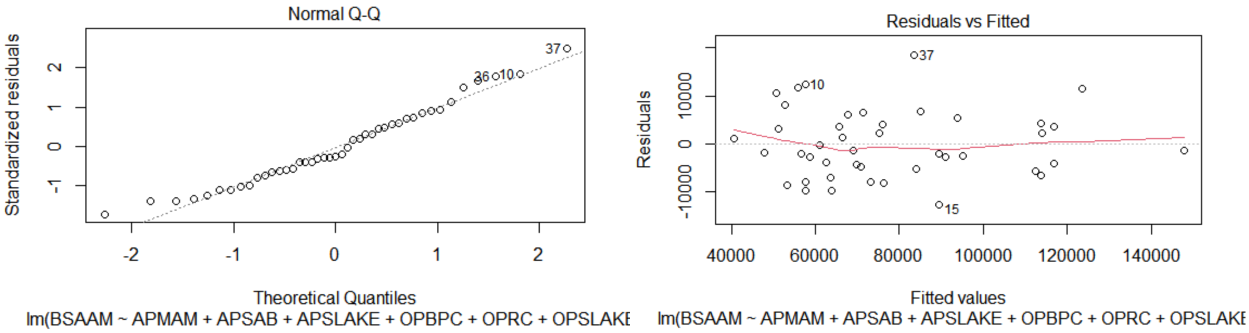
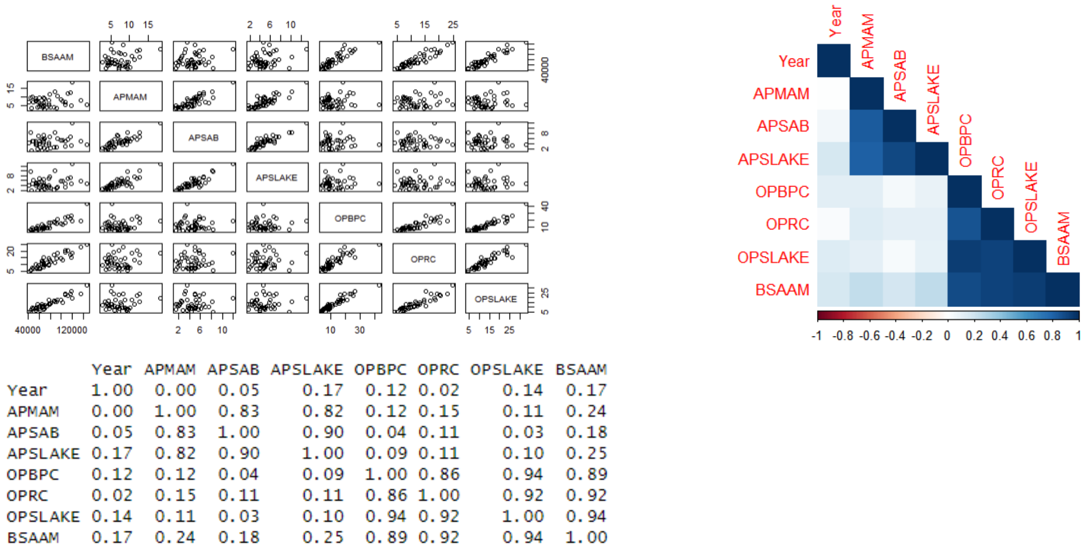


Figure 4: Diagnostic Plots for Full Model

I conclude that the assumptions of normality of the residuals are well-satisfied by visually assessing the points of the model residual values nearly perfectly aligning with the theoretical distribution of perfectly normal residuals on the normal QQ plot. I also confirm normality of the residuals using the Shapiro-Wilk normality test on the initial model (p-value = 0.43 > alpha of 0.05). I conclude that homoscedasticity is satisfied using the studentized Breusch-Pagan test, for which normality is a prerequisite (p-value = 0.948 > alpha of 0.05). I confirm constant variance and homoscedasticity in the residuals of the model by confirming no discernible pattern in the residuals versus fitteds plot; noise hovering around zero within the bounds (-2:2). As a result of my conclusions above, I do not consider transformations or interactions.

The other requirements of multiple linear regression, independence and no multicollinearity, however cannot be assessed by the diagnostic plots. Instead, a scatterplot matrix, correlation matrix, correlation graph, and the variable inflation factor (VIF) were used to assess multicollinearity and independence. It is also worth noting that coefficients of a multiple linear regression model having estimates with values that are negative when we expect a positive, coupled with a high coefficient of determination yet insignificant individual t-test statistics for the predictors, is often an indication of multicollinearity. Given the characteristics of the full model, I highly anticipate that the variables are not independent, and that there is a potentially linear relationship between predictors themselves. The stations are close to one another and thus are collecting similar amounts of snowfall over the same periods of time. This correlation can be visualized by the parallel lines of the time series in Figure 1.

From the scatterplot, correlation matrix and correlation plot (Figure 5) I visually conclude that the subset of stations APMAM, APSAB, APSLAKE are all highly correlated and likely not independent from one another. Similarly, I continue that the subset of stations OPBPC, OPRC and OPSLAKE are highly correlated and likely not independent. Also, the latter subset of predictors are highly correlated to the response variable. From the VIF of each predictor, I confirm my conclusions, as all variable inflation factors are above 5, aside from station APMAM which had a VIF of 3.55 (see R code section VIF). I reject the full model altogether as the violations of the prerequisites for a multiple linear regression are too strongly violated by issues of multicollinearity amongst predictor variables and lack of independence in observations.



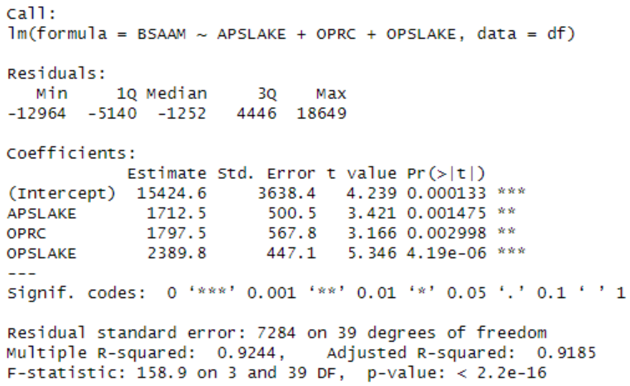
```
        Year APMAM APSAB APSLAKE OPBPC OPRC OPSLAKE BSAAM
Year    1.00  0.00  0.05    0.17  0.12 0.02    0.14  0.17
APMAM   0.00  1.00  0.83    0.82  0.12 0.15    0.11  0.24
APSAB   0.05  0.83  1.00    0.90  0.04 0.11    0.03  0.18
APSLAKE 0.17  0.82  0.90    1.00  0.09 0.11    0.10  0.25
OPBPC   0.12  0.12  0.04    0.09  1.00 0.86    0.94  0.89
OPRC    0.02  0.15  0.11    0.11  0.86 1.00    0.92  0.92
OPSLAKE 0.14  0.11  0.03    0.10  0.94 0.92    1.00  0.94
BSAAM   0.17  0.24  0.18    0.25  0.89 0.92    0.94  1.00
```

Figure 5: Scatterplot, Correlation Plot and Correlation Matrix for Model 1

Before proceeding to build a second linear model for stream runoff and precipitation, I tested the full model for independence of observations by testing for autocorrelated residuals. Autocorrelation amongst residuals is a form of dependence in time series, thus its presence would be a violation of the prerequisites for multiple linear regression. Using the Durbin-Watson test, I find that the residuals of the full model are autocorrelated. The null hypothesis of the Durbin-Watson test in R is that there is no first-order autocorrelation in the residuals of the regression model, i.e., the residuals are independent. The alternative hypothesis is that there is first-order autocorrelation in the residuals. From the p-value of the test (0.026), I reject the null hypothesis in favor of the alternative and conclude the residuals of the first model have autocorrelation and, more importantly, are not independent.

For the final portion of this study I utilize a stepwise variable selection to create a second, reduced model. For the stepwise variable selection I defaulted to the Akaike Information Criterion (AIC) as the criteria for model selection. The aim of this process is not solely variable selection but an attempt to strike a balance between model complexity and fit of the data. The final multiple linear regression model with the lowest AIC, I selected against other options is:

$$\widehat{StreamRunoff} = 15424.6{+}1712.5APSLAKE{+}1797.5OPRC{+}2389.8OPSLAKE$$



```
Call:
lm(formula = BSAAM ~ APSLAKE + OPRC + OPSLAKE, data = df)

Residuals:
   Min     1Q Median     3Q    Max 
-12964  -5140  -1252   4446  18649 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  15424.6     3638.4   4.239 0.000133 ***
APSLAKE       1712.5      500.5   3.421 0.001475 ** 
OPRC          1797.5      567.8   3.166 0.002998 ** 
OPSLAKE       2389.8      447.1   5.346 4.19e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7284 on 39 degrees of freedom
Multiple R-squared:  0.9244,	Adjusted R-squared:  0.9185 
F-statistic: 158.9 on 3 and 39 DF,  p-value: < 2.2e-16
```

Figure 6: Summary for Final Model

The final model employs 3 out of the 6 precipitation stations. In order to ensure that I was not removing any predictors that may contribute to the model, I ran an analysis of variance (ANOVA) test comparing the full and reduced models. The goal of the ANOVA test is to ensure that the subset of coefficients could be removed from the full model, all at once and in the presence of one another. My null hypothesis for this ANOVA test is that the estimated coefficient values associated with the variables (considered for removal) APMAM, APSAB, OPBPC and OPSLAKE are equal to one another as zero. My alternative hypothesis is that at least one of the estimated coefficient values associated with variables APMAM, APSAB, OPBPC and OPSLAKE is different from the others, and not equal to zero. The p-value from the ANOVA test is very large, equal to 0.97. As a result, I fail to reject the null hypothesis and can very confidently conclude that none of the estimated values at the aforementioned stations are significant and can be removed from the full model all at once, and in the presence of all predictors (R code, ANOVA section).

The final model selected was based on the following rationale: (1) all predictors are highly significant in the presence of others (see Figure 6), (2) the coefficient of determination is 0.9185, which shows that the predictors do account for the variability in Y at an almost 100% rate, (3) the estimate values of the model all make sense in the context of the problem and are positive, (4) the model’s diagnostic plots (Figure 7) all satisfy the assumptions of a valid multiple linear regression model, very well.

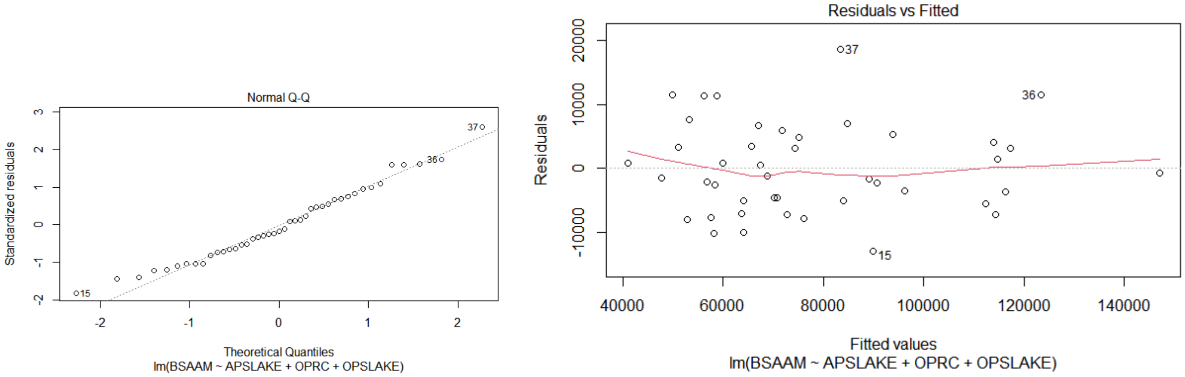


Figure 7: Diagnostic Plots for Final Model

I conclude that stream runoff can be very well forecasted by snow inches collected at precipitation stations, especially in the context of the Owens River Valley region in California. However, when creating multiple linear regression models for the purpose of forecasting and predicting stream runoff using precipitation data, it is important to utilize data that is not redundant. One of the difficulties in forecasting hydrologic phenomena is that, as I concluded from this study, time series data collected from precipitation stations in the same geographic vicinity is often highly autocorrelated and not independent. It is for this reason that any future multiple linear regression analyses used on the similar data be cautious in selecting which predictor variables to employ, and not to overfit the models, as to ensure the most reliable predictions possible.

Appendix

All code used for this paper can be downloaded at:

github.com/gabrieldaiess/STAT-632-Final-Project