# Thomas Allen

# Lakshman Arunachalam

# Group #2

# Week 6 Group Assignment

# University of Arkansas

**Executive Summary**

The data analysis provides a recommendation for implementing precipitation monitoring equipment in Bishop, CA. The following calculations are based upon precipitation measures as the input value to predict what the runoff in Bishop, CA, will be in any given year. Four different Models have been tried and at the end we found model 4 to be the best model to accurately predict the runoff in Bishop, CA, in any given year. Model 4 covers 91% of the variability in the data. And it also has a less percentage deviation from the average value by 10.36%. This predicted model is good. Using this model, following regression equation was made and the BSAAM value in 1991 is predicted to be 78,753.02 units.

Multiple linear regression:

**BSAAM = 16223.1 + 1321.2\*APMAM + 1636.9\*OPRC + 2456.8\*OPSLAKE**

Since measurement of river flow at APMAM (Mammoth LAKE), OPRC (Rock Creek), and

OPSLAKE is important in determining the Bishop runoff levels, these sites should receive the

expensive precipitation monitoring equipment.

Detail work on the model formation and its respective code is explained below.

Model Results

**Model 1**

For model 1 (All independent variables used)

**FIGURE 1**

Call:

lm(formula = BSAAM ~ ., data = train[, -1])

Residuals:

Min 1Q Median 3Q Max

-11971 -6501 -1094 5265 18116

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16211.4 4776.5 3.394 0.00239 \*\*

APMAM 737.1 928.5 0.794 0.43508

APSAB -334.3 1844.9 -0.181 0.85774

APSLAKE 1150.9 1759.3 0.654 0.51921

OPBPC 153.0 727.4 0.210 0.83518

OPRC 1740.5 740.3 2.351 0.02727 \*

OPSLAKE 2226.1 935.9 2.379 0.02568 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8077 on 24 degrees of freedom

Multiple R-squared: 0.9191, Adjusted R-squared: 0.8989

F-statistic: 45.46 on 6 and 24 DF, p-value: 6.099e-12

Equipment

Model 1, APSAB should have the precipitation monitoring equipment implemented. This is due to the intercept and coefficient not being significant and the high probability of the Pr (> | t | ) at 0.85774.

**Model 2**

**FIGURE 2**

Call:

lm(formula = BSAAM ~ APMAM + APSLAKE + OPBPC + OPRC + OPSLAKE,

data = train[, -1])

Residuals:

Min 1Q Median 3Q Max

-11968 -6484 -1194 5089 18117

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16063.4 4614.2 3.481 0.00185 \*\*

APMAM 696.3 883.2 0.788 0.43790

APSLAKE 907.0 1110.5 0.817 0.42178

OPBPC 163.0 711.2 0.229 0.82060

OPRC 1705.2 700.2 2.435 0.02235 \*

OPSLAKE 2254.5 904.7 2.492 0.01969 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7920 on 25 degrees of freedom

Multiple R-squared: 0.919, Adjusted R-squared: 0.9028

F-statistic: 56.74 on 5 and 25 DF, p-value: 7.723e-13

Equipment

OPBPC should have the precipitation monitoring equipment installed. This is due to the intercept and coefficient not being significant and the high probability of the Pr (> | t | ) at 0.82060.

**Model 3**

**FIGURE 3**

Call:

lm(formula = BSAAM ~ APMAM + APSLAKE + OPRC + OPSLAKE, data = train[,

-1])

Residuals:

Min 1Q Median 3Q Max

-12144 -6032 -1200 5301 18267

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 15827.8 4415.5 3.585 0.001367 \*\*

APMAM 730.6 854.5 0.855 0.400387

APSLAKE 890.4 1087.8 0.819 0.420499

OPRC 1697.0 686.4 2.472 0.020297 \*

OPSLAKE 2419.7 536.5 4.510 0.000123 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7774 on 26 degrees of freedom

Multiple R-squared: 0.9188, Adjusted R-squared: 0.9064

F-statistic: 73.6 on 4 and 26 DF, p-value: 8.572e-14

Equipment

APSLAKE should have the precipitation monitoring equipment installed. This is due to the intercept and coefficient not being significant and the high probability of the Pr (> | t | ) at 0.420499.

**Model 4 (best/final model)**

**FIGURE 4**

Call:

lm(formula = BSAAM ~ APMAM + OPRC + OPSLAKE, data = train[, -1])

Residuals:

Min 1Q Median 3Q Max

-13136 -5838 -2109 5630 18718

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 16223.1 4362.1 3.719 0.000926 \*\*\*

APMAM 1321.3 454.6 2.906 0.007218 \*\*

OPRC 1636.9 678.3 2.413 0.022876 \*

OPSLAKE 2456.8 531.3 4.624 8.37e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7726 on 27 degrees of freedom

Multiple R-squared: 0.9168, Adjusted R-squared: 0.9075

F-statistic: 99.12 on 3 and 27 DF, p-value: 1.088e-14

Equipment

OPRC should have the precipitation monitoring equipment installed. This is due to the intercept and coefficient not being significant and the high probability of the Pr (> | t | ) at 0.022876.

Insights in Plain English

In this regression model, BSAAM (Bishop runoff levels) has a positive coefficient, so as the

location’s precipitation increases the BSAAM will increase as well. For each APMAM unit

increase in precipitation, we expect to see, a 1321.3-unit increase in BSAAM. For each OPRC

unit increase in precipitation, we expect to see, a 1636.9-unit increase in BSAAM. For each OPSLAKE unit increase in precipitation, we expect to see, a 2456.8-unit increase in BSAAM.

Predicted BSAAM= 16223.1 + 1321.3 APMAM

Predicted BSAAM= 16223.1 + 1636.9 OPRC

Predicted BSAAM= 16223.1 + 2456.8 OPSLAKE

As a trial and error, we tried removing OPRC based on its P value being high among others to

see how the model would change (**MODEL 5** and **FIGURE 5**).

**MODEL 5**

**FIGURE 5**

Call:

lm(formula = BSAAM ~ APMAM + OPSLAKE, data = train[, -1])

Residuals:

Min 1Q Median 3Q Max

-13937 -5757 -1390 4705 19510

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 19176.1 4533.3 4.230 0.000226 \*\*\*

APMAM 1471.6 487.6 3.018 0.005372 \*\*

OPSLAKE 3610.1 251.3 14.363 1.92e-14 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8365 on 28 degrees of freedom

Multiple R-squared: 0.8988, Adjusted R-squared: 0.8916

F-statistic: 124.3 on 2 and 28 DF, p-value: 1.181e-14

**Best Model**

The adjusted R-value is just 89% for Model 5 which is less than the previous model. So, Model 4 is the best. Since it explains about 91% of the variability in the data. The overall model is

statically significant as the p-value is less than 0.005 (p level) with the value as 1.181e-14 which is very low (0.000000000001%).

When calculating the decided model for testing with test data, we found that RSME to be 8055.026. The mean of the target variable on original data is 77756.05, on an average, the prediction varies by 10.36% from the mean.

Final Model

The calculated model predicts a high 91% variability in the data and it also has a less percentage deviation from the average by 10.36% This predicted model is good.

The linear equation is:

**BSAAM = 16223.1 + 1321.2\*APMAM + 1636.9\*OPRC + 2456.8\*OPSLAKE**

**YEAR 1991**

The BSAAM in 1991, is 78,753.02 units.

Following the linear equation, **BSAAM = 16223.1 + 1321.2\*APMAM + 1636.9\*OPRC + 2456.8\*OPSLAKE**, ends up with that result.

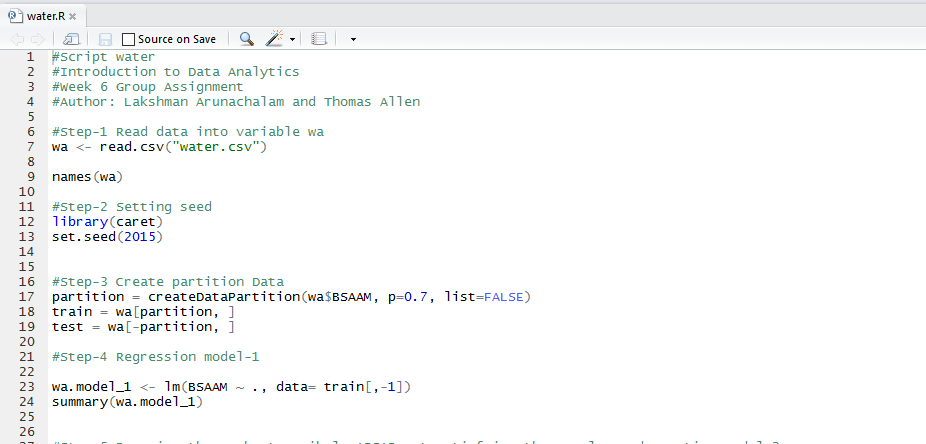
BSAAM= 16223.1 + (1321.2\*APMAM) + (1636.9\*OPRC) + (2456.8\*OPSLAKE)

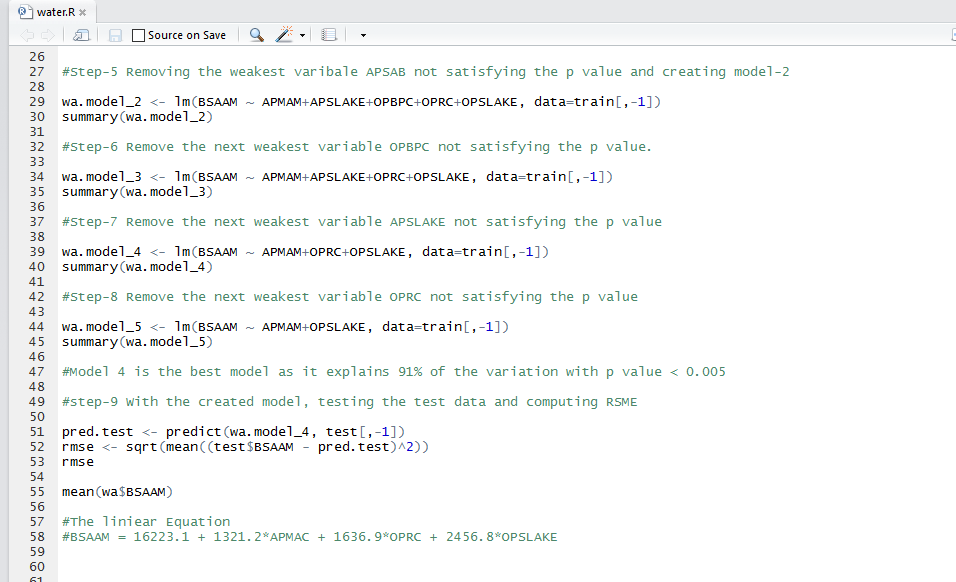
= 16223.1 + (1321.2\* 7.32) + (1636.9 \* 12) + (2456.8\* 13.52)

= 78,753.02 units

R scripts used for This assignment is attached in the following pages.

**R Script**





References

Viswanathan, V. (2015). *Data analytics with R: a hands-on approach* (2nd ed.). East Brunswick, NJ: Infivista Inc.