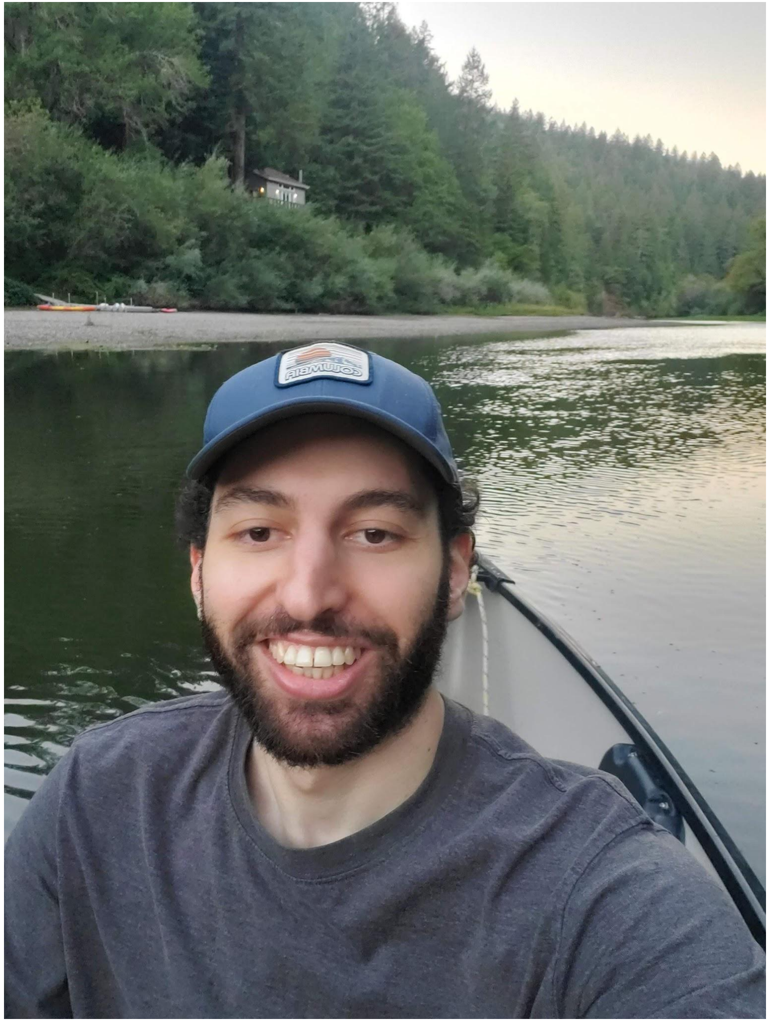
Runoff in the Owens River Valley, CA

Multiple Linear Regression analysis by Gabriel Daiess



CALIFORNIA NEVADA RIVER FORECAST CENTER
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

CLIMATE STATION	SINCE MIDNITE TOTAL	OCT 01- APR 30 2023	PON	OCT 01- APR 30 2022	PON	OCT 01- APR 30 NORMAL	OCT 01- SEP 30 NORMAL
...NORTHERN CALIFORNIA...							
MEDFORD OR	0.00	10.81	70	12.33	80	15.36	18.43
KLAMATH FALLS OR	T	6.88	80	5.26	61	8.64	11.14
CRESCENT CITY	0.06	49.36	94	36.86	70	52.30	57.98
EUREKA	0.05	39.58	107	22.59	61	37.00	40.40
UKIAH	0.00	37.82	114	18.70	57	33.08	34.84
MONTAGUE / SISKIYOU	0.00	6.35	66	3.62	38	9.61	11.99
ALTURAS	0.00	10.52	119	6.11	69	8.81	11.68
MOUNT SHASTA CITY	0.00	40.90	122	17.77	53	33.41	42.63
REDDING	0.00	36.33	120	17.88	59	30.31	33.52
RED BLUFF	0.00	25.88	122	11.92	56	21.13	23.12
SACRAMENTO EXEC AIRPORT	0.00	21.67	127	16.44	97	17.03	18.14
SACRAMENTO - CSUS	0.00	25.84	144	16.39	91	17.96	19.20
BLUE CANYON AIRPORT*	0.00	83.63	148	57.43	102	56.36	62.44
SOUTH LAKE TAHOE	0.00	34.09	194	18.00	103	17.56	20.46
SANTA ROSA	0.00	40.97	128	25.42	79	32.01	33.78
SAN FRANCISCO	0.00	32.54	149	18.44	85	21.82	22.89
SFO INT'L AIRPORT	0.00	30.66	162	18.12	96	18.91	19.64
OAKLAND AIRPORT	0.00	29.77	168	16.87	95	17.73	18.68
LIVERMORE	0.00	20.79	145	12.34	86	14.31	15.18
SAN JOSE INT'L AIRPORT	0.00	14.78	116	7.29	57	12.79	13.48
...CENTRAL CALIFORNIA...							
STOCKTON	0.00	22.87	180	9.75	77	12.69	13.45
MODESTO	0.00	19.31	169	8.99	79	11.40	12.27
MERCED	0.00	20.07	181	7.44	67	11.08	11.80
MADERA	0.00	10.83	107	2.10	21	10.12	10.79
FRESNO	0.00	17.48	171	6.29	61	10.25	10.99
HANFORD	0.00	14.48	189	6.34	83	7.65	8.13
BAKERSFIELD	0.00	9.74	162	5.40	90	6.01	6.36
BISHOP	0.00	13.66	332	4.75	115	4.12	4.84
DEATH VALLEY NP	0.00	1.06	62	M	M	1.72	2.20
SALINAS	0.00	13.73	114	7.31	61	12.04	12.58
PASO ROBLES	0.00	20.51	176	8.70	75	11.65	12.15
SANTA MARIA	0.00	23.18	181	7.79	61	12.79	13.32

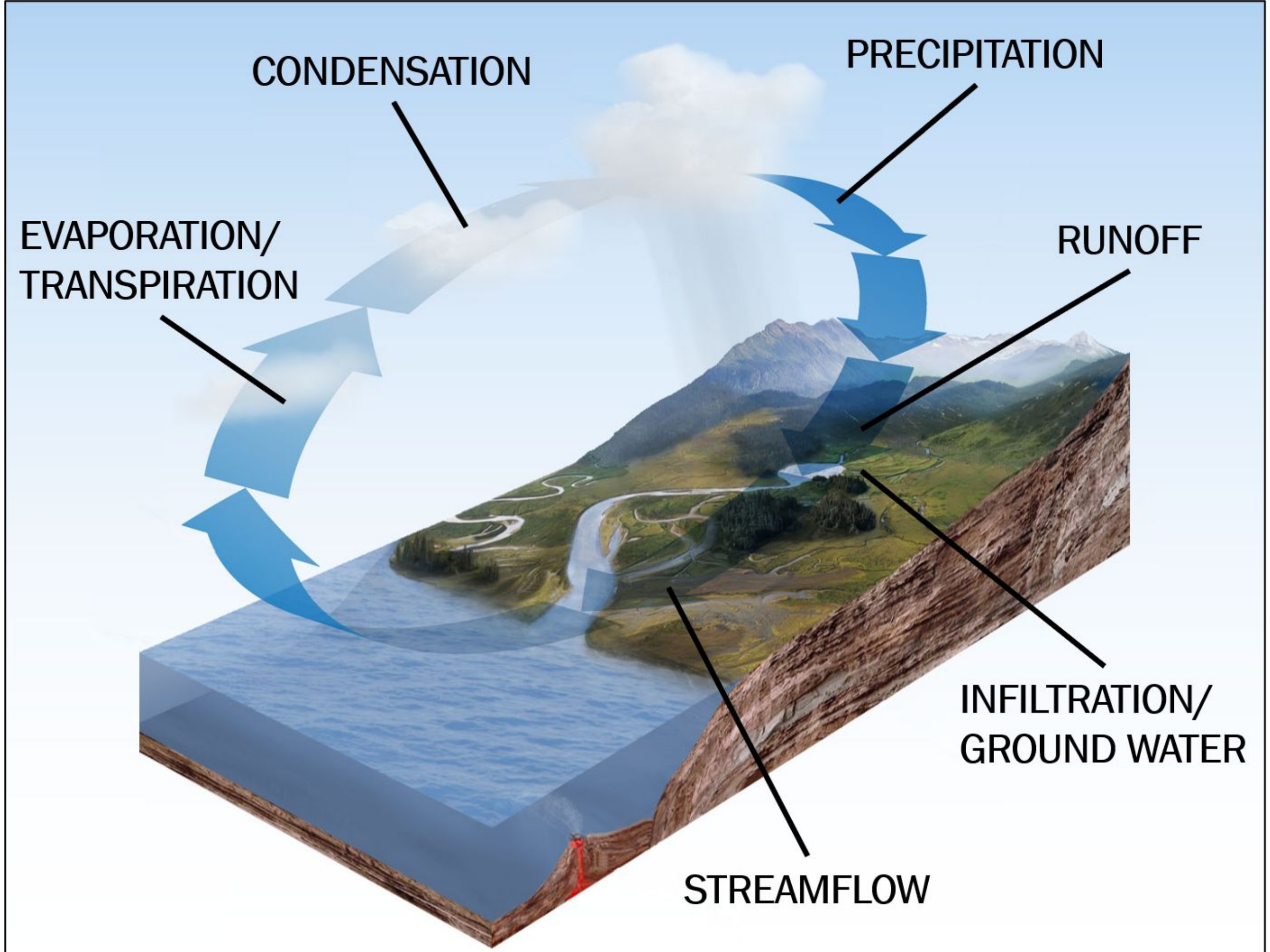
Research Question:

Is there a (linear) relationship between stream runoff and precipitation?

If so, can we use multiple linear regression to explain the relationship?

stream runoff (response Y)

precipitation (predictor X)



The Water Cycle

Variables

Response

43 observations at 1 site near Bishop, CA

Runoff volume (acre-feet) from 1948-1990

- BSAAM

Predictors

43 observations at 6 precipitation stations

Snow (inches) from 1948-1990 at

- Lake Mammoth (APMAM)
- Lake Sabrina (APSAB)
- South Lake (APSLAKE)
- Big Pine Creek (OPBPC)
- Rock Creek (OPRC)
- Rock Creek Lake (OPSLAKE)

Data Source

> library(alr4)

> data(water)

Applied Linear Regression (4th ed.)

Sandy Weisberg

R: California water ▾ Find in Topic

water {alr4} R Documentation

California water

Description

Can Southern California's water supply in future years be predicted from past data? One factor affecting water availability is stream runoff. If runoff could be predicted, engineers, planners and policy makers could do their jobs more efficiently. Multiple linear regression models have been used in this regard. This dataset contains 43 years worth of precipitation measurements taken at six sites in the Owens Valley (labeled APMAM, APSAB, APSLAKE, OPBPC, OPRC, and OPSLAKE), and stream runoff volume at a site near Bishop, California.

Format

This data frame contains the following columns:

Year
: collection year

APMAM
: Snowfall in inches measurement site

APSAB
: Snowfall in inches measurement site

APSLAKE
: Snowfall in inches measurement site

OPBPC
: Snowfall in inches measurement site

OPRC
: Snowfall in inches measurement site

OPSLAKE
: Snowfall in inches measurement site

BSAAM
: Stream runoff near Bishop, CA, in acre-feet

Source

Source: http://www.stat.ucla.edu.

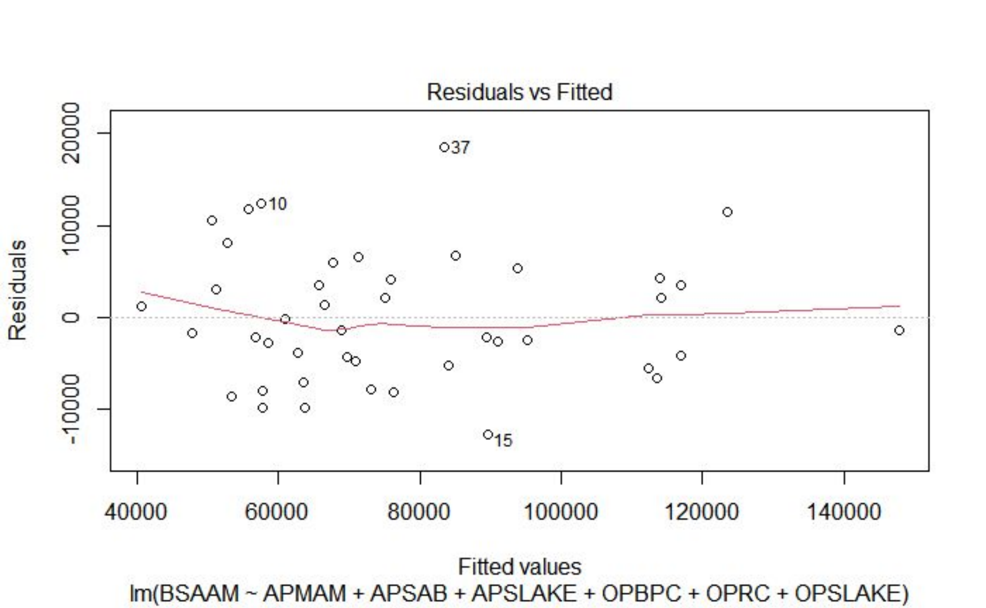
References

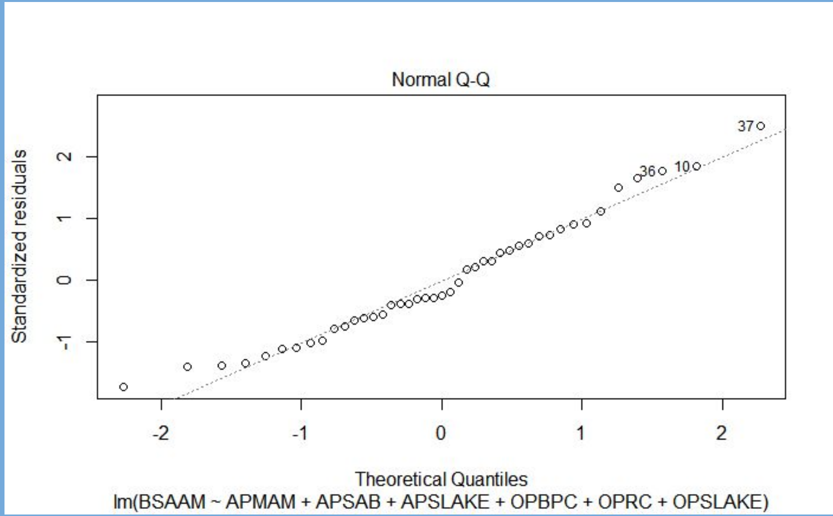
Weisberg, S. (2014). *Applied Linear Regression*, 4th edition. Hoboken NJ: Wiley.

Model 1: Full Model all 6 Stations

```
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
```

Model 1: Diagnostics Satisfied (!?)





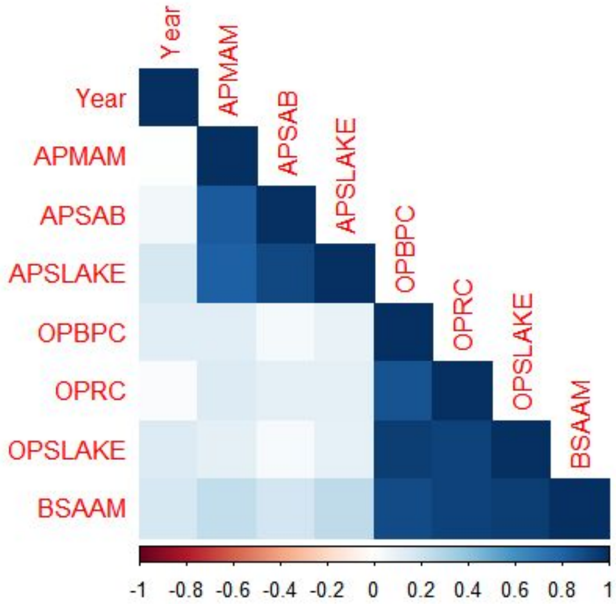

```
        Shapiro-Wilk normality test

data:  resid(lm1)
W = 0.97408, p-value = 0.4327


        studentized Breusch-Pagan test

data:  lm1
BP = 1.6605, df = 6, p-value = 0.9481
```

Multicollinearity

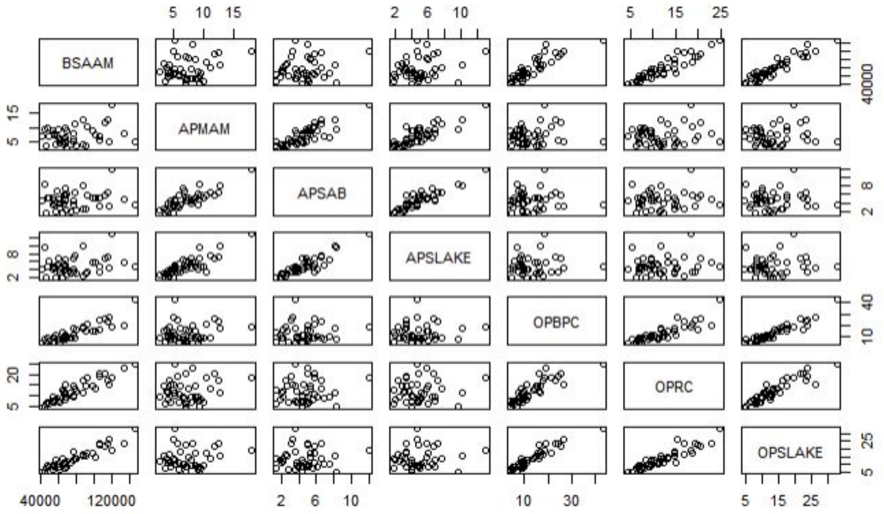
Only 2 of the 6 stations were significant in the initial model despite a very high R^2

The signs of estimates at stations APMAM and APSAB were negative

MLR assumptions well-satisfied despite these issues



Multicollinearity



Station	VIF
APMAM	3.55
APSAB	7.18
APSLAKE	6.75
OPBPC	9.27
OPRC	7.65
OPSLAKE	16.97

LET'S EXPLORE A NEW TOPIC

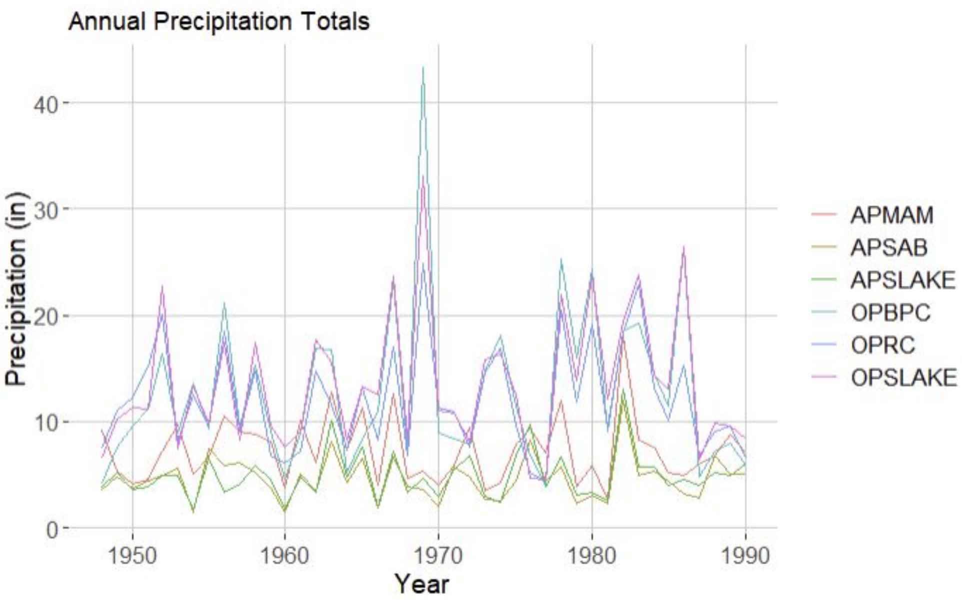
..BRIEFLY

Time Series

Concerned with analyzing observations that are collected over intervals of time (regular or irregular intervals)

Useful for **forecasting:** predicting future values from previous data and their trends

Regression is used in Time Series: Y is the forecast variable and X is the predictor



Autocorrelation

“Just as correlation measures the extent of a linear relationship between two variables, autocorrelation measures the linear relationship between *lagged values* of a time series.”

```
#Durbin-Watson Test for Autocorrelated Residuals
dwtest(lm1)
```
```

	Durbin-Watson test

data:  lm1
DW = 1.4362, p-value = 0.02554
alternative hypothesis: true autocorrelation is greater than 0
```

(Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting : Principles and Practice* (2nd ed.). Otexts. https://otexts.com/fpp2/)

Create a Reduced Model

1. step-wise variable selection
2. ANOVA for subset of stations
3. create a reduced model 2

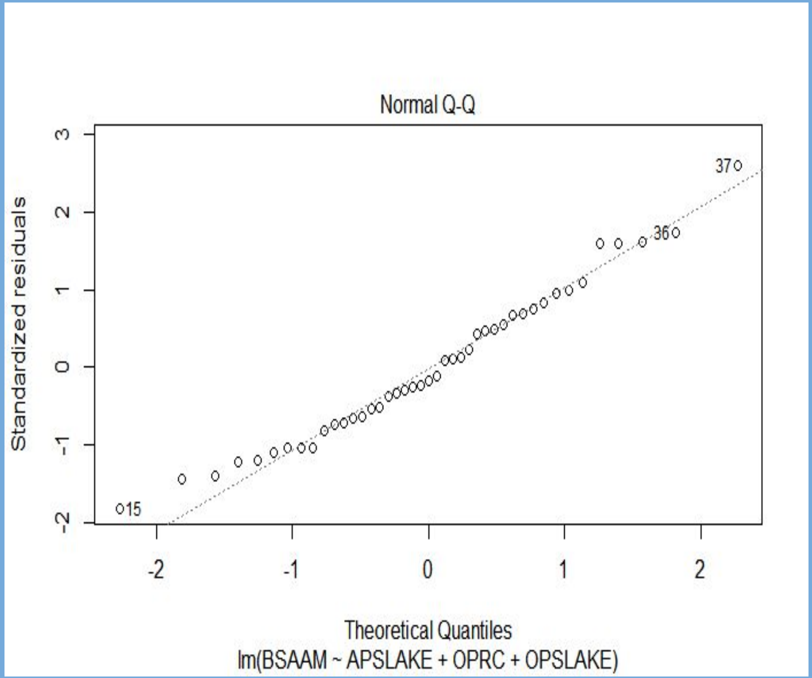
$$\widehat{StreamRunoff} = 15424.6 + 1712.5 APSLAKE + 1797.5 OPRC + 2389.8 OPSLAKE$$

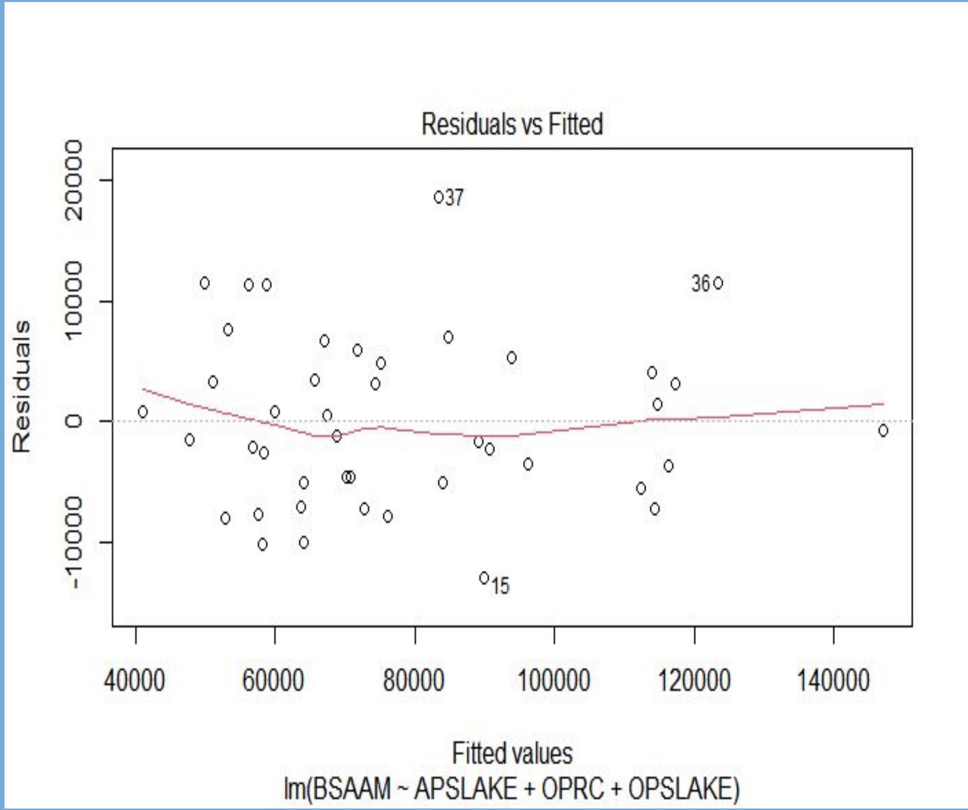
```
lm(formula = BSAAM ~ APSLAKE + OPRC + OPSLAKE, data = df)

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
```

Final Model

Tests for MLR

```
	Shapiro-Wilk normality test

data:  resid(lm2)
W = 0.97377, p-value = 0.4227


	studentized Breusch-Pagan test

data:  lm2
BP = 1.2524, df = 3, p-value = 0.7405
```


Normal Q-Q
Standardized residuals
3
2
1
0
-1
-2
37
36
15
-2
-1
0
1
2
Theoretical Quantiles
lm(BSAAM ~ APSLAKE + OPRC + OPSLAKE)


Residuals vs Fitted
Residuals
20000
10000
0
-10000
37
36
15
40000
60000
80000
100000
120000
140000
Fitted values
lm(BSAAM ~ APSLAKE + OPRC + OPSLAKE)

Limitations and Further Research

- Cookbook/Data background very limited
  - BSAAM response not well defined geographically or in terms of hydrology

- Reproducibility for Future Studies
  - Station names were not given full names for the dataset, which makes it hard to
    - update the dataset and continue with forecasting.
    - validate the model against “real-life” data

- Textbook Example
  - Meant to be open to interpretation, more than ONE right answer (log transformation could work too)
  - Simplified without too many observations (more stations could be added)

In the future I look forward to doing a similar analysis of time series data to forecast hydrologic phenomenon in CA