

G2M81, Statistical Modelling: Project

Summary

The goal of this project is to establish a link between two hydrological measurements for a particular area, namely precipitation (rainfall) and stream runoff (excess water going into streams). By establishing a link, we facilitate a method for extending the time span for which we can forecast the estimated water level of a reservoir. This increases efficiency and saves money. To establish the link, we use data collected over 43 years in Southern California, USA.



Using linear regression, we start by including all possible explanatory variables in the regression model. However, it quickly becomes apparent that some of the explanatory variables are dependent on each other. This multicollinearity (whose presence can be shown by using plots and correlation tests) forces us to rebuild the model, using variable selection techniques such as “Best Subsets” to establish the best regression model for use with the data.

Taking careful note of the Coefficients of Determination R^2 , Mallow’s Statistic C_p , and Residual Mean Square Errors, we select an optimal model that uses three explanatory variables. This model has a high R^2 value (92.4%), significant predictors (at the 0.005% level), and a highly significant F-Value, so the quality of the model is good.

The model does have some drawbacks, such as moderate Variance Inflation Factors, two influential observations and an observation with a large standardised residual. But if we try to remove any of the above imperfections, we obtain other, more severe problems, such as normal probability plots of residuals not following straight lines.

To test the assumptions we make in constructing the model, we use standard methods such as making sure that the *residuals vs. fits* graph has values evenly distributed about zero. The model passes all such tests performed on it. Because the data is ordered with regard to time, there is a possibility that autocorrelation effects such as global warming may come into effect. We disprove this for our model by using the Durbin-Watson statistic.

This leaves us with a valid model that may be used to predict stream runoff for any given year from three distinct precipitation measurements. The equation of the model is as follows:
$$\text{Stream Runoff} = 15425 + (1712 \times \text{APSLAKE}) + (1797 \times \text{OPRC}) + (2390 \times \text{OPSLAKE}).$$
By obtaining the above equation, we satisfy the goal of the project.

Table of Contents

Page	Description
2	<i>Summary</i>
4-5	<i>Introduction</i>
5-16	<i>Analysis:</i>
6	<i>Section 1: Initial Analysis</i>
7	<i>Section 2: Visual Analysis of Regression</i>
8	<i>Section 3: Regression</i>
15	<i>Section 4: Autocorrelation</i>
16	<i>Section 5: Prediction</i>
17	<i>Conclusions; Bibliography; References</i>
18	<i>Appendix 1: Data</i>
19-23	<i>Appendix 2: Minitab Output</i>

Internet Resources

Project Home Page: www.bangor.ac.uk/~mau402/g2m81project

Minitab Statistics Package: www.minitab.com

UCLA Statistics Department: www.stat.ucla.edu

Los Angeles Department of Water and Power: <http://www.dwp.ci.la.ca.us/home.htm>

The Official Bishop, CA Home Page: <http://www.bishopweb.com/>

ABRFC Forecast Methodology: <http://info.abrfc.noaa.gov/fcstmethods.html>

Mammoth Lakes: <http://www.mammothweb.com>

F-Tables: <http://www.psychstat.smsu.edu/introbook/fdist.htm>

Durbin-Watson statistic: <http://www.csus.edu/indiv/j/jensena/mgmt105/durbin.htm>

Data Source: *Background Material:* www.stat.ucla.edu/cases/dwp/index.html

The Data: www.stat.ucla.edu/cases/dwp/dwp-data.html

Introduction

To provide a water supply to any location, we must first have a water source, usually in the form of a reservoir. Monitoring the level of the reservoir is very important, and enables decisions to be made by engineers, planners and policy makers. For example, if the level of the reservoir is low, then a hose pipe ban is usually enforced, reducing the water used from the reservoir, and making sure it doesn't dry out.

Stream runoff

Stream runoff is the amount of rainfall that is carried off an area of land by a stream, or the amount of water that the land cannot absorb. In other words, it is the excess water that goes into a river or stream. Because of this, it directly affects the level of reservoirs. Measured by Gage Houses, Wire Weight Gages or Staff Gages, stream runoff measurements are used to provide river forecasts. These are primarily used for flood forecasting, but are also used to forecast the water level of reservoirs.

Now consider the advantages of being able to *predict the stream runoff* for a given area. This would enable us to forecast further into the future, and decisions like the above could be issued **earlier**, and hence (in the example) more water could be saved. Any accurate technique will improve efficiency and save time and money.

Precipitation

Precipitation occurs earlier in the water cycle than stream runoff, and so if we could predict stream runoff from precipitation, time could be saved. Liquid precipitation or rainfall is traditionally measured using apparatus such as the standard rain gage or a tipping-bucket. These measure precipitation at a point. While the standard gage measures precipitation weight using a funnel, a can, and a measuring tube (to an accuracy of 0.01 inch), tipping-buckets consist of a funnel and rocker mechanism. The rocker tips over once 0.01 inches of precipitation has fallen into the bucket.

Goal of the Project

Our goal in this project is to prove that there is a link between precipitation and stream runoff for a particular area in California, USA. At a location near Bishop, California stream runoff is measured at regular intervals. In the nearby area of Owens Valley, there are six locations that measure precipitation. For each year over a period of 43 years, the stream runoff was measured, together with the total precipitation for each year measured at the six sites. Using this data, if we can show that a link exists between precipitation and stream runoff, then we can use what we obtain to provide longer term forecasts for the reservoir levels, improving the present model.





Owens Valley

An area in South East California, the reservoir we are interested in is Owens Lake, (located near the town of Bishop) which has a river flowing to it for 175 miles before reaching the lake.

In the area around the lake are our precipitation sites, most of which you can see on the road map on the right.



(The picture on the left was taken near Lake Sabrina.)

Analysis

In order to understand the following analysis fully, let us first familiarise ourselves with the data that was collected between 1948 and 1990.

Year	APMAM	APSAB	APSLAKE	OPBPC	OPRC	OPSLAKE	BSAAM
1948	9.13	3.58	3.91	4.1	7.43	6.47	54235
1949	5.28	4.82	5.2	7.55	11.11	10.26	67567
1989	8.8	5.06	4.92	8.05	9.6	9.58	53965
1990	7.1	5.06	6.05	5.8	6.5	8.41	49774

As you can see above, we have for each year seven columns of data, which represent six precipitation measurements and one stream runoff measurement, as follows:

Column 2 (APMAM):	Precipitation Measurement (in inches) taken at <i>Mammoth Lake</i>
Column 3 (APSAB):	Precipitation Measurement (in inches) taken at <i>Lake Sabrina</i>
Column 4 (APSLAKE):	Precipitation Measurement (in inches) taken at <i>South Lake</i>
Column 5 (OPBPC):	Precipitation Measurement (in inches) taken at <i>Big Pine Creek</i>
Column 6 (OPRC):	Precipitation Measurement (in inches) taken at <i>Rock Creek</i>
Column 7 (OPSLAKE):	Precipitation Measurement (in inches) taken at <i>South Lake</i>
Column 8 (BSAAM):	Stream Runoff Measurement (in acre-feet) taken near <i>Bishop</i>

An important distinction has to be made between the two types of precipitation measurement above. We saw in the introduction **two** methods for obtaining precipitation at a point - standard gages and tipping-buckets. The first three sites, prefaced with “AP”, use tipping-buckets while the final three sites, prefaced with “OP”, use standard gages. We would expect differences between the two types of measurements.

Section 1: Initial Analysis

After typing the data into the statistics package Minitab, which will be used to perform all of the analysis in this project, the first thing to do is to run some tests on the data, asking for some basic descriptive statistics and some graphs.

Basic Descriptive Statistics

Variable	Mean	Median	TrMean	StDev	SE Mean	Minimum	Maximum	Q1	Q3
Year	1969.0	1969.0	1969.0	12.6	1.9	1948.0	1990.0	1958.0	1980.0
APMAM	7.323	7.080	7.126	3.098	0.472	2.700	18.080	4.930	9.130
APSAB	4.652	4.460	4.535	2.052	0.313	1.450	11.960	3.260	5.720
APSLAKE	4.930	4.620	4.751	2.258	0.344	1.770	13.020	3.340	5.900
OPBPC	12.84	9.55	12.15	7.69	1.17	4.05	43.37	7.90	16.75
OPRC	12.002	11.110	11.762	5.028	0.767	4.350	24.850	7.600	15.150
OPSLAKE	13.522	12.140	13.143	6.382	0.973	4.600	33.070	8.410	17.420
BSAAM	77756	69177	76297	25519	3892	41785	146345	58942	92715

One thing that is apparent from the above output is that there are some differences between the sites used to measure precipitation. For example, the mean precipitation for OPSLAKE is nearly three times the mean precipitation for APSAB. We would expect some differences to come from what we know about the type of measurements carried out for the first three sites compared to the last three sites. (The first three sites measure precipitation using tipping-buckets while the last three sites use standard gages.)

We notice from the output that mean tipping-bucket precipitation is generally lower than mean standard-gage precipitation. From the standard errors of the mean we notice that tipping-bucket precipitation is also generally more precise. We can conclude most about the differences in measurement technique by analysing the two measurements taken at South Lake (APSLAKE and OPSLAKE). We can clearly see that the mean and standard error are higher here for the standard-gage method of measurement.

This could be explained by the fact that in freezing weather, unlike the standard gage, the tipping-bucket cannot be used as the mechanism or the funnel hole can freeze solid. Hence the lower means for tipping-buckets. Frozen precipitation collected using a standard rain gage is melted into a measuring tube. Maybe this introduces some errors into the measurement. Hence the higher standard error.

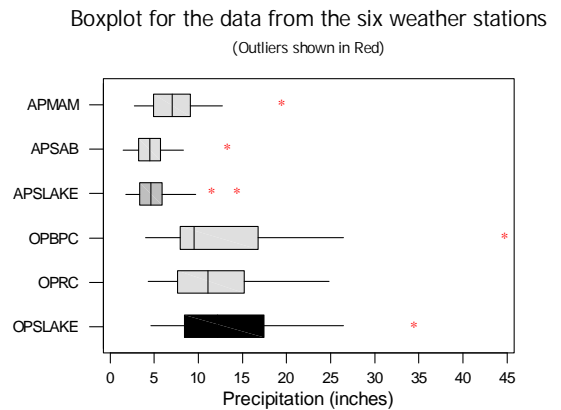
Differences can also be explained by other factors such as location and level of sheltering (trees, buildings, etc.). For example, the weather by the sea compared to inland is usually different as regards to precipitation. We would expect the location of two sites to have a bearing on the difference in mean precipitation between the sites. But as long as these factors remain constant, they shouldn't affect any results we may obtain.

If we wished to test for any significant differences between the means of any two of the above sites, we can use a paired-sample t-test. (We need a *paired* test because the samples from different locations are not independent - weather in one location **can** affect the weather in another location.) This dependence will be importance later when assessing multicollinearity - if the values of one explanatory variable depend on another, then we can remove one of the (thus redundant) variables from the regression model .

Graphs

To further analyse the data, I produced the boxplot shown on the right, showing a summary of the data from the six measurement sites of precipitation. You can clearly see the suggestion of outliers, shown in red, in five out of the six measurement sites.

After looking at the data, it is clear that the most extreme outlier for the first three weather stations comes from 1982, while the outlier for the fourth and sixth weather stations comes from 1969. I will have to take into account these possible outliers while performing regression analysis later on. They may have to be removed from the model.



Section 2: Visual assessment of regression

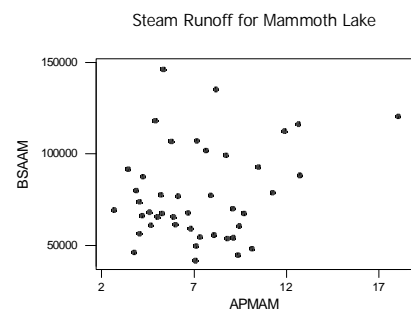
Plotting the data | Assessing the regression visually | Looking for outliers

The first task I took here was to plot the stream runoff values against the precipitation measurements for each weather station in six separate graphs, an example of which is shown on the right, and the rest of which (together with all Minitab output in full) can be found in *Appendix 2*.

I was looking for a rough linear relationship between stream runoff and precipitation, in the form of the values in the graph following a straight line. After plotting the graphs, I came to the conclusion that there **is** a visual linear relationship between stream runoff and the precipitation data for the following weather stations: OPBPC, OPRC and OPSLAKE; while there is **no** strong visual evidence of a linear relationship for the data from weather stations APMAM, APSAB and APSLAKE.

Notice that all the tipping-bucket sites seem to have a strong visual relationship. Maybe this is indicative that these measurements are more accurate or precise. This is backed up by the smaller standard errors we saw in section 1.

A recurring theme ran throughout the graphs, that is the appearance of a possible outlier disjoint from the main group of observations at a relatively high precipitation value. This follows on from our discussion of outliers in the previous section, and adds to the evidence. As an example, in the graph shown above, all precipitation measurements are in the range 3-13 apart from one value isolated at a value of about 18. We shall have to think about removing these outliers if they have a critical effect on the regression model.



Section 3: Regression

Obtaining the Best Model | Checking Unusual Observations | Checking Assumptions

Having assessed visually the relationship between stream runoff and the six explanatory variables, let us now apply a better technique of assessing relationships between variables, **regression**. To start with, let the stream runoff variable be the *response* variable and let the six precipitation related variables be *explanatory* variables.

The regression equation is

BSAAM = 15945 - 13 APMAM - 664 APSAB + 2271 APSLAKE + 70 OPBPC + 1916 OPRC + 2212 OPSLAKE

Predictor	Coef	SE Coef	T	P	VIF
Constant	15945	4100	3.89	0.000	
APMAM	-12.8	708.9	-0.02	0.986	3.5
APSAB	-664	1523	-0.44	0.665	7.2
APSLAKE	2271	1341	1.69	0.099	6.7
OPBPC	69.7	461.7	0.15	0.881	9.3
OPRC	1916.5	641.4	2.99	0.005	7.6
OPSLAKE	2211.6	752.7	2.94	0.006	17.0

S = 7557 R-Sq = 92.5% R-Sq(adj) = 91.2%

Using this model returns an R^2 value of 92.5%, which means that only 7.5% of the variability in the data is unaccounted for. But looking at the above output, we see that some of the predictors do not significantly contribute to the model. The evidence for this comes from looking at the p-values of the predictors.

From the above we see that the variables APMAM, APSAB, APSLAKE and OPBPC have p-values that are greater than 0.05. This means that at the 5% significance level, we would accept the null hypothesis that the coefficients of the above variables in the regression equation are not significantly different from zero to justify the inclusion of the variables in the model. In particular, three p-values are so high that the inclusion of the associated variable in the regression model is totally unjustified.

So what does this suggest? Looking at the variance inflation factors, we have values of up to 17.0. From theory we know that a variance inflation factor (VIF) of more than 10 strongly suggests the presence of multicollinearity. The regression coefficients are poorly estimated in this case. It would therefore be a good idea to assess the level of correlation between the six measurement sites.

Asking Minitab to do this, the output on the next page was gained, where the first value in each cell is the Pearson correlation value (which has a range of -1 to 1) and the second value is the p-value for the hypothesis test of the correlation coefficient being zero.

The interpretation of the correlation value is that if it is significantly different from zero, then we accept the alternate hypothesis that the two variables in question are in fact correlated. The p-value will allow us to see whether a correlation value is *significantly* different from 0.

	APMAM	APSAB	APSLAKE	OPBPC	OPRC	OPSLAKE
APMAM		0.828	0.816	0.122	0.154	0.108
APSAB	0.828		0.900	0.040	0.106	0.030
APSLAKE	0.816	0.900		0.093	0.106	0.101
OPBPC	0.122	0.040	0.093		0.865	0.943
OPRC	0.154	0.106	0.106	0.865		0.919
OPSLAKE	0.108	0.030	0.101	0.943	0.919	
	0.492	0.850	0.521	0.000	0.000	

The presence of high correlation coefficients and low p-values in the yellow blocks strongly suggests that there is some multicollinearity in the data. Note that the high coefficients come when we compare two sites that use the same method of measuring precipitation. When we compare two sites which use different methods, the correlation coefficients are not significantly different from zero at the 5% significance level.

It would follow from this that when picking variables to include in the regression model, we should include at least one variable from each method of measuring precipitation. Using two explanatory variables, this would mean picking one variable from the first three sites and the other from the final three sites.

Obtaining the Best Model

So we have established that we need to remove some variables from our initial model that used six explanatory variables. How do we go about this? The answer lies in examining all subsets of the six explanatory variables and picking the “best” model from the plethora of candidate regressor variables. Using Minitab’s *Best Subsets* command to accomplish this, we get the following output:

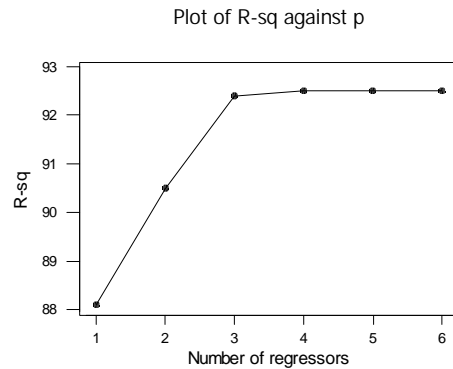
Best Subsets Regression: BSAAM versus APMAM, APSAB, ...

Response is BSAAM

Vars	R-Sq	R-Sq(adj)	C-p	S	A O						
					A	P	S	O	P	O	
1	88.1	87.8	18.2	8922.4							X
1	84.6	84.2	34.9	10145							X
2	90.5	90.0	8.5	8063.3			X				X
2	90.5	90.0	8.6	8065.4	X						X
3	92.4	91.9	1.2	7283.5	X		X				X X
3	91.8	91.2	4.1	7566.1	X						X X
4	92.5	91.7	3.0	7357.7	X	X					X X
4	92.4	91.6	3.2	7376.2	X	X					X X
5	92.5	91.5	5.0	7454.1	X	X	X				X X
5	92.5	91.5	5.0	7456.4	X	X	X				X X
6	92.5	91.2	7.0	7556.9	X	X	X	X			X X

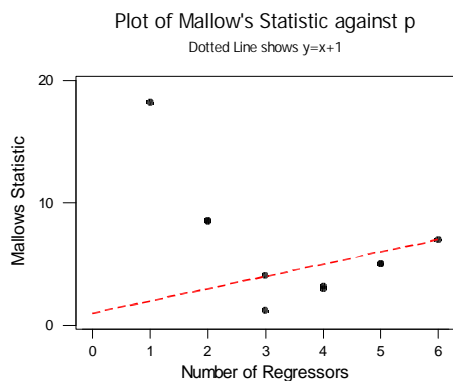
In the output we see several columns of data, each of which is important in deciding which model to choose. The first column shows the number of regressors included in the model. This will affect the complexity of the model obtained. The second column shows the R-sq value for a particular model. This is calculated by dividing the regression sum of squares by the total sum of squares.

To choose the best model in terms of R-sq value, we plot a graph as shown consisting of the highest R-sq value for each p, where p is the number of regressors. We then look for the “knee” of the graph, where the greatest change in slope occurs. In this case the “knee” occurs at $p=3$, and so based on R-sq values, we would choose the model highlighted in red on the output on the previous page.



The third column gives the adjusted R-sq value, and applying the same sort of method as above again returns the model highlighted in red as the “best” model here.

The fourth column gives Mallows’s statistic, C_p . Perhaps the most important statistic to take note of when choosing a model, the statistic is an estimate of the average total error of prediction based on a model containing p terms. A good model (in Minitab) is characterised by when the value of C_p is approximately equal to $p+1$. If it is smaller than $p+1$, then it is an even better model. Going with this argument, on the graph shown opposite, we have six possible models for where $C_p \leq p+1$ (the values beneath the red dotted line). Then choosing the model with the *smallest* value of C_p , we choose one of the models at $p=3$, again this is the model highlighted in red on the previous page.



Finally, the fifth column, headed “S”, is the square root of the Residual Mean Square Error usually found in an analysis of variance table. S^2 is calculated by dividing the residual sum of squares by the residual degrees of freedom. Using the proposal that the lowest value of S will give a favourable model, looking for the lowest value of S in the table again gives us the model highlighted red.

In conclusion, we have been lucky here, as all the above four methods for selecting an advantageous model actually give the same model, where we have **three** regressor or explanatory variables, namely APSLAKE, OPRC and OPSLAKE. This model gives the R-sq and the adjusted R-sq values at the “knee” of a corresponding graph; gives the smallest value of C_p , and gives the smallest value of S. We will therefore feel confident that this is the optimal model for use with this data set.

To be absolutely sure that this is the correct model to use, let us use some stepwise regression techniques for confirmation. As we can see on the right, Minitab output for normal stepwise regression gives us the model we have earmarked for use on the previous page.

Changing the options to allow Minitab to perform Forward Selection and Backward Elimination (changing the Alpha-to-enter and the Alpha-to-remove) gives exactly the same results.

So we have come to the point where we have our optimal model. We can now ask Minitab to perform regression using this model, and this is the (edited) output we obtain:

Step	1	2	3
Constant	27015	19145	15425
OPSLAKE	3752	3690	2390
T-Value	17.39	18.83	5.35
P-Value	0.000	0.000	0.000
APSLAKE		1769	1712
T-Value		3.19	3.42
P-Value		0.003	0.001
OPRC			1797
T-Value			3.17
P-Value			0.003
S	8922	8063	7284
R-Sq	88.07	90.49	92.44
R-Sq(adj)	87.78	90.02	91.85

Regression Analysis: BSAAM versus APSLAKE, OPRC, OPSLAKE

The regression equation is
 BSAAM = 15425 + 1712 APSLAKE + 1797 OPRC + 2390 OPSLAKE

Predictor	Coef	SE Coef	T	P	VIF
Constant	15425	3638	4.24	0.000	
APSLAKE	1712.5	500.5	3.42	0.001	1.0
OPRC	1797.5	567.8	3.17	0.003	6.5
OPSLAKE	2389.8	447.1	5.35	0.000	6.4

S = 7284 R-Sq = 92.4% R-Sq(adj) = 91.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	25282070749	8427356916	158.86	0.000
Residual Error	39	2068947585	53049938		
Total	42	27351018334			

Source	DF	Seq SS
APSLAKE	1	1700437520
OPRC	1	22065714689
OPSLAKE	1	1515918540

Unusual Observations

Obs	APSLAKE	BSAAM	Fit	SE Fit	Residual	St Resid
35	13.0	120463	117277	4355	3186	0.55 X
37	5.7	102001	83352	1218	18649	2.60R
39	4.6	118144	114058	4273	4086	0.69 X

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.46

The first thing we notice is that the P-values of the predictors are *all* less than 0.005, which means that at the 0.005% significance level, we accept the alternate hypothesis that all of the predictors' coefficients are significantly different from zero to justify the inclusion of the predictor in the model. Therefore the model we have chosen has been justified. In addition to being justified, the R² value for the model is 92.4%, so the quality of the fitted model is good.

Looking at the Variance Inflation Factors (VIF), we see that the APSLAKE variable has no linear relationship whatsoever with the remaining predictors, at it has a VIF of 1.0. We *are* slightly concerned with the two other VIF values, 6.4 for OPSLAKE and 6.5 for OPRC. While they do not exceed 10, in which case the presence of multicollinearity would be strongly suggested, the values do suggest some multicollinearity between the standard gage measurements for South Lake and Rock Creek. This is confirmed by looking at previously obtained correlation output, where the correlation coefficient for the test between these two sites was 0.919 with a p-value of 0.000

In my opinion, it is justifiable to leave this multicollinearity in the model because if we were to remove one of the variables from the model, problems arise. If we were to remove the data from Rock Creek from the model, this would mean that we would be wholly dependent on the weather at one location only, South Lake. I feel it is important to have more than one measurement location as the water supply for the reservoir comes from more than one location, and the weather at one location is not necessarily representative of the weather across a large area.

Removing the standard gage measurement for South Lake (OPSLAKE) from the model seems more sensible, but doing this produces a regression model with a reduced R-Sq value (86.9%) and, crucially, a Normal Probability Plot that does not follow a straight line. We will analyse these kind of plots in more detail later.

So, given the alternatives, I feel it is best to stick with the present model.

Continuing with the analysis, we come to the Analysis of Variance. To assess whether the regression is highly significant, let us analyse the F-Ratio for the output. The upper 5 percent point of the F-distribution on 3 and 39 degrees of freedom is 2.8451 while the upper 1 percent point is 4.3274. The calculated ratio of 156.86 is well in excess of both critical values, so we conclude that the regression is highly significant with useful explanatory variables. The p-value given of 0.000 confirms this.

Towards the end of the output, we see three “unusual observations”. Given the potential outliers we saw earlier in the report, these unusual observations are to be expected. The suspect data occurs at observations 35, 37 and 39. These correspond to years 1982, 1984 and 1986.

So the question arises as to whether we should remove these observations from the data. Let us start by analysing observation 37 (1984) which has a large standardised residual. The values for this row in the data set don’t particularly seem to stand out as outliers, and the leverage for the row is very small (about 0.028). This coupled with the fact that we didn’t notice observation 37 as an outlier in sections 1 and 2 makes me think that we should leave this observation in the data set.

Year	APMAM	APSAB	APSLAKE	OPBPC	OPRC	OPSLAKE	BSAAM
1981	2.7	2.22	2.48	8.99	9.45	12.14	69177
1982	18.08	11.96	13.02	18.55	18.4	19.45	120463
1983	8.2	4.98	5.76	19.25	22.9	23.86	135043
1984	7.65	5.3	5.74	14.45	13.15	14.42	102001
1985	5.22	4.42	4.04	11.45	10.16	13.06	77790
1986	4.93	3.26	4.58	26.47	15.33	26.46	118144
1987	5.99	2.76	3.98	4.8	6.85	6.36	61229

Let us now look at the observations with large influence, observations 35 (1982) and 39 (1986). From the data, these look as if they could be possible outliers, and we noticed this for observation 35 (1982) in section 1 of this report. This leads us to removing these observations from the data set and recalculating the regression analysis:

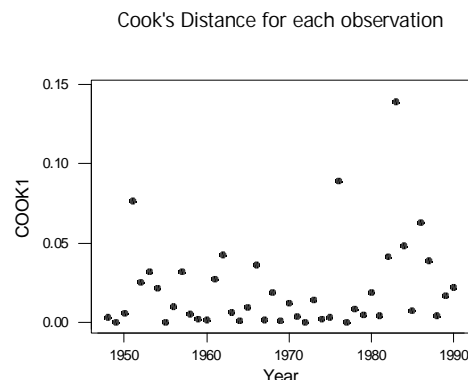
The regression equation is
 $BSAAM = 16336 + 1556 \text{ APSLAKE} + 1965 \text{ OPRC} + 2213 \text{ OPSLAKE}$

Predictor	Coef	SE Coef	T	P	VIF
Constant	16336	4193	3.90	0.000	
APSLAKE	1556.4	613.4	2.54	0.016	1.0
OPRC	1964.5	677.8	2.90	0.006	8.4
OPSLAKE	2212.7	555.9	3.98	0.000	8.4

S = 7406 R-Sq = 91.4% R-Sq(adj) = 90.8%

But instead of improving the model, removing the observations from the data actually worsens it! Looking at the above output, the R-Sq value has marginally decreased and the Variance Inflation Factors have increased for OPRC and OPSLAKE. In addition (see Appendix 2), the F-Value has decreased, some predictors are not now significant at the 1% level and new unusual observations are introduced. Removing these from the model worsens it further, for instance decreasing the F-Value to 98.13. While still large, the F-Value has decreased around 37% compared with the original F-Value.

Another way of assessing whether to remove an observation from the model is to use Cook's Distance, a statistic combining leverages and Studentised residuals into a measurement of how unusual an observation is. It is suggested that we check any observation with a Cook's distance of more than the value of the upper 50 percent point of the F-distribution on p and $n-(p+1)$ degrees of freedom. In this case, $p=3$ and $n-(p+1) = 39$, and the F value from tables is 0.8026. As you can see in the graph, no observation in our data set has a Cook's distance of more than 0.15. This suggests that we should *not* remove any observation from the data.



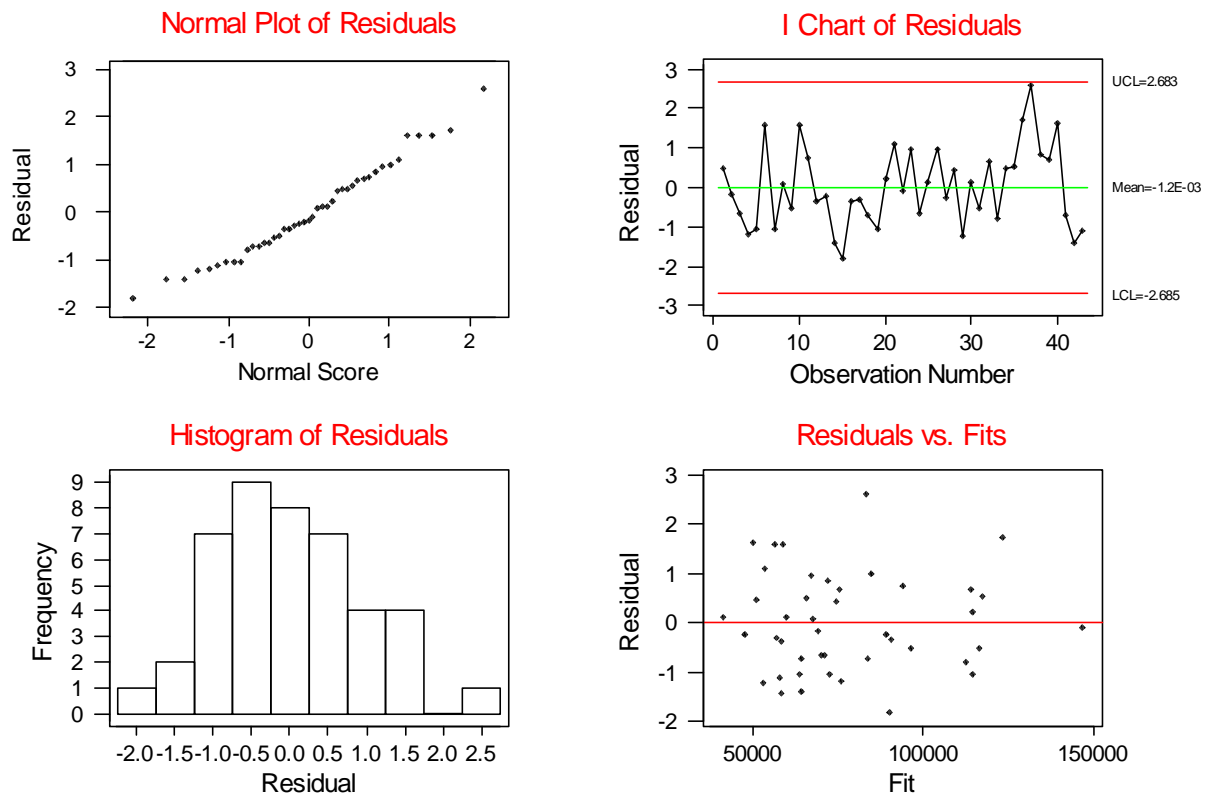
In my opinion, it is better to stick with the original model because of the reasons outlined above. Other reasons for sticking with the model include the lack of any direct evidence that the observations are in fact outliers, and the fact that the unusual observations are close to the end of the data, towards the present day, for which we will be trying to predict stream runoff for. Another important reason is that the "outliers", which are unusually high precipitation measurements, may in fact be valid, and be representative of a year when there was, for example, severe flooding. By removing the "outliers" from the data we would be excluding the extrapolation of the model to these values in future years when flooding might occur again.

In conclusion then, the original data set and the explanatory variables APSLAKE, OPRC and OPSLAKE will be used to predict stream runoff, and the equation to do this is given by
 $Stream\ Runoff = 15425 + 1712 \text{ APSLAKE} + 1797 \text{ OPRC} + 2390 \text{ OPSLAKE}$

Checking Assumptions

Now that we have a model, it is important to test the assumptions we make in constructing the model. In summary, we must test that the residuals are normally distributed, have constant variance and are independent. Minitab provides an excellent way to perform these tests, allowing us to perform a residual analysis with the **standardised** residuals and the fitted values generated by the regression output:

Residual Analysis for our Regression Model



The first thing we must check for is the assumption of normality of the residuals. We can confirm the normality by looking at a normal plot of the residuals and looking for a straight line. The top left graph is sufficiently straight in my opinion, so we conclude that the residuals are normally distributed. Also, as only 2 out of the 43 values are outside the range ± 2.0 on the x-axis, this satisfies the requirement that 95% of the values are inside the range ± 2.0 if the data is normally distributed.

Secondly, to check on the assumption of constant variance, we look at the graph of the Residuals vs. the Fits. As the points appear evenly scattered around zero without any systematic changes in spread, I conclude that the data has passed this check and the residuals are of constant variance.

Finally, to check on the assumption of independence, we look for the same kind of systematic changes in spread in the "Chart of Residuals". This is a plot of the standardised residuals against the order in which the observations were taken. As the data is again evenly spread, I conclude that the residuals pass the test of independence.

Section 4: Autocorrelation

The data set we are working with consists of 43 observations, ordered from 1948 to 1990. Over a relatively long period of time such as this, it is fair to say that changes could have happened to the values of the variables *because* of the long period of time. If changes have occurred, then the effects of **autocorrelation** have been felt.

We hear a lot about global warming these days, how the temperature of the earth is rising due to decreased ozone levels. Perhaps a side-effect of global warming is a change in precipitation behaviour. If a change exists, global warming is therefore a form of autocorrelation that might be present in our data.

To test for (first-order) autocorrelation in our data, and more specifically in our regression model, Minitab can calculate the Durbin-Watson statistic for us. This tests for the presence of autocorrelation in regression residuals by determining whether or not the correlation between two adjacent error terms is zero.

For our regression model, the value of the Durbin-Watson statistic was 1.46. To reach a conclusion about this value, we must construct the following table, which shows the possible conclusions that can be made. Note that the null hypothesis is that there is no autocorrelation and that d_1 and d_u are critical points taken from tables.

Regions of Acceptance and Rejection of the Null Hypothesis

Zero to d_1	d_1 to d_u	d_u to $(4-d_u)$	$(4-d_u)$ to $(4-d_1)$	$(4-d_1)$ to 4
Reject Null Hypothesis: Positive Autocorrelation	Neither accept or reject	Accept the Null Hypothesis	Neither accept or reject	Reject Null Hypothesis: Negative Autocorrelation

When $n=43$ and the number of explanatory variables in the regression is 3, from tables we have approximate values for d_1 and d_u : $d_1 = 1.39$; $d_u = 1.65$. As our value, 1.46, is in between d_1 and d_u , we neither accept or reject the Null Hypothesis. At least from this we do not conclude that autocorrelation is *present* in our model, and so we do not have to worry about incorporating its effects in the regression model. In this case, we conclude that factors such as global warming do not affect our data and our regression model.

Interestingly, if we had removed influential observations from our regression model, then we would have obtained Durbin-Watson values of less than 1.39. This would have led to the problem of having to deal with autocorrelation in the regression model. This is therefore another reason why not to remove the influential or unusual observations from our regression model. We want to obtain as accurate and as **simple** a model as we can

(*Aside:* Another possible way to test for autocorrelation would be to break the data into six groups and perform a one-way Anova test with BSAAM as the response and the groups as the factor. Doing this (see Appendix 2) gives us an F-Value which is not significant at the 30% level. So we again discount autocorrelation as a possible influence.)

Section 5: Prediction

Now that our regression model has passed all of the tests performed on it, we can now use it for the purpose it was designed for - predicting *stream runoff* from *precipitation measurements*. So, given a precipitation measurement from the tipping-bucket at South Lake (APSLAKE); a precipitation measurement from the standard gage at Rock Creek (OPRC); and a precipitation measurement from the standard gage at South Lake (OPSLAKE), we can now predict the stream runoff near Bishop, California, using the following equation:

$$\text{Stream Runoff} = 15425 + (1712 \times \text{APSLAKE}) + (1797 \times \text{OPRC}) + (2390 \times \text{OPSLAKE})$$

For example, let us suppose that the 1991 precipitation measurements were as follows: APSLAKE = 4.93, OPRC = 12.00, OPSLAKE = 13.52. Using Minitab to calculate the fitted value for our new data, specifying a 95% Confidence Interval, gives us the following output

Predicted Values for New Observations

New Obs	Fit	SE Fit	95.0% CI	95.0% PI
1	77747	1111	(75501, 79994)	(62845, 92650)

The above interprets as saying that for the precipitation measurements obtained, we are 95% confident that the stream runoff will lie in the interval (75501, 79994). Note that this interval is based on predicting the mean response μ of all individuals with these values of the regressor variables.

As always, the model is best used for collected data which is similar to data that has already been collected. It would be dangerous to use the equation for values outside of the maximum-minimum values shown in the basic descriptive statistics on page 6, as we have no evidence that the equation is valid for these values.

Closing our analysis, we now *know* that there is a definite link between precipitation and stream runoff, and have found a linear equation to calculate this relationship. We must remember, however, that this equation is only valid for values collected over a period of one year. To derive some practical use for the relationship, we need an equation that will be valid for smaller periods of time, perhaps days or weeks.

That would be the next step if we were to carry the work on this project forward - to calculate equations that can be used to extend the prediction range of the river forecasts the relationship is intended for use with. Only then would we begin to see the effects and advantages of knowing about the relationship.

To do this, we would use similar techniques to what we have been using here, and expect to get equations with smaller coefficients. We would also expect to get a more accurate model - the smaller time intervals involved will yield bigger volumes of data.

Conclusions

The goal of this project was to prove the existence of a relationship between stream runoff and precipitation. Using multiple linear regression, we have used data specific to an area in California, USA, to prove that a relationship **does** exist for this particular area. The relationship is given by the equation

$$\text{Stream Runoff} = 15425 + (1712 \times \text{APSLAKE}) + (1797 \times \text{OPRC}) + (2390 \times \text{OPSLAKE})$$

The above expression fits the data well, has significant predictors, and residuals which satisfy all assumptions (independence, normality, constant variance). The negative sides of the model, such as influential observations, are unavoidable.

All in all, we have provided a basis for predicting stream runoff using just three separate precipitation measurements. This provides a basis for which further work can now go ahead in producing equations valid for smaller time intervals, which can then be used to extend the time intervals of river forecasts. The work in this project has shown that such work is definitely feasible.

Bibliography

1. Krzanowski, W. J. (1998)
An Introduction to Statistical Modelling, 1st ed., Arnold Publishers
2. McCloskey, M.; Blythe, S.; Robertson, C. (1997)
Quercus: Statistics for Bioscientists: A Student Guidebook, 1st ed., Arnold Publishers.
3. (2000) Minitab Help File and StatGuide, Release 13.1
4. The Internet Sites listed on page 3

References

Page 2	Picture taken from http://www.rain.org/~phoheck/BigPineHike/BPine02.jpg
Page 4	Methods of measuring stream runoff & precipitation taken from http://info.abrfc.noaa.gov/fcstmethods.html .
	Map taken from <u>Microsoft Encarta 1997</u>
Page 5	Picture taken from http://www.gravityhome.com/jk/pix/sierra97/sabrina/1-700.html
	Map taken from http://www.395.com
Page 8	Interpretation of Variance Inflation Factors taken from the <u>Minitab Help File</u>
Page 10	Krzanowski, W. J., <u>An Introduction to Statistical Modelling</u> , 1st ed., Arnold Publishers pp. 93-99 (Variable Selection: examining all subsets)
Page 13	Interpretation of Cook's Distance taken from the <u>Minitab Help File</u>
Page 14	Krzanowski, W. J., <u>An Introduction to Statistical Modelling</u> , 1st ed., Arnold Publishers pp. 102-103 (Model Assumptions)
Page 15	Information on the Durbin-Watson statistic taken from http://www.csus.edu/indiv/j/jensena/mgmt105/durbin.htm
Page 16	Krzanowski, W. J., <u>An Introduction to Statistical Modelling</u> , 1st ed., Arnold Publishers pp. 93-99 (Prediction)

Appendix 1: The Data

Year	APMAM	APSAB	APSLAKE	OPBPC	OPRC	OPSLAKE	BSAAM	Group	SRES1	FITS1	COOK1
1948	9.13	3.58	3.91	4.1	7.43	6.47	54235	1	0.46618	50938	0.003285
1949	5.28	4.82	5.2	7.55	11.11	10.26	67567	1	-0.17573	68819	0.000348
1950	4.2	3.77	3.67	9.52	12.2	11.35	66161	1	-0.64912	70763	0.005835
1951	4.6	4.46	3.93	11.14	15.15	11.13	68094	1	-1.19426	75985	0.076690
1952	7.15	4.99	4.88	16.34	20.05	22.81	107080	1	-1.04112	114333	0.025234
1953	9.7	5.65	4.91	8.88	8.15	7.41	67594	1	1.60394	56191	0.031890
1954	5.02	1.45	1.77	13.57	12.45	13.32	65356	1	-1.04257	72667	0.021430
1955	6.7	7.44	6.51	9.28	9.65	9.8	67909	1	0.08014	67339	0.000077
1956	10.5	5.85	3.38	21.2	18.55	17.42	92715	2	-0.51172	96187	0.009986
1957	9.1	6.13	4.08	9.55	9.2	8.25	70024	2	1.59856	58664	0.032290
1958	8.75	5.23	5.9	15.25	14.8	17.48	99216	2	0.74254	93905	0.005104
1959	8.1	3.77	4.56	9.05	6.85	9.56	55786	2	-0.36999	58393	0.002345
1960	3.75	1.47	1.78	4.57	6.1	7.65	46153	2	-0.22659	47720	0.001410
1961	10.15	5.09	4.86	8.9	7.15	9	47947	2	-1.43157	58108	0.027199
1962	6.15	3.52	3.3	16.9	14.75	17.68	76877	2	-1.82470	89841	0.042465
1963	12.75	8.17	10.16	16.75	11.55	15.53	88443	3	-0.34134	90698	0.006267
1964	7.35	4.33	4.85	5.25	7.45	8.2	54634	3	-0.29245	56718	0.000957
1965	11.25	6.56	7.6	8.4	13.2	13.29	78806	3	-0.72793	83927	0.009528
1966	4.05	1.9	2	10.85	8.25	12.56	56542	3	-1.04497	63695	0.036084
1967	12.65	6.62	7.14	23.25	17	23.66	116244	3	0.22016	114752	0.001883
1968	4.65	3.84	3.34	7.1	6.8	8.28	60857	3	1.09062	53155	0.018940
1969	5.35	3.62	4.62	43.37	24.85	33.07	146345	3	-0.11026	147035	0.001075
1970	4.05	1.98	2.94	8.95	11.25	11	73726	4	0.95273	66969	0.012405
1971	5.9	5.72	5.42	8.45	10.9	10.82	65530	4	-0.64610	70157	0.003606
1972	9.45	4.82	6.79	7.9	7.6	8.06	60772	4	0.11297	59975	0.000212
1973	3.45	2.63	2.88	14.8	14.7	15.86	91696	4	0.99056	84682	0.014251
1974	4.25	2.54	2.36	18.05	16.9	16.42	87377	4	-0.24881	89084	0.001960
1975	7.9	4.42	6.78	11.5	9.55	12.56	77306	4	0.43750	74217	0.003083
1976	9.38	8.3	9.7	6.8	5.25	4.73	44756	4	-1.22507	52776	0.089198
1977	7.08	4.4	3.9	4.05	4.35	4.6	41785	5	0.12449	40916	0.000339
1978	11.92	5.78	6.7	25.3	20.55	21.94	112653	5	-0.52629	116269	0.008563
1979	3.88	2.26	3.1	15.97	11.83	13.88	79975	5	0.67362	75168	0.004751
1980	5.8	3.1	3.34	24.4	19.15	23.78	106821	5	-0.80823	112396	0.018772
1981	2.7	2.22	2.48	8.99	9.45	12.14	69177	5	0.49787	65670	0.004293
1982	18.08	11.96	13.02	18.55	18.4	19.45	120463	5	0.54576	117277	0.041441
1983	8.2	4.98	5.76	19.25	22.9	23.86	135043	5	1.73015	123472	0.139241
1984	7.65	5.3	5.74	14.45	13.15	14.42	102001	6	2.59692	83352	0.048465
1985	5.22	4.42	4.04	11.45	10.16	13.06	77790	6	0.83717	71817	0.007355
1986	4.93	3.26	4.58	26.47	15.33	26.46	118144	6	0.69269	114058	0.062934
1987	5.99	2.76	3.98	4.8	6.85	6.36	61229	6	1.62138	49752	0.038652
1988	6.83	6.82	5.18	7.2	9.01	9.88	58942	6	-0.72025	64102	0.004357
1989	8.8	5.06	4.92	8.05	9.6	9.58	53965	6	-1.40159	64000	0.017099
1990	7.1	5.06	6.05	5.8	6.5	8.41	49774	6	-1.10778	57567	0.022067

Key:

APMAM:	Precipitation Measurement (Tipping-Bucket) in inches taken at <i>Mammoth Lakes</i>
APSAB:	Precipitation Measurement (Tipping-Bucket) in inches taken at <i>Lake Sabrina</i>
APSLAKE:	Precipitation Measurement (Tipping-Bucket) in inches taken at <i>South Lake</i>
OPBPC:	Precipitation Measurement (Standard-Gage) in inches taken at <i>Big Pine Creek</i>
OPRC:	Precipitation Measurement (Standard-Gage) in inches taken at <i>Rock Creek</i>
OPSLAKE:	Precipitation Measurement (Standard-Gage) in inches taken at <i>South Lake</i>
BSAAM:	Stream Runoff Measurement in acre-feet taken near <i>Bishop</i>
Group:	Used in the one-way Anova test for Autocorrelation
SRES1:	<i>Standardised residual</i> for each observation in the Regression Model
FITS1:	<i>Fitted value</i> for each observation in the Regression Model
HI1:	<i>Cook's Distance</i> for each observation in the Regression Model

Note: The **green** columns are used to denote the explanatory variables that are used in the *optimal regression model*, while the **purple** column denotes the response variable.

Appendix 2: Minitab Output

1. Basic Descriptive Statistics.

Descriptive Statistics: Year, APMAM, APSAB, APSLAKE, OPBPC, OPRC, OPSLAKE, BSAAM

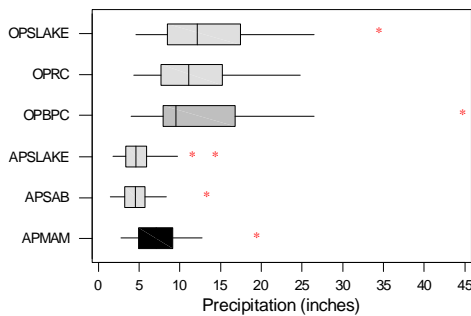
Variable	N	Mean	Median	TrMean	StDev	SE Mean
Year	43	1969.0	1969.0	1969.0	12.6	1.9
APMAM	43	7.323	7.080	7.126	3.098	0.472
APSAB	43	4.652	4.460	4.535	2.052	0.313
APSLAKE	43	4.930	4.620	4.751	2.258	0.344
OPBPC	43	12.84	9.55	12.15	7.69	1.17
OPRC	43	12.002	11.110	11.762	5.028	0.767
OPSLAKE	43	13.522	12.140	13.143	6.382	0.973
BSAAM	43	77756	69177	76297	25519	3892

Variable	Minimum	Maximum	Q1	Q3
Year	1948.0	1990.0	1958.0	1980.0
APMAM	2.700	18.080	4.930	9.130
APSAB	1.450	11.960	3.260	5.720
APSLAKE	1.770	13.020	3.340	5.900
OPBPC	4.05	43.37	7.90	16.75
OPRC	4.350	24.850	7.600	15.150
OPSLAKE	4.600	33.070	8.410	17.420
BSAAM	41785	146345	58942	92715

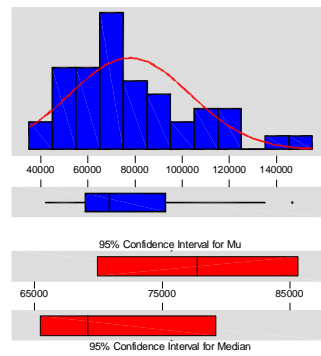
2. Graphs of the data

Boxplot for the data from the six weather stations

(Outliers shown in Red)



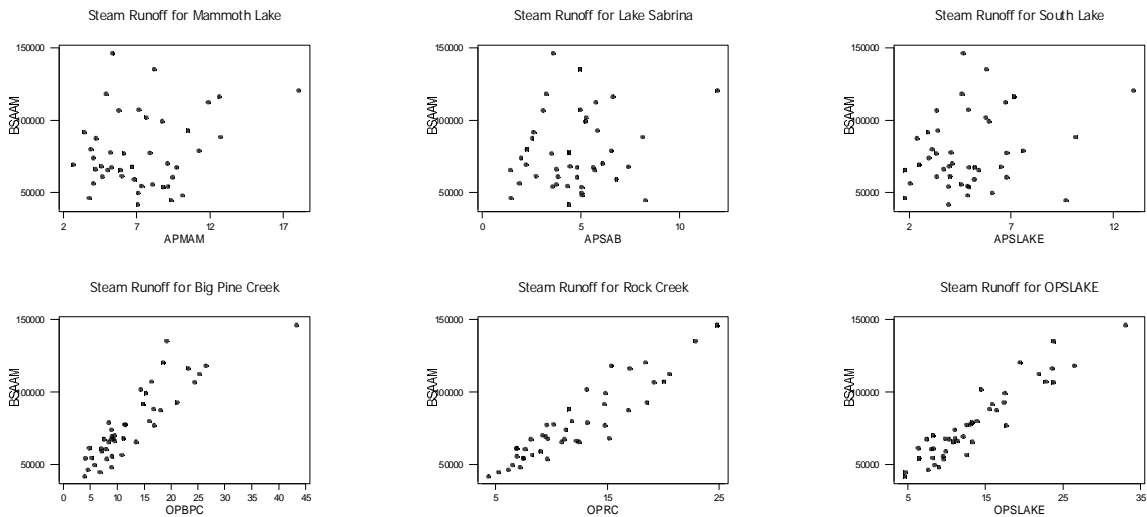
Descriptive Statistics



Variable: BSAAM

Anderson-Darling Normality Test	
A-Squared:	1.073
P-Value:	0.007
Mean:	77756.0
StDev:	25518.9
Variance:	6.51E+08
Skewness:	0.876467
Kurtosis:	0.126515
N:	43
Minimum:	41785
1st Quartile:	58942
Median:	69177
3rd Quartile:	92715
Maximum:	146345
95% Confidence Interval for Mu:	6902 85610
95% Confidence Interval for Sigma:	21041 32435
95% Confidence Interval for Median:	65475 79177

3. Visual Assessment of Regression



4. Regression output using all six explanatory variables.

Regression Analysis: BSAAM versus APMAM, APSAB, ...

The regression equation is

$$\text{BSAAM} = 15945 - 13 \text{ APMAM} - 664 \text{ APSAB} + 2271 \text{ APSLAKE} + 70 \text{ OPBPC} + 1916 \text{ OPRC} + 2212 \text{ OPSLAKE}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	15945	4100	3.89	0.000	
APMAM	-12.8	708.9	-0.02	0.986	3.5
APSAB	-664	1523	-0.44	0.665	7.2
APSLAKE	2271	1341	1.69	0.099	6.7
OPBPC	69.7	461.7	0.15	0.881	9.3
OPRC	1916.5	641.4	2.99	0.005	7.6
OPSLAKE	2211.6	752.7	2.94	0.006	17.0

S = 7557 R-Sq = 92.5% R-Sq(adj) = 91.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	6	25295187601	4215864600	73.82	0.000
Residual Error	36	2055830733	57106409		
Total	42	27351018334			

Source	DF	Seq SS
APMAM	1	1556694820
APSAB	1	17427382
APSLAKE	1	661257428
OPBPC	1	19990640428
OPRC	1	2576154606
OPSLAKE	1	493012936

Unusual Observations

Obs	APMAM	BSAAM	Fit	SE Fit	Residual	St Resid
22	5.4	146345	147746	5634	-1401	-0.28 X
37	7.7	102001	83459	1381	18542	2.50R

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

5. Correlation Output

Correlations: APMAM, APSAB, APSLAKE, OPBPC, OPRC, OPSLAKE

	APMAM	APSAB	APSLAKE	OPBPC	OPRC
APSAB	0.828 0.000				
APSLAKE	0.816 0.000	0.900 0.000			
OPBPC	0.122 0.434	0.040 0.801	0.093 0.551		
OPRC	0.154 0.323	0.106 0.500	0.106 0.497	0.865 0.000	
OPSLAKE	0.108 0.492	0.030 0.850	0.101 0.521	0.943 0.000	0.919 0.000

Cell Contents: Pearson correlation
 P-Value

6. Best Subsets Regression

Best Subsets Regression: BSAAM versus APMAM, APSAB, ...

Response is BSAAM

Vars	R-Sq	R-Sq(adj)	C-p	S	A O P P A A S O S P P L P O L M S A B P A A A K P R K M B E C C E							
1	88.1	87.8	18.2	8922.4								X
1	84.6	84.2	34.9	10145								X
2	90.5	90.0	8.5	8063.3				X				X
2	90.5	90.0	8.6	8065.4				X				X
3	92.4	91.9	1.2	7283.5				X		X	X	
3	91.8	91.2	4.1	7566.1				X		X	X	
4	92.5	91.7	3.0	7357.7				X	X	X	X	
4	92.4	91.6	3.2	7376.2				X	X	X	X	
5	92.5	91.5	5.0	7454.1				X	X	X	X	X
5	92.5	91.5	5.0	7456.4				X	X	X	X	X
6	92.5	91.2	7.0	7556.9				X	X	X	X	X

7. Stepwise Regression (Normal; Forward Selection; Backward Elimination)

Stepwise Regression: BSAAM versus APMAM, APSAB, ...

Alpha-to-Enter: 0.15
Alpha-to-Remove: 0.15

Forward selection.
Alpha-to-Enter: 0.25

Backward elimination.
Alpha-to-Remove: 0.1

Response is BSAAM
on 6 predictors, with N = 43

Response is BSAAM
on 6 predictors, with N = 43

Response is BSAAM
on 6 predictors, with N = 43

Step	1	2	3	Step	1	2	3	Step	1	2	3	4	
Constant	27015	19145	15425	Constant	27015	19145	15425	Constant	15945	15931	15750	15425	
OPSLAKE	3752	3690	2390	OPSLAKE	3752	3690	2390	APMAM		-13			
T-Value	17.39	18.83	5.35	T-Value	17.39	18.83	5.35	T-Value		-0.02			
P-Value	0.000	0.000	0.000	P-Value	0.000	0.000	0.000	P-Value		0.986			
APSLAKE		1769	1712	APSLAKE		1769	1712	APSAB		-664	-673	-651	
T-Value		3.19	3.42	T-Value		3.19	3.42	T-Value		-0.44	-0.47	-0.47	
P-Value		0.003	0.001	P-Value		0.003	0.001	P-Value		0.665	0.638	0.643	
OPRC			1797	OPRC			1797	APSLAKE		2271	2264	2245	1712
T-Value			3.17	T-Value			3.17	T-Value		1.69	1.78	1.80	3.42
P-Value			0.003	P-Value			0.003	P-Value		0.099	0.083	0.080	0.001
S	8922	8063	7284	S	8922	8063	7284	OPBPC		70	69		
R-Sq	88.07	90.49	92.44	R-Sq	88.07	90.49	92.44	T-Value		0.15	0.15		
R-Sq(adj)	87.78	90.02	91.85	R-Sq(adj)	87.78	90.02	91.85	P-Value		0.881	0.880		
C-p	18.2	8.5	1.2	C-p	18.2	8.5	1.2	OPRC		1916	1916	1910	1797
								T-Value		2.99	3.03	3.07	3.17
								P-Value		0.005	0.004	0.004	0.003
								OPSLAKE		2212	2213	2295	2390
								T-Value		2.94	2.99	4.64	5.35
								P-Value		0.006	0.005	0.000	0.000
								S		7557	7454	7358	7284
								R-Sq		92.48	92.48	92.48	92.44
								R-Sq(adj)		91.23	91.47	91.69	91.85
								C-p		7.0	5.0	3.0	1.2

8. Regression output using the optimal model.

Regression Analysis: BSAAM versus APSLAKE, OPRC, OPSLAKE

The regression equation is

$$\text{BSAAM} = 15425 + 1712 \text{ APSLAKE} + 1797 \text{ OPRC} + 2390 \text{ OPSLAKE}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	15425	3638	4.24	0.000	
APSLAKE	1712.5	500.5	3.42	0.001	1.0
OPRC	1797.5	567.8	3.17	0.003	6.5
OPSLAKE	2389.8	447.1	5.35	0.000	6.4

S = 7284 R-Sq = 92.4% R-Sq(adj) = 91.9%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	25282070749	8427356916	158.86	0.000
Residual Error	39	2068947585	53049938		
Total	42	27351018334			

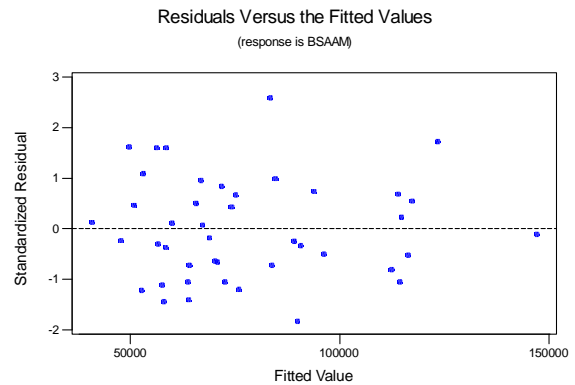
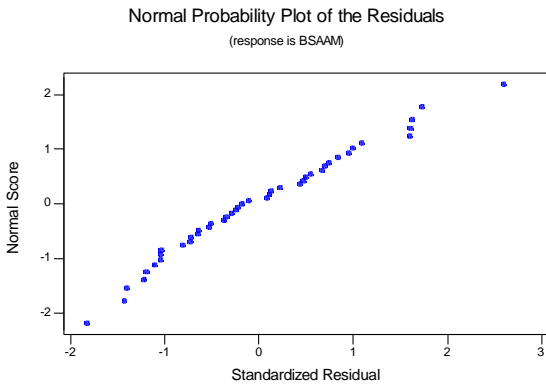
Source	DF	Seq SS
APSLAKE	1	1700437520
OPRC	1	22065714689
OPSLAKE	1	1515918540

Unusual Observations

Obs	APSLAKE	BSAAM	Fit	SE Fit	Residual	St Resid
35	13.0	120463	117277	4355	3186	0.55 X
37	5.7	102001	83352	1218	18649	2.60R
39	4.6	118144	114058	4273	4086	0.69 X

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.46



9. Attempting to Remove Unusual Observations

Regression Analysis: BSAAM versus APSLAKE, OPRC, OPSLAKE (Having Removed Observations from 1982 and 1986)

The regression equation is

$$\text{BSAAM} = 16336 + 1556 \text{ APSLAKE} + 1965 \text{ OPRC} + 2213 \text{ OPSLAKE}$$

Predictor	Coef	SE Coef	T	P	VIF
Constant	16336	4193	3.90	0.000	
APSLAKE	1556.4	613.4	2.54	0.016	1.0
OPRC	1964.5	677.8	2.90	0.006	8.4
OPSLAKE	2212.7	555.9	3.98	0.000	8.4

S = 7406 R-Sq = 91.4% R-Sq(adj) = 90.8%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	21698078524	7232692841	131.86	0.000
Residual Error	37	2029460279	54850278		
Total	40	23727538803			

Source	DF	Seq SS
APSLAKE	1	447055272
OPRC	1	20381913305
OPSLAKE	1	869109947

Unusual Observations

Obs	APSLAKE	BSAAM	Fit	SE Fit	Residual	St Resid
22	4.6	146345	145519	4303	826	0.14 X
36	5.7	102001	83010	1360	18991	2.61R

R denotes an observation with a large standardized residual
 X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.43

**Regression Analysis: BSAAM versus APSLAKE, OPRC, OPSLAKE
 (Having removed observations from 1982, 1986 and 1969)**

The regression equation is
 BSAAM = 16459 + 1561 APSLAKE + 1989 OPRC + 2177 OPSLAKE

Predictor	Coef	SE Coef	T	P	VIF
Constant	16459	4347	3.79	0.001	
APSLAKE	1561.1	622.7	2.51	0.017	1.0
OPRC	1988.9	710.2	2.80	0.008	7.4
OPSLAKE	2177.3	621.3	3.50	0.001	7.5

S = 7506 R-Sq = 89.1% R-Sq(adj) = 88.2%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	16587873657	5529291219	98.13	0.000
Residual Error	36	2028429635	56345268		
Total	39	18616303292			

Source	DF	Seq SS
APSLAKE	1	478338947
OPRC	1	15417506333
OPSLAKE	1	692028377

Unusual Observations

Obs	APSLAKE	BSAAM	Fit	SE Fit	Residual	St Resid
35	5.7	102001	82970	1410	19031	2.58R

R denotes an observation with a large standardized residual

Durbin-Watson statistic = 1.38

10. Autocorrelation

One-way ANOVA: BSAAM versus Group

Analysis of Variance for BSAAM

Source	DF	SS	MS	F	P
Group	5	3.787E+09	757406004	1.19	0.333
Error	37	2.356E+10	636864549		
Total	42	2.735E+10			

Individual 95% CIs For Mean
 Based on Pooled StDev

Level	N	Mean	StDev	CI
1	8	70500	15480	(-----*-----)
2	7	69817	21109	(-----*-----)
3	7	85982	34394	(-----*-----)
4	7	71595	16148	(-----*-----)
5	7	95131	32730	(-----*-----)
6	7	74549	26214	(-----*-----)

Pooled StDev = 25236 60000 80000 100000 120000