Problem 6: Correlation and regression

It is often useful to look for significant relationships among variables in a dataset. Does TP concentration vary with flow? Are SS and TP concentrations related? Such questions are usually addressed by determining if there are correlations between variables and by testing for significant linear relationships with regression. Linear regression quantitatively describes the relationship in a way that can be used to predict values of one variable from the other.

a. Correlations among variables

Using Dataset 1 in file Sampledata.xlsx, evaluate significant correlations among flow (Q_2), TP concentration (TP_2), and SS concentration (SS_2) at Station 2 across all periods.

The following correlation matrix is generated by applying the parametric correlation r statistic (where $\pm$ 1.0 indicates a perfect correlation and 0.0 represents no correlation) to the log-transformed data:

	log Q_2	log TP_2	log SS_2
log Q_2	**1.000**	0.008	0.026
log TP_2	0.008	**1.000**	**0.782**
log SS_2	0.026	**0.782**	**1.000**

Values of r in bold are statistically significant at $P \leq 0.05$

The above table indicates that TP (log TP_2) and SS (log SS_2) concentrations are significantly correlated (r = 0.782) and that the correlation is positive, i.e., higher TP concentrations are associated with higher SS concentrations. There were no significant correlations between flow (log Q_2) and either TP or SS concentrations.

The following result is obtained by applying the nonparametric Spearman's rho (*p*) (where $\pm$ 1.0 indicates a perfect correlation and 0.0 represents no correlation) to the raw data:

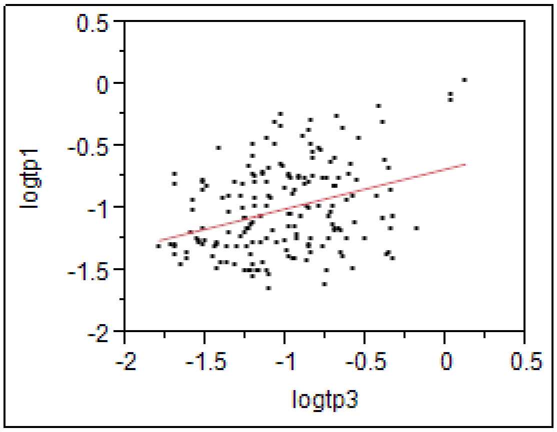
Variable	By variable	Spearman's *p*	*P* value
TP_2	Q_2	-0.059	0.286
SS_2	Q_2	-0.032	0.557
SS_2	TP_2	**0.821**	**<0.001**

The above table indicates a significant positive correlation between TP and SS concentrations at Station 2, but no significant correlation between flow and either TP or SS.

b. Test for regression relationship between the same variable at two sites

In Dataset 1, Station 3 represents the control watershed, while Stations 1 and 2 represent two treatment watersheds. Using Dataset 1, and assuming that all data satisfy the requirements for parametric statistics using a log transformation, determine if significant regression relationships exist between TP measured at Station 3 and TP measured at each of the other stations during the Calibration Period (Treatment=CAL).

TP_1 vs. TP_3



Summary of Fit

RSquare	0.115583
Root Mean Square Error	0.336782
Observations (or Sum Wgts)	181

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	2.653319	2.65332	23.3933
Error	179	20.302585	0.11342	**Prob > F**
C. Total	180	22.955905		<.0001*

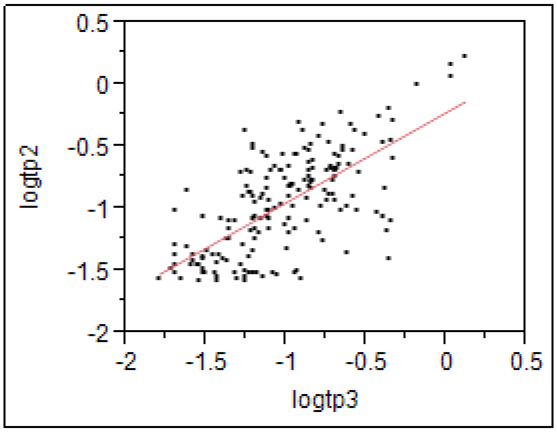
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.6767	0.071772	-9.43	<.0001*
logtp3	0.3200551	0.066173	4.84	<.0001*

The regression statistics (F ratio and associated *P* value) indicate that there is a significant relationship ($P \leq 0.001$) between logTP_3 and logTP_1, although the relationship is relatively weak because the regression model, logTP_3 explains only ~11% (RSquare = 0.115583) of the variation in logTP_1. Both the slope and intercept are significantly different from zero ($P \leq$ 0.001). The regression equation is:

$$logTP\_1 = -0.6767(logTP\_3) + 0.3200$$

TP_2 vs. TP_3



Summary of Fit

RSquare	0.464501
Root Mean Square Error	0.296943
Observations (or Sum Wgts)	181

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	13.690687	13.6907	155.2674
Error	179	15.783309	0.0882	**Prob > F**
C. Total	180	29.473997		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-0.220853	0.063282	-3.49	0.0006*
logtp3	0.7270138	0.058345	12.46	<.0001*

The regression statistics (F ratio and associated *P* value) indicate that there is a significant relationship ($P \leq 0.001$) between logTP_3 and logTP_2, one that may be more meaningful than that from Station 1, as the regression model explains 46% of the variation in logTP_2. Both the slope and intercept are significantly different from zero ($P \leq 0.001$). The regression equation is:

logTP_2 = -0.2208(logTP_3) + 0.7270