

Appendix B

Statistical Hydrology

Statistics is the science of understanding uncertainty. Will it rain today? Given that it has not rained for three months, what is the probability that it might rain in the next week? How does a dam (or ground-water pumping, wetland construction, timber harvesting) affect stream-flows? What are the health risks from drinking contaminated water? These are all questions that are commonly asked.

While the goal of science is to separate fact from fiction, we are often limited to providing statistical measures of truth and error. Thus science often rests on the edge of certainty, not entirely sure, yet nor entirely unsure. Many of the early studies in statistics were performed by compulsive gamblers who wished to improve their odds of winning. Rather than accept the “roll of the dice”, these individuals wanted to better understand the risks they were taking, and place their bets in ways that maximized their likelihood of winning.

B.1 Probability

Probability analysis is used to describe random behavior, such as the *chance* that an event will occur, or the *likelihood* that an event will exceed a certain magnitude. While much of nature is not entirely random, we can often apply probability models to natural systems. We can make these applications more readily in cases where:

- Events are independent of each other, and do not affect each other. That is, the result of one coin toss does not affect the following coin toss.
- Events are stationary - they are not a function of time. That is, heads are not more likely in the morning than in the evening.
- Events are identically distributed. That is, the variability of heads is the same under all conditions.

If these assumptions are satisfied, then we can say that the likelihood of either one event or another occurring is just the sum of the individual events:

$$P(A \cup B) = P(A \text{ or } B) = P(A) + P(B) \quad (\text{B.1})$$

We can also say that the probability of two events, A and B , occurring together is just the product of the probability of each event:

$$P(A \cap B) = P(A \text{ and } B) = P(A) \cdot P(B) \quad (\text{B.2})$$

For example, let us assume that the probability of landing either a heads (H) or a tails (T) when a coin is flipped are equal, so that there are two outcomes, with the probability of obtaining one or the other being $p = 0.5$.

When the coin is tossed twice, $n = 2$, there are four outcomes; $H \cap H$, $T \cap T$, $H \cap T$, and $T \cap H$. Each outcome has an equal probability because these are independent events. The probability of two heads in a row is:

$$P(H \cap H) = p^2 = 0.25 \quad (\text{B.3})$$

which is the same for landing two tails. The probability of landing one of each has two outcomes, so that:

$$\begin{aligned} P((T \cap H) \cup (H \cap T)) &= P(T \cap H) + P(H \cap T) \\ &= 2 p^2 = 0.5 \end{aligned} \quad (\text{B.4})$$

We can write this mathematically for any number of tosses, n , to determine the number of heads, m , and tails, $n - m$

$$\begin{aligned} P(H = m, T = n - m) &= \binom{n}{m} P(H)^m P(T)^{n-m} \\ &= \binom{n}{m} p^n \end{aligned} \quad (\text{B.5})$$

where

$$\binom{n}{m} = \frac{n!}{m! (n - m)!} \quad (\text{B.6})$$

is the combinatorial operator that accounts for the number of opportunities for getting the same outcome, and where $n! = n \times (n - 1) \times (n - 2) \times \dots \times 1$ is the factorial of n .

For example, we can calculate the probability of getting exactly 5 heads and 5 tails in ten tosses:

$$P(H = 5, T = 5) = \binom{10}{5} \left(\frac{1}{2}\right)^5 \left(\frac{1}{2}\right)^5$$

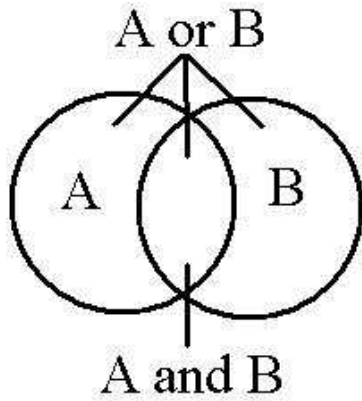


Figure B.1: Venn diagram of events A and B

$$= \binom{10}{5} \left(\frac{1}{2}\right)^{10} = 0.246 \quad (\text{B.7})$$

which means that we have a chance of only about 1 in 4 of getting an equal number of heads and tails.

If events are not independent of each other, then we can still calculate their probability using:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{B.8})$$

For example, if the probability of a rainy day is thirty percent, $P(A) = 0.3$, the probability of snow is ten percent, $P(B) = 0.1$, and the probability of getting both rain and snow in a day is five percent, $P(A \cap B) = 0.05$, then the probability of getting either rain or snow in a day is:

$$P(A \cup B) = 0.3 + 0.1 - 0.05 = 0.35 \quad (\text{B.9})$$

Frequency vs. probability. The frequency of an outcome is the *observed* number of occurrences based on a finite sample size, $f_i = n_i/N$, while a probability is the *likelihood* of that outcome based on an infinite sample size, $p_i = \lim_{N \rightarrow \infty} n_i/N$. We might say that the probability is the *predicted* frequency based on complete information.

Conditional probability. Conditional probabilities arise when an event, A , may be more (or less) likely given that another event, B , has happened. In this case:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (\text{B.10})$$

where $P(A|B)$ means the probability of Event A given that Event B has already occurred.

For example, if the probability of rain and snow are again 30 and 10 percent, respectively, and the probability of snow given that rainfall has occurred is 20 percent, then the probability of rain given that snow has occurred is:

$$P(A|B) = \left(\frac{0.3}{0.1}\right) 0.2 = 0.6 \quad (\text{B.11})$$

or sixty percent.

Problems

- For a daily rainfall probability of 30 percent, find the probability that it will:
 - Rain three days in a row.
 - Not rain for seven days in a row.
 - Rain three days in a week.
 - Rain at least 1 day in a week.
 - Rain greater than one inch if the probability is 25 percent if it rains.

Digression: Bayes' Theorem

Bayes' theorem can be used to predict events by incorporating the relationship between events. For example, if Joe and Susan are frequently found together, then if you see Susan, you have a high likelihood of seeing Joe.

The theorem is derived by noting that:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (\text{B.12})$$

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad (\text{B.13})$$

If A and B are independent events, then $P(A \cap B) = P(A) \cdot P(B)$, so that:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \cdot P(B)}{P(B)} = P(A) \quad (\text{B.14})$$

and

$$P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) \cdot P(B)}{P(A)} = P(B) \quad (\text{B.15})$$

which makes sense because knowing A doesn't help knowing B , and vice versa.

If, on the other hand, A and B are related to each other, then we can use information about one to help with our prediction. Combining the first two equations yields:

$$P(A \cap B) = P(B|A) P(A) = P(A|B) P(B) \quad (\text{B.16})$$

or:

$$P(A|B) = P(B|A) \frac{P(A)}{P(B)} \quad (\text{B.17})$$

Example. Say we have a relationship between temperature and snowfall. Let $P(A)$ represent the probability that temperature is below freezing and $P(B)$ represent the snowfall probability, so that:

- $P(A) = 20\%$, probability of cold weather
- $P(B) = 5\%$, probability of snow
- $P(A|B) = 100\%$, probability that it's cold when it's snowing
- $P(B|A) = P(A|B) \times P(B) / P(A) = 1 \times 0.05 / 0.20 = 25\%$, is the probability of snow when it's cold outside.

B.2 Statistics

Expected Value. We are often asked what the outcome of an uncertain event is likely to be, such as today's expected high temperature. The *expected value* – equivalent to the mean or average – is calculated using:

$$\begin{aligned}\bar{x} &= E(x) = \int_{-\infty}^{\infty} x f(x) dx \\ &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}\quad (\text{B.18})$$

where n is the number of observations, and $f(x)$ is the frequency distribution of x . The frequency distribution represents the likelihood of individual observation. Just as each observation is weighted by $1/n$, the distribution of individual observations will have a weight associated with them that corresponds to their frequency.

Variance. The *variance* of individual observations (when the expected value is known, μ_x) is calculated using:

$$V(x) = \sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x) dx$$

where σ_x is the *standard deviation* of x . A case with a known mean is when a coin is tossed - the mean number of heads is always $\mu_x = 1/2$. The variance is sometimes referred to as the second-moment about the mean, $V(x) = M_2(x)$.

If the mean is unknown, then:

$$V(x) = \tilde{x}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{B.19})$$

Coefficient of Variation. The *coefficient of variation* is the ratio of the standard deviation to the mean:

$$CV = \frac{\tilde{x}}{\bar{x}} \quad (\text{B.20})$$

An example would be the variation in height with age - is there more variation as people grow taller? Plotting the coefficient of variation would indicate how much variation there is as people age. Can you guess which age has the greatest variation?

A hydrologic example of the coefficient of variation is streamflow; low-flow variability is probably much smaller than under flood conditions, but their coefficient of variation may be similar.

Covariance. The *covariance* between two variables is found using:

$$C(x, y) = \iint_{-\infty}^{\infty} (x - \mu_x) (y - \mu_y) f(x, y) dx dy$$

$$= \frac{1}{n} \sum_{i=1}^n (x_i - \mu_x) (y_i - \mu_y) \quad (\text{B.21})$$

$$= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}) (y_i - \bar{y}) \quad (\text{B.22})$$

where $f(x, y)$ is the joint frequency distribution between x and y .

Variogram. The *variogram* describes the similarity between measurements separated by a distance, r :

$$\gamma(r) = \frac{1}{n} \sum_{i=1}^n \Delta x_i(r)^2 \quad (\text{B.23})$$

where $\Delta x_i(r)$ is the difference between two nearby observations. For example, this could be used to describe how rainfall varies spatially, where Δx is the difference in rainfall between the two points. Note that $\gamma \rightarrow 0$ as $r \rightarrow 0$, but this is not always the case, especially if there are measurement errors or large changes over small distances.

Correlation. The *correlation* between two variables is:

$$\begin{aligned}r(x, y) &= \frac{C(x, y)}{\sqrt{V(x) V(y)}} \\ &= \sum_{i=1}^n \frac{(x_i - \mu_x)}{\sigma_x} \frac{(y_i - \mu_y)}{\sigma_y}\end{aligned}\quad (\text{B.24})$$

Note that if $x = y$, then $C(x, y) = C(x, x) = V(x)$ and $r(x, y) = r(x, x) = 1$.

Standardized Variables. We can standardize the variable by subtracting the mean and dividing by the standard deviation:

$$X_i = \frac{(x_i - \mu_x)}{\sigma_x} = \frac{(x_i - \bar{x})}{\tilde{x}} \quad (\text{B.25})$$

Using this notation, the correlation is just:

$$r(x, y) = r(X, Y) = \sum_{i=1}^n X_i Y_i \quad (\text{B.26})$$

Skewness. Data is *skewed* when there are unbalanced high and low observations. For example, a stream may have an average discharge of 10 L/s, an extreme low flow of 1 L/s (9 L/s below the mean), and an extreme high flow of 100 L/s (90 L/s above the mean). This is an example of a *positive skew*, in that the larger observations are farther from the mean than the smaller observations. The skew - also called *third moment about the mean*, M_3 - is calculated using:

$$M_3(x) = \int_{-\infty}^{\infty} (x - \bar{x})^3 f(x) dx$$

$$= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^3 \tag{B.27}$$

The *skew coefficient*, $G(x)$, can also be standardized using:

$$G(x) = \frac{n}{n-2} \frac{M_3(X)}{\tilde{x}^3}$$

This adjusts the skew by how variable the data are - one needs a greater skew when there is a greater variability to arrive at the same skew coefficient.

Kurtosis. The *kurtosis*, or *fourth-moment about the mean*, M_4 , is used to describe the frequency of low-probability events - both extremely high and low:

$$M_4(x) = \int_{-\infty}^{\infty} (x - \bar{x})^4 f(x) dx$$

$$= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^4 \tag{B.28}$$

Variation of Sample Statistics. The sample statistics calculated above were for individual observations. Once we have calculated these statistics, they may still be uncertain. For example, we may have calculated the mean and variance of daily rainfall for each month of the year. Yet the rainfall in each month varies from year to year. Clearly, there is still a variation in monthly averages.

Example. Here's some student height data (cm).

	Year						
	1990	1991	1992	1993	1994	1995	1996
	180	182	172	169	158	177	173
	190	175	172	165	175	167	182
	166	175	172	173	169	156	184
	167	178	174	176	178	162	169
	174	169	187	160	188	179	158
	161	174	175	176	159	165	162
	172	170	150	175	162	186	192
	157	177	182	188	186	169	152
	189	167	175	188	181	165	183
	174	170	162	188	160	191	167
	195	169	170	162	182	174	171
	172	180	171	184	196	167	174
	169	186	163	162	159	170	161
	156	175	172	167	178	178	161
	156	169	178	186	175	177	169
	174	171	156	155	164	183	177
	169	178	172	177	184	187	173
	171	173	182	176	181	188	183
	189	170	182	168	187	178	161
	179	182	165	178	183	178	196
	176	177	168	174	182	183	184
	173	154	179	168	177	179	190
	152	183	181	167	180	179	181
	181	171	179	182	182	172	177
	163	166	173	152	199	178	180
\bar{x}	172.1	173.7	172.5	172.7	177.0	175.6	174.4
\tilde{x}	11.3	6.7	8.6	10.1	11.4	8.8	11.3

- $\bar{x} = 174.0$
- $\tilde{x} = 1.78$
- $\tilde{\tilde{x}} = 9.7$
- $\tilde{\tilde{x}}/\sqrt{n} = 1.94$

For groups of events, the mean of the group, $E(\bar{x})$ should equal the mean of individual events within the group, $E(x)$:

$$E(\bar{x}) = E(x) \tag{B.29}$$

but is not true if the calculated mean is *biased*.

Bias occurs when the average of a distribution does not converge to the true mean, usually because certain outcomes are more likely measured than others. Bias is relatively common in hydrology because we tend to sample more heavily in good weather, and avoid taking measurements when the weather is awful.

The variance of the sample mean will be different from the variance of individual events because the average behavior is commonly much less variable than the individual events.

$$V(\bar{x}) = \frac{V(x)}{n} \quad \text{and} \quad \tilde{\tilde{x}} = \frac{\tilde{x}}{\sqrt{n}} \tag{B.30}$$

The variation in the variance is described using the χ^2 (pronounced *kī*) statistic, which is mathematically equivalent to a *gamma* distribution (below), with $\beta = 2$ and

$\alpha = \nu/2$, and where ν (pronounced *nu*) are the degrees of freedom, equal to the number of observations minus the number of statistics (such as the mean and variance) that are being estimated. This is equivalent to:

$$\chi^2(x, \nu) = \frac{2^{-\nu/2} x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2)} \quad (\text{B.31})$$

Problems

1. Find and download water quality data for a minimum of five variables.
2. Calculate the means, standard deviations, standard errors of the means, coefficient of variation, skewness, kurtosis, and correlation and covariance matrices.

Digression: Taylor Series

The Taylor Series expansion of a function of one variable is defined using:

$$f(x) = f(a) + (x - a) f'(a) + \frac{(x - a)^2}{2!} f''(a) + \frac{(x - a)^3}{3!} f'''(a) + \dots + \frac{(x - a)^n}{n!} f^{(n)}(a) + \dots \quad (\text{B.32})$$

where $f(x)$ is an arbitrary function that is infinitely differentiable, x is a variable, and a is a constant.

The variance of an arbitrary function can be approximated using:

$$V[f(x)] = V[f(a)] + V[(x - a) f'(a)] \quad (\text{B.33})$$

where only the first terms are used. Note that the variance of a constant can be neglected, $V[f(a)] = 0$, and $V(x - a) = V(x)$, so that we have the general expression:

$$V[f(x)] = (f'(a))^2 V[x] \quad (\text{B.34})$$

For the case where $x = 100.57$ and $b = -99.94$, we have:

$$f(x) = \frac{x}{x + b} = \frac{100.57}{100.57 - 99.94} = 159.6 \quad (\text{B.35})$$

The derivative of this is:

$$f'(x) = \frac{1}{x + b} - \frac{x}{(x + b)^2} = \frac{b}{(x + b)^2} \quad (\text{B.36})$$

so that:

$$V[f(x)] = \left[\frac{b}{(x + b)^2} \right]^2 V[x] \quad (\text{B.37})$$

and:

$$\sigma_{f(x)} = \left| \frac{b}{(x + b)^2} \right| \sigma_x = \frac{99.94}{(0.63)^2} 0.01 = 2.5 \quad (\text{B.38})$$

where $V[x] = \sigma_x^2$ and $\sigma_x = 0.01$. The final expectation is, therefore:

$$f(x) = 159.6 \pm 2.5 \quad (\text{B.39})$$

The Taylor series for a function of two variables is:

$$f(x, y) = f(a, b) + \left[(x - a) \frac{\partial f(x, y)}{\partial x} \right]_{x=a} + (y - b) \frac{\partial f(x, y)}{\partial y} \Big|_{y=b} + \dots + \frac{1}{n!} \left[(x - a) \frac{\partial}{\partial x} + (y - b) \frac{\partial}{\partial y} \right]^n f(x, y) \Big|_{x=a, y=b} + \dots \quad (\text{B.40})$$

which, for the previous example, yields:

$$f(x, y) = \frac{x}{x + y} \quad (\text{B.41})$$

$$\frac{\partial f(x, y)}{\partial x} = \frac{y}{(x + y)^2} \quad (\text{B.42})$$

$$\frac{\partial f(x, y)}{\partial y} = -\frac{x}{(x + y)^2} \quad (\text{B.43})$$

so that:

$$V[f(x, y)] = \left[\frac{b}{(a + b)^2} \right]^2 V[x] - \left[\frac{a}{(a + b)^2} \right]^2 V[y] \quad (\text{B.44})$$

For the case where $V[x] = V[y] = \sigma_x^2$, we have:

$$V[f(x, y)] = \left[\frac{b - a}{(a + b)^2} \right]^2 \sigma_x^2 \quad (\text{B.45})$$

or:

$$\sigma_{f(x, y)} = \frac{b - a}{(a + b)^2} \sigma_x \quad (\text{B.46})$$

Which, for the above example with $x = 100.57$, $b = -99.94$, and $\sigma_x = 0.01$, yields:

$$\sigma_{f(x, y)} = \frac{100.57 + 99.94}{(100.57 - 99.94)^2} \times 0.01 = 5.05 \quad (\text{B.47})$$

so that:

$$f(x) = 159.6 \pm 5.0 \quad (\text{B.48})$$

which is twice the error range estimated above.

B.3 Distributions

Various probability distributions can be used to describe how observations vary, including:

Uniform. A two-parameter distribution (upper and lower limits) with the property:

$$p(x) = \frac{1}{b - a} \quad (\text{B.49})$$

Discrete vs. Continuous Distributions. Discrete distributions have a countable number of outcomes, while continuous distributions have a smooth (infinite) number of outcomes. Analogous to the factorial vs. the Γ function.

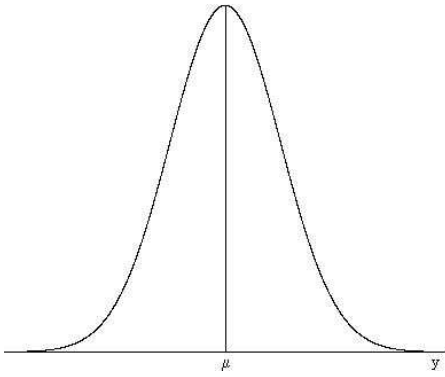


Figure B.2: Frequency plot for the normal distribution.

Bounded vs. Unbounded Distributions. Bounded distributions have a maximum and/or minimum limit (such as the uniform distribution), while unbounded distributions (such as the normal distribution) are unlimited in their range. Some distributions (such as the lognormal distribution) may be bounded on one side, and not on the other.

Normal. A continuous, two-parameter distribution (mean and standard deviation) that is widely used:

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{X^2}{2}\right) \quad (\text{B.50})$$

which is also called a *Gaussian* or bell-shaped distribution.

Exponential. A continuous, single-parameter distribution:

$$p(x) = \frac{1}{\bar{x}} \exp(-x/\bar{x}) \quad (\text{B.51})$$

Poisson. A discrete, one-parameter distribution:

$$p(x) = \frac{\bar{x}^x \exp(-\bar{x})}{x!} \quad (\text{B.52})$$

Gamma. A continuous, two-parameter distribution:

$$p(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x) \quad (\text{B.53})$$

where $\alpha = (\bar{x}/\tilde{x})^2$ and $\beta = \bar{x}/\tilde{x}^2$.

Beta. A continuous, two-parameter distribution:

$$p(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha) \cdot \Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad (\text{B.54})$$

where $\bar{x} = \alpha/(\alpha + \beta)$ and $\tilde{x}^2 = \alpha\beta/(\alpha + \beta + 1)(\alpha + \beta)^2$.

Weibull. A continuous, two-parameter distribution:

$$p(x) = \lambda^k (\lambda x)^{k-1} \exp(-(\lambda x)^k) \quad (\text{B.55})$$

where $\bar{x} = \Gamma(1 + 1/k)/\lambda$ and $\tilde{x}^2 = \Gamma(1 + 2/k)/\lambda^2 - \bar{x}^2$.

Problems

1. Plot the probability of x , $p(x)$, for a range of x using at least three of the above distributions. Identify which of the distributions are skewed.

B.4 Hypothesis Testing

One reason for estimating statistics is to be able to make a definitive statement about a situation. Did this timber harvest cause this flooding? Did this dam destroy this habitat?

To answer a question definitively, we need some process to decide whether a chance occurrence is sufficiently unlikely that we can safely say that it is improbable. For example, we might think the number of heads and tails should be approximately equal, yet we sometimes observe twenty heads in twenty tosses. We can calculate the probability as $0.5^{20} = 10^{-6}$, or one chance in a million. We might suspect that this coin is not fair.

Furthermore, we might say that any coin that has a rare outcome (say, less than one chance per thousand, 10^{-3}) should not be used. Using this rejection statistic, we recommend that this coin should be rejected as being unfair. In other words, outcomes that are beyond normal expectation are rejected, while outcomes that behave normally are not rejected.

There are two errors when this method is used. Even a fair coin has a slight chance of yielding a rare outcome. Thus, by rejecting the coin, we may be making a mistake. We call the rejection of a fair coin a *Type 1 Error*. On the other hand, an unfair coin may still give normal results and not be detected. We call the failure to reject an unfair coin a *Type 2 Error*.

For example, let's say that a river flooded following a timber cut. We observe that the flood is far worse than any observed flood, and conclude that the timber harvest caused the flood. We may then be making a Type 1 error. On the other hand, there may actually be adverse flooding from timber harvests, but the effects were too small to notice with everything else going on. This is the Type 2 error.

In statistical testing, we can assume that observations are normally distributed - the familiar bell-shaped curve. If the observation is too far from the mean, then we might think it is fundamentally different from the other observations. To check, we first find the standard normal variable:

$$z = \frac{x - \bar{x}}{\tilde{x}} \quad (\text{B.56})$$

and then use this variable to make a decision. Using the normal distribution, we can calculate the likelihood of this variable. If the probability is too small, then we might decide that it does not belong.

Problems

1. Indicate which of the coefficients in the following equation are statistically significant:

$$y = (1.09 \pm 0.05)x + (6.33 \pm 4) \quad (\text{B.57})$$

2. Can you reject the hypothesis that $y = x$?

B.5 Regression

We normally use a linear regression equation of the form:

$$\mathbf{y} = y_o + a \mathbf{x} \quad (\text{B.58})$$

where \mathbf{x} is the column vector, $(n \times 1)$, of independent observations, \mathbf{y} is the column vector, $(n \times 1)$ of dependent observations, y_o is the intercept, and a is the slope.

Surprisingly, this equation is not a linear function. A linear function, $y = f(x)$, has the property that:

$$f(2x) = 2 f(x) \quad (\text{B.59})$$

which means that if you double the input, you should double the output. To check to see if our equation, above, is linear, we have:

$$2 y = 2 (y_o + a x) = 2y_o + 2ax \neq y_o + a (2x) \quad (\text{B.60})$$

so that it is clearly *nonlinear*! To linearize this function we note that:

$$\begin{aligned} \bar{y} &= E(\mathbf{y}) = E(y_o + ax) \\ &= y_o + aE(x) = y_o + a\bar{x} \end{aligned} \quad (\text{B.61})$$

We make a linear equation by subtracting this second equation from the original equation:

$$\begin{aligned} \mathbf{Y} &= \mathbf{y} - \bar{y} = y_o + ax - (y_o + a\bar{x}) \\ &= a(\mathbf{x} - \bar{x}) = a\mathbf{X} \end{aligned} \quad (\text{B.62})$$

where the new variables, $\mathbf{Y} = \mathbf{y} - \bar{y}$ and $\mathbf{X} = \mathbf{x} - \bar{x}$, are just the original data with their means subtracted.

For multiple independent variables, then we can establish the multiple regression equation:

$$\mathbf{y} = a_o + a_1 \mathbf{x}_1 + a_2 \mathbf{x}_2 + \dots + a_m \mathbf{x}_m \quad (\text{B.63})$$

Deconvolution can be performed using this approach, such that the convolution equation:

$$y(t) = h_o x(t) + h_1 x(t-1) + \dots + h_m x(t-m) \quad (\text{B.64})$$

can be solved for the unknown values of h_i when both $x(t)$ and $y(t)$ are observed.

We normally have a large number of observations of x and y from which we wish to estimate the regression

coefficients. To do this easily we can make a set of vector equations:

$$\mathbf{Y} = \mathbf{X} \beta \quad (\text{B.65})$$

where \mathbf{Y} is a $(n \times 1)$ column vector of n observations of the dependent variable, \mathbf{X} is a $n \times m$ matrix of n observations of each of the m variables, and β is the $(m \times 1)$ column vector of unknown regression coefficients. This is the same as:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1m} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2m} \\ X_{31} & X_{32} & X_{33} & \cdots & X_{3m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & X_{n3} & \cdots & X_{nm} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_m \end{bmatrix} \quad (\text{B.66})$$

Ordinary Least Squares. Common multiple regression uses *ordinary least squares*, OLS, to estimate the regression coefficients. To begin, we first pre-multiply both sides by the transpose of \mathbf{X} , \mathbf{X}^T :

$$\mathbf{X}^T \mathbf{Y} = \mathbf{X}^T \mathbf{X} \beta \quad (\text{B.67})$$

where $(\mathbf{X}^T \mathbf{X})$ is a square matrix, diagonally dominant, and symmetric, and is closely related to the variance-covariance matrix:

Note that the variance of x_j is:

$$V_j = \frac{1}{n-1} \sum_{i=1}^n X_{ij}^2 \quad (\text{B.68})$$

and the covariance between X_j and X_k is:

$$C_{jk} = \frac{1}{n-1} \sum_{i=1}^n X_{ij} X_{ik} \quad (\text{B.69})$$

so that the diagonal of the $(\mathbf{X}^T \mathbf{X})$ matrix is related to the variance of each of the variables and the off-diagonal elements are the covariances.

$$\frac{\mathbf{X}^T \mathbf{X}}{(n-1)} = \begin{bmatrix} V_1 & C_{12} & C_{13} & \cdots & C_{1m} \\ C_{21} & V_2 & C_{23} & \cdots & C_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{2m} & C_{3m} & \cdots & V_m \end{bmatrix} \quad (\text{B.70})$$

and

$$\frac{\mathbf{X}^T \mathbf{Y}}{(n-1)} = \begin{bmatrix} C_{1Y} \\ C_{2Y} \\ C_{3Y} \\ \vdots \\ C_{mY} \end{bmatrix} \quad (\text{B.71})$$

where C_{iY} is the covariance between \mathbf{X}_i and \mathbf{Y} .

Continuing with our effort to estimate the unknown regression coefficients, $\hat{\beta}$, we now pre-multiply both sides by the inverse of the $(\mathbf{X}^T \mathbf{X})$ matrix, $(\mathbf{X}^T \mathbf{X})^{-1}$, yielding:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} (\mathbf{X}^T \mathbf{Y}) \quad (\text{B.72})$$

$$\mathbf{X}^T \mathbf{X} = \begin{bmatrix} \sum_{i=1}^n X_{1i}^2 & \sum_{i=1}^n X_{1i} X_{2i} & \sum_{i=1}^n X_{1i} X_{3i} & \cdots & \sum_{i=1}^n X_{1i} X_{mi} \\ \sum_{i=1}^n X_{2i} X_{1i} & \sum_{i=1}^n X_{2i}^2 & \sum_{i=1}^n X_{2i} X_{3i} & \cdots & \sum_{i=1}^n X_{2i} X_{mi} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n X_{mi} X_{1i} & \sum_{i=1}^n X_{mi} X_{2i} & \sum_{i=1}^n X_{mi} X_{3i} & \cdots & \sum_{i=1}^n X_{mi}^2 \end{bmatrix}$$

The *residuals* are the difference between the estimated and observed dependent variables:

$$\mathbf{e} = \mathbf{Y} - \tilde{\mathbf{Y}} \quad (\text{B.73})$$

The mean of the residuals is zero, $E(\mathbf{e}) = \bar{\mathbf{e}} = 0$, and the fitting error is the standard error of the residuals:

$$V(\mathbf{e}) = \tilde{\mathbf{e}}^2 = \mathbf{e}^T \mathbf{e} \quad (\text{B.74})$$

The variance-covariance matrix of $\hat{\beta}$ is:

$$V(\hat{\beta}) = \tilde{\mathbf{e}}^2 (\mathbf{X}^T \mathbf{X})^{-1} \quad (\text{B.75})$$

The variance-covariance matrix is very important because it contains information about how your model parameters are correlated. For example, if rainfall and temperature are inversely correlated, then a negative covariance is expected. Variables that are highly correlated are called *col-linear*, and result in large values in the variance-covariance matrix. *Multicolinearity* is the general problem of using variables that are highly correlated to each other, causing large prediction uncertainties.

Predicted outcomes, \mathbf{Y}_p , for specific *prediction* conditions, \mathbf{X}_p (as opposed to *observed* conditions), is given by:

$$\mathbf{Y}_p = \mathbf{X}_p \hat{\beta} \quad (\text{B.76})$$

and the prediction error is:

$$\begin{aligned} V(\mathbf{Y} - \mathbf{Y}_p) &= V(\mathbf{e}) + \mathbf{X}_p V(\hat{\beta}) \mathbf{X}_p^T \\ &= \tilde{\mathbf{e}}^2 + \tilde{\mathbf{e}}_{\beta}^2 \end{aligned} \quad (\text{B.77})$$

which means that the total prediction error depends on the fitting errors, $\tilde{\mathbf{e}}$, along with errors due to parameter errors, $\tilde{\mathbf{e}}_{\beta}$.

The confidence intervals about Y_p are given by:

$$Y_p \pm t(n, \alpha) V(\mathbf{Y} - \mathbf{Y}_p)^{1/2} \quad (\text{B.78})$$

Problems

1. Using downloaded data, select one of the variables as the dependent variable and find the regression relationship. Plot the observed and predicted values against each other.
2. Which model is better, one in which $V(e) \rightarrow 0$ or $V(\hat{\beta}) \rightarrow 0$?

B.6 Transformations

Ordinary least squares, the primary approach for forming a regression equation, assumes that the data are *homoscedastic*, meaning that the errors are constant, and are independent of the magnitude of the observation. In reality, many stream discharge measurements are variable - a function of the magnitude of the observation. These are *heteroscedastic* errors, which must be eliminated by transforming the data prior to performing a regression.

Hetero- vs. Homoscedasticity. Homoscedastic errors are uncorrelated to either magnitude of x or y , while heteroscedastic errors occur when they increase (or decrease) with the magnitude of x or y , and must be transformed (perhaps by taking logarithms) to make them homoscedastic.

Exponential Model. A linear reservoir drains according to the following exponential equation:

$$y = y_o \exp(-kt) = y_o e^{-kt} \quad (\text{B.79})$$

It is common that any errors in tail are much smaller than errors near peak - that is, the magnitude of any errors are a function of the magnitude of the observation. This implies that the errors are *heteroscedastic*. Recall that

For this exponential model, taking the log of the data eliminates heteroscedasticity:

$$\ln(y) = \ln(y_o) - kt \quad (\text{B.80})$$

which is now a linear model between $\ln(y)$ and t :

$$Y = Y_o - kt \quad (\text{B.81})$$

where $Y = \ln y$ and $y_o = e^{Y_o}$.

Log-Log Model. A non-linear reservoir model, such as a weir, has the general form:

$$y = y_o x^n \quad (\text{B.82})$$

which can be linearized by taking the logarithm of both sides:

$$\ln y = \ln y_o + n \ln x \quad (\text{B.83})$$

Introducing new variables, $Y = \ln y$ and $X = \ln x$, yields the general linear equation:

$$Y = Y_o + n X \quad (\text{B.84})$$

where $Y_o = \ln y_o$ or $y_o = e^{Y_o}$.

Polynomial Model. A general polynomial model is of the form:

$$y = a_o + a_1 x + a_2 x^2 + a_3 x^3 + \dots + a_n x^n \quad (\text{B.85})$$

To make this linear, we create new variables, $x_1 = x$, $x_2 = x^2$, etc., and then use these instead:

$$y = a_o + a_1 x_1 + a_2 x_2 + a_3 x_3 + \dots + a_n x_n \quad (\text{B.86})$$

and estimate the coefficients using OLS.

Power Model. A commonly used multiplicative, or power, model is:

$$y = a_o x_1^{a_1} x_2^{a_2} x_3^{a_3} \dots x_m^{a_m} \quad (\text{B.87})$$

To linearize this, we again use the logarithmic transform, $Y = \ln y$, $X_1 = \ln x_1$, etc., yielding:

$$Y = A_o + a_1 X_1 + a_2 X_2 + a_3 X_3 + \dots + a_m X_m \quad (\text{B.88})$$

where $A_o = \ln a_o$.

Examination of Residuals. One method for determining whether a transformation is needed is to examine the residuals. To begin, we first make a prediction using a linear model, $\hat{y} = \beta x$, and then calculate the residuals, $e = \hat{y} - y$, using the difference between the predicted and observed dependent variables. We then plot these residuals against the observed dependent variable, y vs e . Finally, we check to see if there is any correlation between y and e . If so, then we must transform of y until the correlation with the residuals is eliminated.

Example: Watershed Characteristics

Empirical models can be used, such as the *Benson* model, can be used to estimate peak flows, Q_p :

$$Q_p = a A^b S^c S_t^d I_b^e L^f \quad (\text{B.89})$$

where A is the watershed area, S is the main channel slope, S_t is the area of lakes and ponds, I_b is the maximum 24-hour, 10-year precipitation depth, and L is the main channel length.

Another model, proposed by *Scott*, is:

$$Q_p = a A^b E_m^c S_t^d P_s^e I^f T^g \quad (\text{B.90})$$

where E_m is the mean altitude, P_s is the average May to September precipitation depth, I is the maximum 24-hour,

2-year precipitation depth, and T is the mean January temperature.

And finally, the *Borland* model is:

$$Q_p = a A^b E_m^c S_h^d S^e S_t^f P_a^g P_s^h I^j L_a^j L_o^k \quad (\text{B.91})$$

where S_h is the watershed shape, P_a is the average October to April precipitation depth, L_a is the latitude, and L_o is the longitude.

These models are formed by using peak stormflows at gages where watershed information is available. Normally, these models are regional - in that they can only be applied in the area where they were developed.

To form your own model, logarithmic transforms are used to linearize the equations:

$$y = y_o + b x_1 + c x_2 + d x_3 + \dots \quad (\text{B.92})$$

where $y = \log Q_p$, $y_o = \log a$, $x_1 = \log(A)$, etc. Note that any number of factors can be added, thus improving the *fitting* or *calibration* error. *Multicollinearity* is major problem with this approach, however, causing large *prediction* errors. One must always check the prediction uncertainties to make sure that each added parameter further decreases the prediction error.

Problems

1. Using the same downloaded data from the preceding section, find the regression relationship using log-transformed values. Plot the observed and predicted values against each other.
2. Can $y = a(x - x_o)^b$ be linearized?

B.7 Time Series Analysis

A time-series model can be used to forecast variables that change with time. One can use previous observations of a time series to predict future values, called an *autoregressive* model. Adding previous prediction errors to the model can help improve forecasts, called a *moving average* model. Other variables, called *external inputs* often help to predict the forecast. And finally, converting the observed variables by either *differencing* or *integrating* them also can improve the prediction accuracy.

Autoregressive (AR). The autoregressive model uses previous observations to predict future observations:

$$\begin{aligned} \hat{y}(t) &= y_o + a_1 y(t-1) + a_2 y(t-2) + \dots + a_n y(t-n) \\ &= y_o + \sum_{i=1}^n a_i y(t-i) \end{aligned} \quad (\text{B.93})$$

where n is the memory of the autoregressive system. This is often written using $AR(n)$.

Moving Average (MA). The moving-average model adds the previous prediction errors, meaning that if one is off by a certain amount on the last time step, then the error, $e(t) = \hat{y}(t) - y(t)$, is added to the next prediction:

$$\begin{aligned}\hat{y}(t) &= y_o + b_1 e(t-1) + b_2 e(t-2) + \dots + b_m e(t-m) \\ &= y_o + \sum_{i=1}^m b_i e(t-i)\end{aligned}\quad (\text{B.94})$$

where n is the memory of the moving-average error. This is often written as $MA(m)$, or $ARMA(n, m)$ if the moving-average model is combined with the autoregressive model.

External Inputs (X). Another time-series approach is to use external inputs, such as precipitation or upstream flows, to predict downstream events. This takes the form:

$$\begin{aligned}\hat{y}(t) &= y_o + h_o x(t) + h_1 x(t-1) + \dots + h_p x(t-p) \\ &= y_o + \sum_{i=1}^p h_i x(t-i)\end{aligned}\quad (\text{B.95})$$

where p is the memory of the external input. This is the same as the convolution operator, where h_i is the unit response function. This is written as $X(p)$ or $ARMAX(n, m, p)$ when combined with the $ARMA(n, m)$ model:

$$\begin{aligned}\hat{y}(t) &= y_o + \sum_{i=1}^n a_i y(t-i) \\ &+ \sum_{i=1}^m b_i e(t-i) + \sum_{i=1}^p h_i x(t-i)\end{aligned}\quad (\text{B.96})$$

Differenced (D). Some processes are better described using their changes, such as the change in water level in a wetland. In this case, we take the first difference of the data:

$$\Delta y(t) = y(t) - y(t-1)\quad (\text{B.97})$$

The resulting time-series model could be a combination of $DAR(m)$, $DARMA(n, m)$, or $DARMAX(n, m, p)$, which would just use the Δy instead of y in the prediction equations.

Integrated (I). Yet other processes might be described using cumulative observations, such as using cumulative inflows to predict reservoir volumes:

$$Y(t) = Y(t-1) + y(t)\quad (\text{B.98})$$

The resulting time-series model could be a combination of $IAR(n)$, $IARMA(n, m)$, or $IARMAX(n, m, p)$, which would just use the Y instead of y in the prediction equations.

Partial Correlations. Partial correlations refers to the marginal contribution of the next coefficient. For example, if you have an $AR(1)$ model, find the net correlation of the next autoregressive term. If there is no partial correlation, and if $r_1 = 0.9$, then $r_2 = r_1^2 = 0.9 \times 0.9 = 0.81$. If we find $r_2 > r_1^2$, then there is a positive partial correlation. We can also say that if $r_2 < r_1^2$, then we have a negative partial correlation.

In general:

$$r_{ab|c} = \frac{r_{ab} - r_{bc} r_{ac}}{\sqrt{1 - r_{ac}^2} \sqrt{1 - r_{bc}^2}}\quad (\text{B.99})$$

For example, let $r_{ab} = 0.79$ and $r_{ac} = r_{bc} = 0.90$, then

$$r_{ab|c} = \frac{0.79 - (0.90) \times (0.90)}{\sqrt{1 - 0.90^2} \sqrt{1 - 0.90^2}} = -0.10\quad (\text{B.100})$$

Problems

1. Download three time-series of streamflow data from the U.S. Geological Survey water data website, preferably three stations within the same watershed.
2. For the most upstream location, find the $AR(2)$ model using both the original and log-transformed discharges. Plot the predicted and observed discharges.
3. For the most downstream location, find the $ARX(2, p)$ model using the other two stations. Plot the predicted and observed discharges.

B.8 Other Statistical Tools

Factor analysis. Factor analysis, and its cousin Principal Components Analysis (PCA), are commonly used to identify the relationships between variables using the covariance or correlation matrix. If C is the variance-covariance matrix, then we can form the equation:

$$\lambda C = D C\quad (\text{B.101})$$

where λ are the eigenvalues of C and D are the eigenvectors, with one eigenvector per eigenvalue. The eigenvalues correspond to the relative importance of a linear set of combinations of each of the original variables in explaining the correlation between these variables.

Kriging. Kriging uses the semivariogram to estimate both the most likely value along with the uncertainty in the estimate. The predicted value is interpolated based on nearby observations, with the closer observations being given greater weights than more distant observations.

$$\hat{x} = \sum_{i=1}^n \lambda_i x(i)\quad (\text{B.102})$$

where x_i are the nearby observations and λ_i are the weights.

Model updating. We can establish the regression equation $Y = X \beta$ between model outputs, Y , and model inputs, X . Our objective in regression is to minimize the errors, $J = \sum (y_o - y_p)^2$. Note that the predicted values of y , y_p , are a function of model parameters, β .

A recursive estimation can be used to update the model parameters, β , as the system changes (such as happens when floods change the cross-sectional area of the channel, a dike breaks, etc.):

$$\Delta\beta_i = (p_i)^{-1} g_i \quad (\text{B.103})$$

where

$$p_i = p_{i-1} + (X^T X) \quad (\text{B.104})$$

and

$$g_i = -\frac{\Delta J_i}{\Delta\beta} \quad (\text{B.105})$$

This approach means that we update the parameters by examining what the prediction errors are and coupling that with coefficients that tell which parameters are most useful for minimizing the error.

Kalman Filtering. Kalman Filtering is another means for updating your model. In this case we have system outputs, Y , and state variables, S . We say that we can predict the outputs based upon predictions of the state variables, which are only approximately known, using:

$$Y_{i,p} = H S_{i,p} \quad (\text{B.106})$$

We predict the state variables for the next time step based on the state variables from the previous time step using:

$$X_{i,p} = F S_{i-1,p} \quad (\text{B.107})$$

Once we have an observation of output, Y_i , we find the error between that and our prediction:

$$e_i = Y_{i,p} - Y_i \quad (\text{B.108})$$

We update the state variables using:

$$S_i = S_{i,p} + K e_i \quad (\text{B.109})$$

where K is the Kalman gain matrix:

$$K_{i,p} = P_{i,p} H^T [R + H P_{i,p} H^T]^{-1} \quad (\text{B.110})$$

where $R = C[e]$ and $E = C[\eta]$ in which $e = y_{i,p} - y_i$, $\eta = S_{i,p} - S_i$, and P is the covariance matrix associated with the estimation error of the states, which is dynamic and is predicted using:

$$P_{i,p} = F P + i - 1 F^T + Q \quad (\text{B.111})$$

The final value of this covariance matrix is:

$$P_i = P_{i,p} - P_{i,p} H^T [R + H P_{i,p} H^T]^{-1} H P_{i,p} \quad (\text{B.112})$$

Problems

1. Using downloaded data, calculate the eigenvectors and eigenvalues. Plot the two most significant eigenvectors against each other.