Chapter 9

Introduction to Multivariate Analysis

"The ocean: multiple, to a blinding
oneness."- Archie Randolph Ammons

As we learned in the previous chapter, the ability to mathematically test the statistical significance of values within a single variable certainly has its uses, and forms the foundation for hypothesis testing in the sciences. But the natural world is a very complex place, and rarely will a single variable be sufficient for our investigative needs. In truth, it is far more realistic to expect that the phenomena we seek to describe shall require mathematical models that imply some kind of interrelationship between two or more variables.

Multivariate analyses allow us to perform more complicated (and more meaningful) investigations. Multivariate methods typically try to establish either correlation or causation using a variety of statistical tests. Correlation simply attempts to establish that two (or more) variables are somehow related to each other, and that the value of one variable somehow exerts an influence on the value of a different variable. Correlation analysis doesn't explore how or why the variables are associated—it simply seeks to establish that an association exists. Causation analysis is a different and more powerful tool, because it seeks to define how the variables are related to each other, so we can use that relationship in a predictive manner (so the values of one variable can be used to predict the values of an associated variable). Not only does this process provide insight as to how the natural world functions, but it also allows us to develop mathematical models of complex phenomena we observe in the field or in the laboratory. If you can model it, you can simulate it.

Key Concepts

- Investigations of multivariate dynamics require that two or more variables be analyzed simultaneously.
- Correlation analyses are required to establish whether the variables under consideration are interrelated and covary in some way.
- Linear regression analyses can be used to develop a mathematical formula to predict the value of a dependent variable based on the measurement of a different, independent variable.
- Multiple linear regression analyses employ several different variables simultaneously to predict the value of a dependent variable.
- Curve estimation is used when the mathematical model that describes interrelated variables is nonlinear.

Introduction to Correlation Analysis

As we learned in the previous chapter, there are a wide variety of statistical methods (particularly the *t* tests) that we can use if all we want to do is focus on a single variable among different samples and simply compare one central tendency against another. In the context of statistics, this is necessary because we must be able to distinguish between two measurements that are "close enough" to be equal, and those that are truly different from each other. Where (and how) you draw the line between "statistically similar" and "statistically different" is the reason why *t* tests are so important, because they establish the statistical foundation that allows us to compare different values within a single variable. Then it is a relatively simple matter to expand our comparisons to include measurements of more than just one variable.

As a practical matter, the natural world is a complicated place so most scientific investigations require some kind of **multivariate analysis**, when we must compare two or more variables at the same time if we want to make any sense about the way the natural world functions. But it's also important that we limit our analyses to include only those variables that are germane to the hypotheses we are trying to test. In other words, don't waste your time measuring every variable under the sun—try instead to focus only on those variables that have some relevance to the question at hand.

That means the first order of business in any multivariate analysis is to establish whether the variables under investigation are (1) associated with each other and (2) associated with the phenomenon you are attempting to explain. Ideally, both conditions should be satisfied, since it is of little use to discover two variables that are associated with each other but have nothing to do with your hypothesis.

So how do we establish **correlation**? Fortunately, there are a host of statistical methods that we can use to mathematically define whether the variables in our dataset are indeed associated with each other. These same statistical tests can also be used to determine the relative "strength" of each association. This is especially useful in cases where there are several variables associated with each other, because it allows the researcher to immediately determine which associations are most significant and focus on those.

BE CAUTIOUS WITH CORRELATIONS

Correlations only imply that the variables are somehow associated with each other. Correlations can provide insight as to the strength of those associations, but they do not imply causation.

For example, an analysis of the chemical constituents of seawater would indicate an extremely strong correlation between dissolved Cl^- and dissolved SO_4^{2-}, simply because they are the two most common anions in the world ocean. Although we might be able to use that association to predict the relative concentrations of Cl^- and SO_4^{2-} in other seawater samples, it would be incorrect for us to claim that the Cl^- concentration is somehow caused by SO_4^{2-}.

Correlation Analysis Can Only Be Performed Among Continuous Variables

Demonstrating correlation between variables is a very powerful tool in scientific research, because correlation allows us to explore the connection between those variables, based on the results of our statistical tests. Not only can correlation analyses provide us with information as to how strong or weak the correlation is, they can also help us understand whether the association is positive or negative.

A **positive correlation** implies direct association between two or more variables. If we're dealing with continuous data, a positive correlation means that the measurements of the variables will "follow" each other (**Figure 9.1**). In other words, if the measurements in one variable get larger, the measurements of the other variable(s) will follow suit and also increase. Likewise, if the numerical values of one variable decline, the others would also decline.

A **negative correlation** implies an inverse relationship; that is, as the measurements of one variable increase, the measurements of the other variable(s) will decrease. Of course, if there is zero correlation (that is, neither positive nor negative), it simply means that the variables are completely independent of each other and are not associated at all. This condition would essentially look like the data for each variable are distributed randomly with respect to each other.

Correlation Analysis Can Be Used on No More Than Two Dependent Variables at a Time

At the heart of every correlation analysis is the intent to establish some kind of association between variables. If we take a moment to think about the logic of association, we have two fundamental possibilities. On one hand, it is conceivable that we might have a particular variable whose measured values are completely unaffected by any of the other variables we are measuring. We would call this an **independent variable** (IV). In contrast, there is also the distinct possibility that the value of some variable is being influenced by another variable. Not surprisingly, the variable exposed to this influence is called the **dependent variable** (DV). These distinctions are critically important, because correlation analysis requires that we first define what we think our DVs and IVs really are, so we can use correlation analysis to test whether our DVs are really determined by (dependent on) the IVs we've selected.

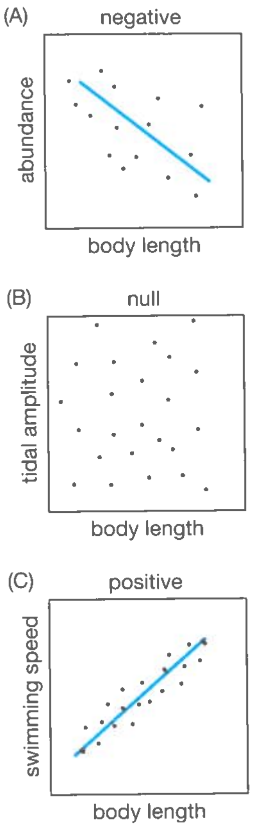


Figure 9.1 Correlation between variables is most easily visualized when the association is linear. Note that when the data are plotted on a Cartesian graph, negative correlations exhibit a negative slope (A) whereas positive correlations exhibit a positive slope (C). Technically speaking, it is not possible to describe a line for a null correlation (B) because *x* and *y* vary independently of each other.

As an example, let's consider the density of seawater (ρ) as measured in kilograms of seawater per cubic meter. Based on our previous education and knowledge of ocean properties, we might be tempted to assert that ρ will vary as a function of many different measureable parameters, such as

- Water temperature (because of thermal expansion or contraction of the water volume)
- Salinity (due to increased or decreased mass of dissolved solutes)
- Hydrostatic pressure (due to the compressibility of the water volume)
- Evaporation (due to the loss of fresh water from the water volume)
- Precipitation (due to the gain of fresh water to the water volume)

Since we are already making the claim that ρ is influenced by these variables, we have already defined ρ as our DV. For the other variables listed above, they can only be considered to be IVs if they are not influenced by each other in any way. We may be getting ourselves into trouble by using evaporation and precipitation here, simply because they could each affect our measures of water temperature and/or salinity. However, if we remove them from our list of IVs, we can still assume their effects would be included in measures of temperature and salinity. And if we make things even simpler and take our measurements from relatively shallow water, we can likely ignore the influence of hydrostatic pressure.

So that leaves us with only two IVs: water temperature and salinity. If we can convince ourselves that the salt content of a seawater sample would not influence its temperature, and that the temperature of a seawater sample would in no way determine its salt content, we can confidently claim that they are indeed independent. In this rather simple example, we count one DV (ρ) and two IVs (temperature and salinity). In the context of correlation, we would first test whether the two IVs are even associated with our DV, ρ. If an association can be demonstrated as either a positive or negative correlation (**Figure 9.2**), we would then use correlation analysis to explore whether either (or both) of the IVs can actually predict the value of our DV.

It is important to note that although we used only two independent variables in our simple example, there is no theoretical limit to the number of independent variables you can use in a correlation analysis (although practicality

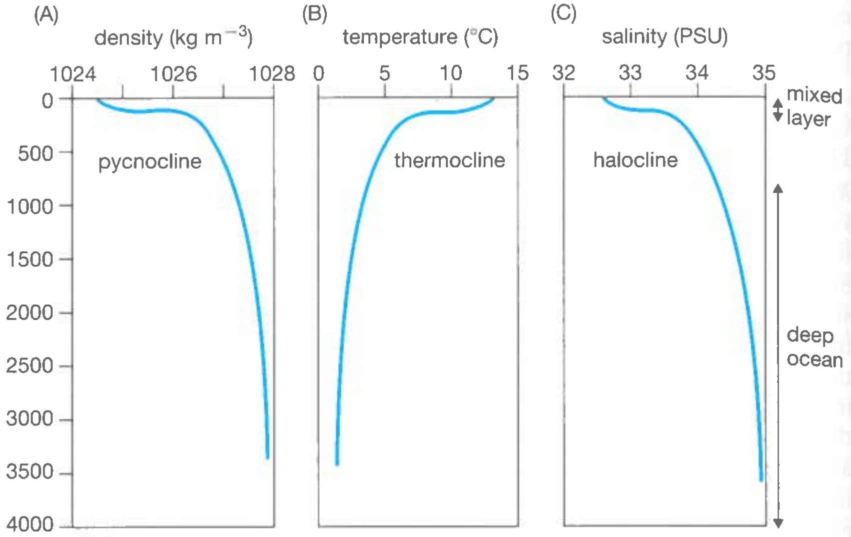


Figure 9.2 The measure of seawater density, as a function of temperature and salinity, is a well-established correlation in marine science. Note how the depth profile of seawater density (the pycnocline, A) shows an inverse relationship with temperature (the thermocline, B). This is indicative of a negative correlation between density and temperature. In contrast, the depth profile of salinity (the halocline, C) closely follows the same pattern of the pycnocline and is thus indicative of a positive correlation between density and salinity. Although these relationships may be easy to demonstrate by simply plotting the measurements, correlation analysis can be used to test the statistical significance of these relationships and may even be used to predict density values based on the strength of its positive correlation with salinity or its negative correlation with temperature.

will likely limit you to only a handful). And although you may use any number of independent variables, the statistical methods most commonly used for correlation analyses will limit you to no more than two dependent variables at a time. So take care to define the number of dependent variables you wish to test, as well as the number of different independent variables you will use as predictors of each dependent variable. When it comes time to choose the most appropriate type of correlation analysis, it is easy to provide guidance as long as you know how many dependent and independent variables you wish to include (Figure 9.3). There are several different correlation methods to explore, but the basic structure of the linear regression provides the foundation for all of the other correlation analyses.

Linear Regression Analysis

Statistically, the simplest model of association is the **linear regression**. For two variables (*X* and *Y*), we can use Equation 9.1 to express their association simply as

$$Y = mX + b \qquad (9.1)$$

where *X* and *Y* represent the measurements for each of the two variables we want to compare, *m* represents the slope of the resultant line, and *b* represents the *Y*-intercept (the special condition that defines the magnitude of *Y* when *X* is set to zero). In reality, our measurements will always exhibit some degree of random error (ε), so the linear correlation function is more correctly defined in Equation 9.2 as

$$Y = mX + b + \varepsilon \qquad (9.2)$$

It is important to note that not all associations are linear, so the linear regression model described by Equation 9.2 cannot be applied in all cases. However, it does represent the simplest example of a multivariate association between a single independent variable and a single dependent variable. Using the linear regression model (see Equation 9.2), we might even be able to use the independent variable *X*) to predict the value of the dependent variable (*Y*, presuming a strong correlation exists between *X* and *Y*. So before we get ahead of ourselves, it would seem that the first task of our correlation analysis is to determine whether an association between *X* and *Y* even exists.

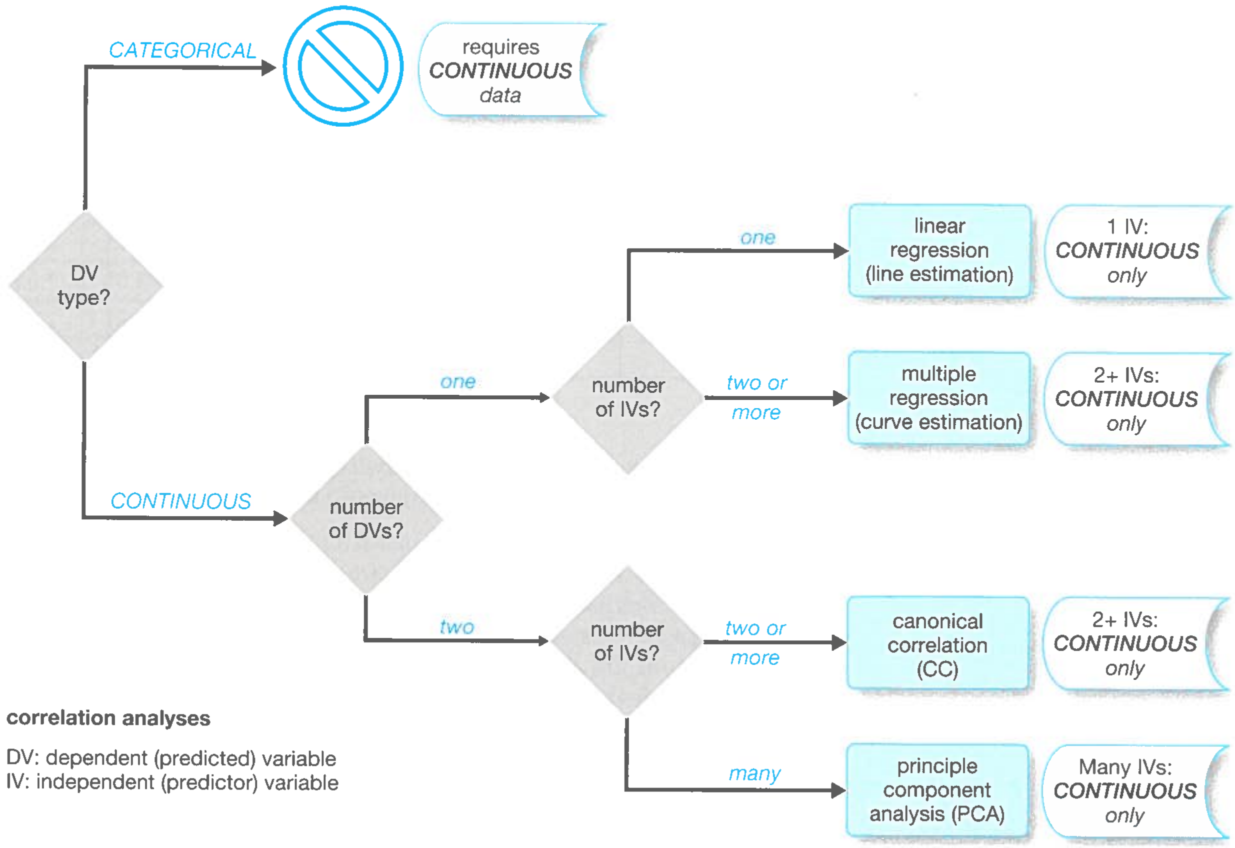


Figure 9.3 A simplified flowchart of the multivariate correlation analyses most commonly used in the natural sciences. Although all correlation analyses require continuous data, the most appropriate method is determined by the number of dependent variables (DVs) being predicted by correlation, as well as the number of independent variables (IVs) being used as predictors.

Correlations Among Normal Data Can Be Tested Using Pearson's Correlation Coefficient

The most common statistical method used to test the strength of a linear correlation between two variables is Pearson's correlation coefficient r, as defined in Equation 9.3:

$$r = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)\left(X_i - \overline{X}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2} \cdot \sqrt{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}} \qquad (9.3)$$

where X_i and Y_i represent the ith measurements of the X and Y variables, respectively, and $\overline{X}$ and $\overline{Y}$ are the means. Don't let the complexity of Equation 9.3 spook you: Pearson's coefficient is actually quite easy to calculate from spreadsheet data using modern statistical software.

As an example of correlation analysis, let's use a simple hydrography dataset that contains four variables: water temperature T, salinity S, density ρ, and wind speed U. In reality, we would want to use a much more robust dataset, but for the sake of this example, we can get away with using such a small dataset (**Table 9.1**).

If we assume that our data are normal, and that our pairwise correlations are indeed linear, we can use Pearson's r to investigate how our variables are

Table 9.1 Physical Properties Measured at Select Oceanographic Stations, Including Water Temperature (T, °C), Salinity (S, PSU), In-Situ Density (ρ, kg m^{-3}), and Average Wind Speed (U, m s^{-1})

Station	T (°C)	S (PSU)	ρ (kg m^{-3})	U (m s^{-1})
1	28.1761	19.4654	1010.87	3.3975
2	24.3896	15.9064	1010.90	3.4422
3	29.4930	24.8525	1013.28	2.8610
4	20.8722	29.1413	1021.07	2.0564
5	22.3083	33.0060	1022.42	2.9505
6	19.8067	31.7843	1023.15	3.1740

associated. Based on our example hydrography data in Table 9.1, the results of Pearson's r (using Equation 9.3) would be

$$\begin{array}{ll} T\,|\,\rho & r = -0.829 \\ & p = 0.041 \\ S\,|\,\rho & r = 0.949 \\ & p = 0.004 \\ U\,|\,\rho & r = -0.513 \\ & p = 0.298 \end{array}$$

For all datasets, the value and sign of Pearson's coefficient r will always fall in the range of $-1 < r < +1$. A positive correlation is indicated when $r > 0$ and a negative correlation when $r < 0$. The strength of the correlation can be determined by the magnitude of r. Very weak associations are indicated as $r \rightarrow 0$; thus, very strong associations can be deduced as $r \rightarrow \pm 1$.

When performing correlation analyses, our null (H_o) and alternate (H_a) hypotheses are always stated as

$$\begin{array}{ll} H_o\!: r = 0 & (X \text{ and } Y \text{ are not correlated}) \\ H_a\!: r \neq 0 & (X \text{ and } Y \text{ are correlated}) \end{array}$$

From the results of Pearson's analysis, we can see that water temperature T is strongly associated with water density ρ, as a negative correlation ($r = -0.829$). In other words, the inverse association indicates that as water temperatures increase, water density will decrease. The fact that $p < 0.05$ indicates that the result is statistically significant, so we must accept H_a and reject H_o with respect to the association between T and ρ.

The correlation between salinity S and water density ρ is even stronger ($r = +0.949$). Note that for this particular correlation, the sign of r indicates a positive correlation between salinity and density; therefore, as salinity values increase, we should expect density values to increase as well. The highly significant p-value ($p = 0.004$) indicates an even stronger statistical significance, so we must accept H_a and reject H_o with regard to the association between S and ρ as well.

At first glance, it would seem as though wind speed U and water density are negatively correlated ($r = -0.513$). However, since our p-value is much larger than 0.05, we cannot confirm that the result is statistically significant. In other words, we cannot confirm that the correlation does indeed exist, so we must reject H_a and accept H_o and therefore treat wind speed and water density as truly independent variables that are not at all associated with each other.

ASSUMPTIONS OF PEARSON'S LINEAR CORRELATION

1. Linearity: Variables X and Y must be associated according to the mathematical function $Y = mX + b + \varepsilon$.
2. Normality: The distribution of values for X and Y must follow a normal (Gaussian) curve, with a clearly discernible central tendency and standard deviation.

Table 9.2 Physical Properties Measured at Select Oceanographic Stations and Ranked According to Ascending Magnitudes of Water Temperature (T, °C), Salinity (S, PSU), and In-Situ Density (ρ, kg m^{-3})

Station	T (°C, rank)	S (PSU, rank)	ρ (kg m^{-3})	ρ (rank)
1	28.1761, 5	19.4654, 2	1010.87	1
2	24.3896, 4	15.9064, 1	1010.90	2
3	29.4930, 6	24.8525, 3	1013.28	3
4	20.8722, 2	29.1413, 4	1021.07	4
5	22.3083, 3	33.0060, 6	1022.42	5
6	19.8067, 1	31.7843, 5	1023.15	6

Correlations Among Nonnormal Data Can Be Tested Using Spearman's Ranked Correlation Coefficient

If our data are presumed to exhibit a linear correlation but are not normally distributed (or their distribution is unknown), we must rank our data from smallest to largest and compute Spearman's correlation coefficient based on the ranks rather than the actual measurements. For the sake of simplicity, let's use our earlier hydrography dataset as an example, but rank only the temperature T, salinity S, and density ρ measurements (**Table 9.2**).

The results of Spearman's correlation would then be

$$\begin{array}{ll} T\,(\text{ranked})\,|\,\rho\,(\text{ranked}) & r = -0.771 \\ & p = 0.072 \\ S\,(\text{ranked})\,|\,\rho\,(\text{ranked}) & r = 0.886 \\ & p = 0.019 \end{array}$$

From the results of Spearman's correlation analysis, we can see that the ranked values for the water temperature T are still negatively associated with water density (ρ), but those results are no longer statistically significant ($p > 0.05$). This is most likely a consequence of using so few data for our example of ranked correlation analysis; if we included more data (that is, more ranks), Spearman's results would likely comport well with Pearson's. Of course, if we had more data, we might even satisfy the central limit theorem and therefore be safe in assuming our data were normally distributed (allowing us to use Pearson's correlation in the first place).

With regard to our ranked salinity (S) data, salinity is still positively correlated with water density ($r = 0.886$), and the result remains significant ($p = 0.019$) even with such a small dataset. Hence, salinity and density are clearly the strongest association, based on our available data and the diversity of correlation analyses used.

Correlations Can Also Be Used to Make Predictions Using Regression Analysis

These sorts of correlation analyses are extremely critical to field researchers because they (ideally) allow us to eliminate superfluous variables and focus only on those that have the strongest associations. If we want to use those associations to make predictions about how the natural world functions, it is not enough to simply say they are correlated—what is really important is to define a mathematical model where the measured values of certain variables can be used to predict the values of some other variable. That process is commonly called **regression analysis**. In the simplest case, that mathematical model would be a familiar one: the equation for a line, where $Y = mX + b + \varepsilon$.

Recall that for linear regression analyses, we must assume that our data are continuous, and that we are investigating the correlation between two variables (*X* and *Y*) that exhibit a linear relationship according to the mathematical model defined earlier in Equation 9.2. Also keep in mind that the variables *X* and *Y* are true unknowns—these are the measurements we must take in the field or laboratory. The way Equation 9.2 is written, we can see the value of *Y* can only be determined if we know the value of *X*, as well as the slope *m*, the intercept *b*, and the random error ε associated with the $X|Y$ relationship. In this context, *Y* is considered to be the dependent variable, because its value is dependent on the function involving *m*, *X*, *b*, and ε. That means that *X* is considered to be the independent variable, because its value is completely unaffected by *Y*, *m*, *b*, and ε.

If this is the first time we're taking a crack at solving the linear regression equation (as defined in Equation 9.2), we will first have to take measurements of variables *X* and *Y*, but we won't know the values for *m*, *b*, or ε. Although the values of *m*, *b*, and ε are currently unknown, they are not variable—they actually represent different mathematical constants. A **constant** is a mathematical value that does not vary and is critical to the solution of a mathematical equation. And the nice thing about constants is that once you figure out what their value should be, that value will never change.

So let's take another look at Equation 9.2 and think about this for a bit. We know that both *X* and *Y* are variable, but we should be able to predict the value for *Y* (the dependent variable) if we can measure the value of *X* (the independent variable) and somehow know the constant values for *m* and *b*. If the correlation between *X* and *Y* is significant and ordered, we can use that association to estimate the values of *m* and *b* by simply gathering enough measurements of *X* and *Y* and letting that relationship "reveal itself" to us (**Figure 9.4**).

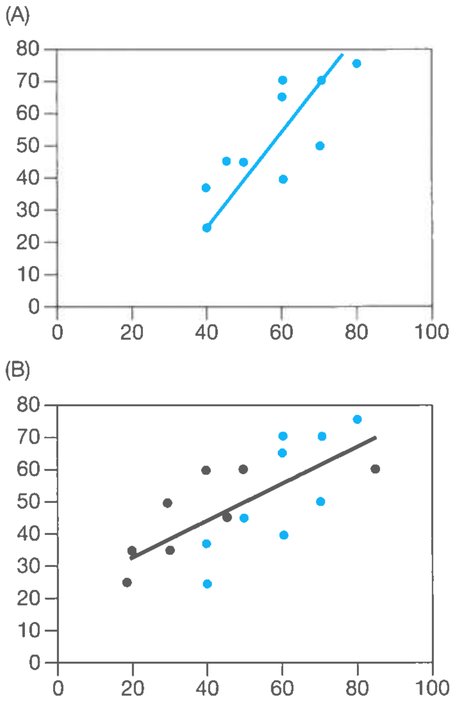


Figure 9.4 As more data are gathered, the true relationship between *X* and *Y* is revealed in the data themselves. Although there will always be error (ε) associated with our measurements, our original estimates of the slope *m* and intercept *b* of a linear regression (A) are based on few data points (blue). The original regression will improve (B) with the incorporation of more data points (black) as these new data will "correct" the original estimates of the slope *m* and intercept *b* of the regression. When the addition of new data does not significantly change the slope or intercept, it means we have finally determined them as constants.

Note that we have completely ignored any attempts to mathematically define ε. This is because ε represents the random errors in all our measurements of *X* and *Y* and in the relationship defined in Equation 9.2. Since ε represents random error, its randomness also makes it unknowable—so there's no point in trying to quantify it. In mathematics, we can cheat a little bit and say that our expectation for error, $E(\varepsilon)$, is zero. So that also means it is impossible to determine the "true" value of *Y*; what we're really trying to do is estimate the value of *Y*, based on all our measurements of *X* and *Y* (error and all). So we should modify Equation 9.2 and assert that our predicted value of *Y* (as *Y*′) is more correctly stated in Equation 9.4, now as

$$Y' = \hat{m}X + \hat{b} \tag{9.4}$$

where $\hat{m}$ and $\hat{b}$ are the statistical estimates for the slope and intercept, respectively.

The method for estimating $\hat{m}$ and $\hat{b}$ is a statistical one that uses our measurements of *X* and *Y* to estimate the slope $\hat{m}$ and intercept $\hat{b}$, and then tests how well those estimates predict the value of *Y* (as *Y*′) compared to the actual measurements of *Y*. Mathematically, $\hat{m}$ and $\hat{b}$ can be determined using Equations 9.5–9.6:

$$\hat{m} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2} \tag{9.5}$$

$$\hat{b} = \overline{Y} - \hat{m}\overline{X} \tag{9.6}$$

As you can see, the number of measurements (n) of X and Y affects the estimate of $\hat{m}$, which in turn is used to estimate $\hat{b}$ (using the mean values of X and Y). Fortunately, these estimates are routinely calculated by statistical software as part of the linear regression analysis, so one would rarely need to calculate these values by hand. But it is nonetheless instructive to take a look at the mathematical model used in Equation 9.5 to estimate $\hat{m}$, because it assumes that any random errors in our measures of X and Y will eventually cancel themselves out ($\varepsilon \rightarrow 0$) if enough observations (n) are made. Since $\hat{m}$ is then used to estimate $\hat{b}$ (in Equation 9.6), this same assumption applies to our estimate of $\hat{b}$.

Regression Analysis Is Used to Test Hypotheses and Define Confidence Intervals Within Predictive Mathematical Models

The most common reason for testing a hypothesis using linear regression is to test whether X and Y are truly associated with each other. As discussed earlier, the null hypothesis H_o in any correlation analysis assumes that X and Y are not at all associated with each other, in which case the estimated slope $\hat{b}$ would be zero. The alternate hypothesis H_a assumes that X and Y are indeed associated, and that Y can be predicted by using some function of X and the constants $\hat{m}$ and $\hat{b}$ (as defined in Equation 9.4). For brevity, this can be written as

H_o: $\hat{m} = 0$ (no association between X and Y')
H_a: $\hat{m} \neq 0$ (association exists; X predicts Y')

As standard output for any linear regression, the statistical software will provide the regression estimates of both the slope $\hat{m}$ and intercept $\hat{b}$, as well as the results of the two-tailed t test for $\hat{m}$. As usual, any $p < \alpha$ would indicate a statistically significant result, and the investigator would be compelled to accept H_a and reject H_o. This would also mean that the estimates of $\hat{m}$ and $\hat{b}$ could then be used with confidence to predict Y, according to Equation 9.4.

Depending on the statistical software used to conduct the linear regression analysis, the output might include the correlation coefficient r or the r-squared (r^2) value. The r^2 is simply that—the square of the correlation coefficient r, which represents the total percentage of Y values that are accurately predicted from X, using the regression estimates $\hat{m}$ and $\hat{b}$ (**Figure 9.5**).

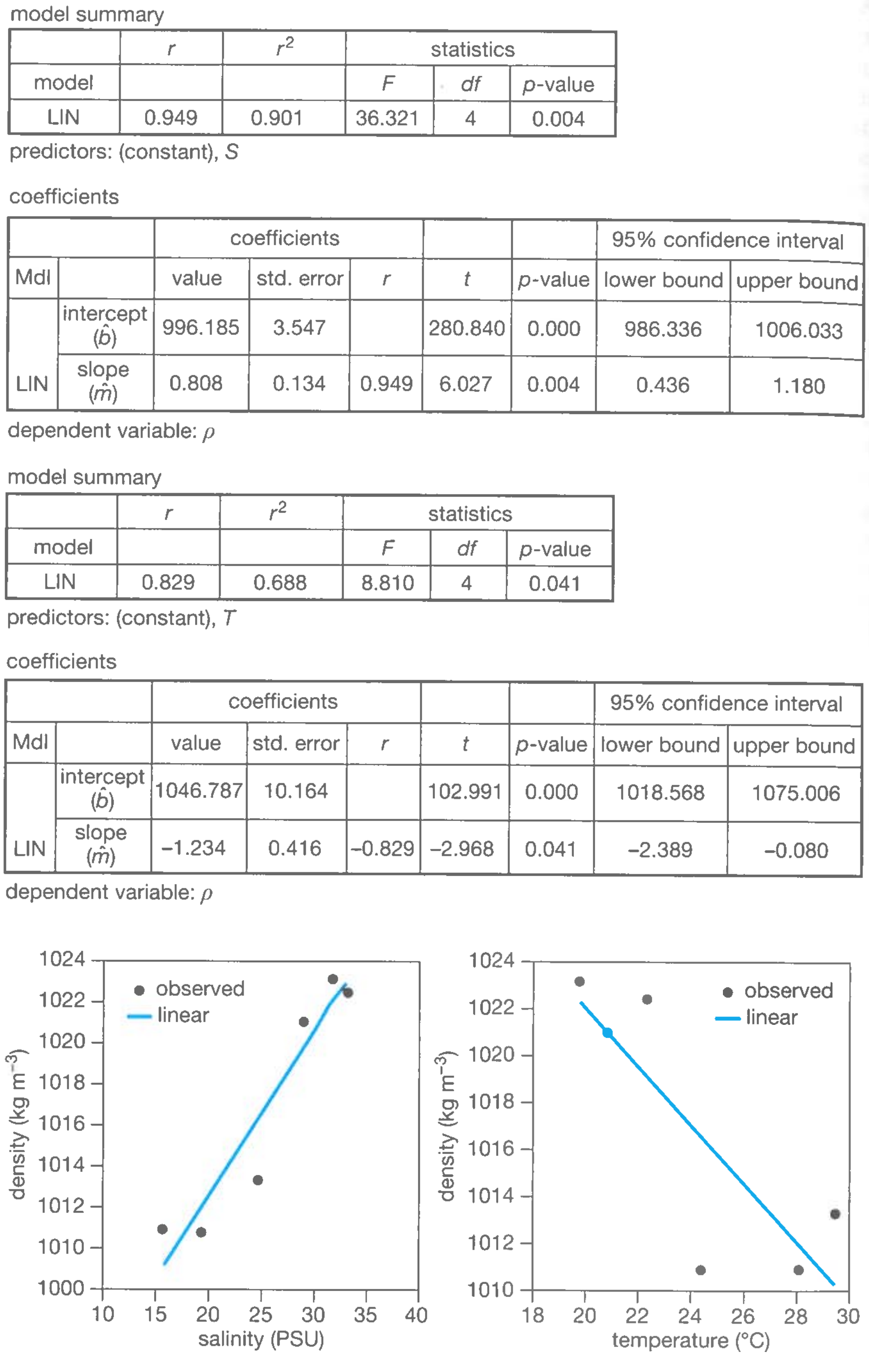
Let us take another look at the hydrography data we used in Table 9.1 when discussing correlation analyses. If these same data are used in a linear regression analysis, the results are shown in Figure 9.5.

In the first case, our values for salinity S are predictive of the density ρ according to the model described in Equation 9.7, based on the fundamental linear regression model (see Equation 9.4) fitted with the statistical estimates of $\hat{m}$ and $\hat{b}$ given in Figure 9.5:

$$\rho = 0.808\,S + 996.185 \tag{9.7}$$

Since the p-value for the slope is 0.004, we must accept H_a and conclude that the mathematical relationship expressed in Equation 9.7 is statistically significant. In fact, the r-squared (r^2) value of 0.901 indicates that 90.1% of all our measurements for ρ can be accurately predicted from our measurements of S. Generally speaking, all values of r^2 will range between 0 and 1, so the higher the r^2 value, the better the predictive power of the regression equation.

This is easily demonstrated by the results of the regression analysis for the temperature T and density ρ. When Equation 9.4 is refitted with our estimates



model summary

	r	r^2	statistics		
model			F	df	p-value
LIN	0.949	0.901	36.321	4	0.004

predictors: (constant), S

coefficients

		coefficients					95% confidence interval	
Mdl		value	std. error	r	t	p-value	lower bound	upper bound
LIN	intercept ($\hat{b}$)	996.185	3.547		280.840	0.000	986.336	1006.033
	slope ($\hat{m}$)	0.808	0.134	0.949	6.027	0.004	0.436	1.180

dependent variable: ρ

model summary

	r	r^2	statistics		
model			F	df	p-value
LIN	0.829	0.688	8.810	4	0.041

predictors: (constant), T

coefficients

		coefficients					95% confidence interval	
Mdl		value	std. error	r	t	p-value	lower bound	upper bound
LIN	intercept ($\hat{b}$)	1046.787	10.164		102.991	0.000	1018.568	1075.006
	slope ($\hat{m}$)	−1.234	0.416	−0.829	−2.968	0.041	−2.389	−0.080

dependent variable: ρ

Figure 9.5 Typical output from a statistical software program, indicating the estimated slope $\hat{m}$ and intercept $\hat{b}$ for each regression analysis, as well as the tests of statistical significance (p-values), r-squared (r^2) values, and 95% confidence intervals for the predictive functions. In this example, our measurements of salinity S and water temperature T were analyzed for their ability to predict water density ρ, the dependent variable, using a linear regression.

of $\hat{m}$ and $\hat{b}$, based now on the $\rho \mid T$ association (see Figure 9.5), an alternate regression equation for ρ can be defined in Equation 9.8, now as

$$\rho = -1.234\ T + 1046.787 \qquad 9.8$$

In this case, the result of our regression analysis is still significant ($p = 0.041$), so we must accept H_a and assume that Equation 9.8 is also a significant predictor of ρ. That being said, the weaker value for r^2 indicates that only 68.8% of our measurements of ρ can be accurately predicted from T.

Perhaps the most powerful aspect of the regression analysis is its predictive power. Although it is always preferable to collect real-world measurements of critical variables, our regression analyses indicate that we might not need to measure density at all. If we measure salinity or temperature instead, we can use the regression equations (see Equations 9.7 and 9.8) to predict the density values without actually having to measure them! Since Equation 9.7 exhibits a smaller p-value (stronger significance) and a higher r^2 value (better predictive capability), our regression analyses indicate that we would be well advised to use the salinity S as a more accurate predictor of the density ρ, rather than the water temperature T.

As standard output in most statistical programs, the linear regression will also provide 95% confidence intervals for the estimates of the slope $\hat{m}$ and intercept $\hat{b}$. By examining the upper and lower bounds of $\hat{m}$ and $\hat{b}$, the investigator can demonstrate the **confidence belt** that describes 95% of the data, where the main regression line serves as the central tendency (mean) of the regression analysis (**Figure 9.6**).

ASSUMPTIONS OF THE LINEAR REGRESSION METHOD

1. Linearity: Y must be a linear function of X and can be estimated as $Y' = \hat{m}X + \hat{b}$.
2. Independence: Each measurement of Y must be independent of all other measurements of Y.
3. Normality of Error: The random errors ε associated with the relationship between X and Y must be normally distributed with the same standard deviation.
4. Equal Variance: The variance s^2 must be equal for all ε.

Multiple Regression and Curve Estimation

Linear regression analysis is most useful when the dependent variable in the correlation can reliably be estimated with a single independent variable. However, natural systems are usually quite complex and are rarely estimable using a single variable. More often than not, we are faced with the extremely daunting task of predicting some natural behavior that is forced by several different variables, all operating at the same time and in very different ways.

Fortunately for us, the simple linear regression can be expanded to include more than one independent variable. Theoretically, it is possible to have an infinite number of independent variables that could act on and therefore determine the behavior of the dependent variable. A slight modification to Equation 9.4 is needed to account for this possibility, resulting in the multiple regression equation defined in Equation 9.9:

$$Y' = \hat{m}_1 X_1 + \hat{m}_2 X_2 + \hat{m}_3 X_3 + \cdots \hat{m}_i X_i + \hat{b} \tag{9.9}$$

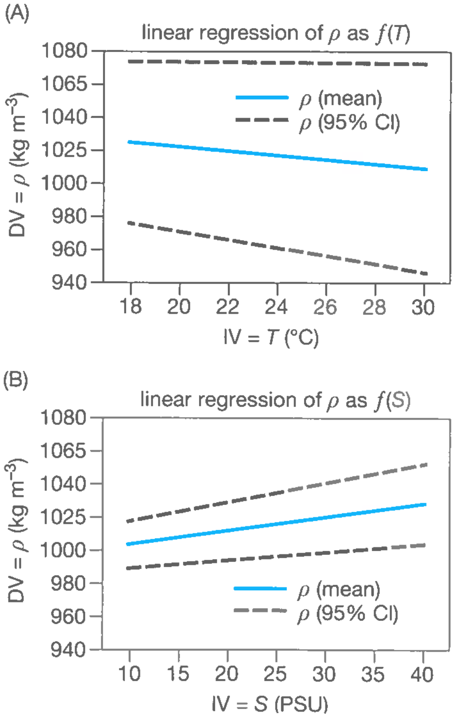


Figure 9.6 The regression equation that predicts the water density ρ as a function of the salinity S had a stronger statistical significance (lower p-value) and better correlation coefficient (higher r^2) than the function using the water temperature T (see Figure 9.5). We can plot the central regression line (mean) and the upper/lower bounds of the 95% confidence intervals of the y-intercept ($\hat{b}$) for each of the T (A) and S (B) regressions and analyze the variance predicted by each model. In this case, we can plainly see that the S regression (B) has a much lower predicted variance, as illustrated by the narrow confidence belt constrained between the upper/lower bounds of the 95% confidence intervals for $\hat{b}$. In contrast, the T regression (A) has a very broad confidence belt and would therefore be a less reliable predictor of ρ.

This is fundamentally the same as the simple linear regression in Equation 9.4, except that we now have to analyze several different slopes (and several different variables) instead of just one. That also means each slope must be tested (using a *t* test) to determine whether it is statistically different than zero. Just as before,

H_o: $\hat{m}_i = 0$ (no association between X_i and Y')
H_a: $\hat{m}_i \neq 0$ (association exists; X_i predicts Y')

As an example of a multiple regression, let us return to the simple hydrographic dataset in Table 9.1 that includes the unranked measurements of temperature *T*, salinity *S*, density ρ, and wind speed *U*. Earlier in this chapter, we were able to demonstrate through Pearson's correlation analysis that within our example dataset, wind speed is not correlated with density, but both salinity and temperature are. We then performed two separate linear regressions: the first to determine the linear equation where density is predicted using salinity, and the second where density is predicted instead by water temperature. Both salinity and temperature were found to be correlated with, and predictive of, density. However, they were each analyzed separately. Perhaps we could improve our ability to predict density if salinity and temperature were considered together.

Now, instead of analyzing all of the independent variables piecemeal, we might instead opt to perform a **multiple regression** analysis to determine if (and how) our measurements of density can be predicted using measurements of salinity, temperature, and wind speed (that is, multiple independent variables) simultaneously. Assuming they are all relevant to density, the relationship could be expressed as the multiple regression equation defined in Equation 9.10:

$$\rho' = \hat{m}_S S + \hat{m}_T T + \hat{m}_U U + \hat{b} \qquad (9.10)$$

When all three independent variables (*S*, *T*, *U*) are analyzed together in a statistical software package, the output will look something like **Figure 9.7**.

It would at first seem like ρ can be predicted perfectly ($r^2 = 1.00$) using measurements of water temperature, salinity, and wind speed together (see Figure 9.7A). However, since the slope for wind speed ($\hat{m}_U$) has $p > 0.05$, we cannot confidently claim that this particular correlation is significant and we must accept the null hypothesis H_o that $\hat{m}_U = 0$ and can therefore be ignored with regard to the multiple regression equation for ρ (see Equation 9.10). That would give us a new generalized equation for ρ, which is now more accurately described as Equation 9.11:

$$\rho' = \hat{m}_S S + \hat{m}_T T + \hat{b} \qquad (9.11)$$

If the multiple regression analysis is run again, this time by eliminating the wind speed *U* as one of the independent variables, the standard error of the estimate for ρ does unfortunately increase. The good news is that the increase in the standard error of the estimate for ρ is very slight, our r^2 is still 1.00, and all the p-values have remained very strongly significant now that the prediction equation has been streamlined (see Figure 9.7B), such that only the relevant variables (namely, salinity, and water temperature) are included. The result is a prediction equation for ρ, where Equation 9.11 is now fitted with our estimates of $\hat{b}$, as well as our simultaneous estimates of both $\hat{m}_S$ and $\hat{m}_T$, to become Equation 9.12:

$$\rho' = 0.601S - 0.592T + 1015.817 \qquad (9.12)$$

This is a vast improvement over the simple linear regression analysis, because we have been able to establish that ρ is actually correlated with

(A)

model summary

model	r	r^2	std. error of the estimate
LIN	1.000	1.000	4.837E-02

predictors: intercept ($\hat{b}$), windspd, T, S
dependent variable: ρ

coefficients

		coefficients				95% confidence interval	
Mdl		value	std. error	t	p-value	lower bound	upper bound
LIN	intercept ($\hat{b}$)	1016.215	0.312	3260.418	0.000	1014.187	1017.556
	S	0.597	0.004	136.047	0.000	0.578	0.616
	T	–0.592	0.007	–85.385	0.000	–0.622	–0.562
	U	–0.103	0.050	–2.056	**0.176**	–0.318	0.112

dependent variable: ρ

(B)

model summary

model	r	r^2	std. error of the estimate
LIN	1.000	1.000	6.969E-02

predictors: (constant), T, S
dependent variable: ρ

coefficients

		coefficients				95% confidence interval	
Mdl		value	std. error	t	p-value	lower bound	upper bound
LIN	intercept ($\hat{b}$)	1015.817	0.352	2886.337	0.000	1014.697	1016.937
	S	0.601	0.006	105.257	0.000	0.583	0.619
	T	–0.592	0.010	–59.311	0.000	–0.624	–0.560

dependent variable: ρ

Figure 9.7 Typical output from a statistical software program, indicating the estimated slope and intercept for every independent variable analyzed in the multiple regression analysis (A). By eliminating variables that are not significantly correlated with the dependent variable (B), the multiple regression can be run again and improved.

water temperature and salinity together, as variables that are complementary to each other in their power to predict the dependent variable ρ. In fact, if we use the upper and lower bounds of the confidence interval of the y-intercept ($\hat{b}$) of the multiple regression compared to Figure 9.6, we can visually demonstrate the superiority of our new multiple regression model that now boasts a very narrow confidence belt (**Figure 9.8**). Had we not performed the multiple regression analysis, we would have erroneously used either salinity or water temperature to estimate density, when we should have been using both.

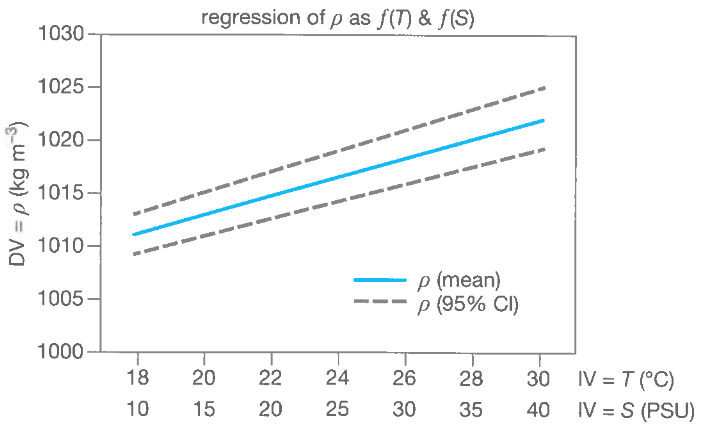


Figure 9.8 A plot of the central tendency and the upper/lower bounds of the 95% confidence interval of the y-intercept ($\hat{b}$) for the dependent variable ρ, as predicted by the multiple regression of both independent variables T and S, considered simultaneously. Indeed, the multiple regression of ρ boasts a much lower predicted variance, as illustrated by the narrow confidence belt constrained between the upper/lower bounds of the 95% confidence interval of $\hat{b}$. Note the difference in scale of the y-axis for the multiple regression, where the variance in the predicted ρ values ranges from 1009 to 1025 kg m^{-3}, while the ρ values from the linear regressions in Figure 9.6 range anywhere from 940 to 1080 kg m^{-3}.

Multiple Regressions That Are Nonlinear Must Be Determined Using Curve Estimation

Most statistical applications also possess the capability to perform more complicated, nonlinear regressions as a function of a single independent variable. Although nonlinear multiple regression methods do exist, they require a significant level of mastery to perform and are admittedly beyond the introductory scope of this text. The reader is invited to consult more advanced statistical methods for nonlinear regression analyses involving two or more independent variables.

Within the variety of nonlinear mathematical relationships that require a single independent variable for predictive capability, the most common are defined as follows:

$$Y' = \hat{b} + \left(\hat{m} \cdot ln\ X\right) \qquad \text{Log} \tag{9.13}$$

$$Y' = \hat{b} + \frac{\hat{m}}{X} \qquad \text{Inverse} \tag{9.14}$$

$$Y' = \hat{b} + \hat{m}_1 X + \hat{m}_2 X^2 \qquad \text{Quadratic} \tag{9.15}$$

$$Y' = \hat{b} + \hat{m}_1 X + \hat{m}_2 X^2 + \hat{m}_3 X^3 \qquad \text{Cubic} \tag{9.16}$$

$$Y' = \hat{b} X^{\hat{m}} \qquad \text{Power} \tag{9.17}$$

$$Y' = \hat{b} \hat{m}^X \qquad \text{Compound} \tag{9.18}$$

$$Y' = e^{\hat{b} + \left(\frac{\hat{m}}{X}\right)} \qquad \text{S} \tag{9.19}$$

$$Y' = \frac{1}{\left(1/u\right) + \left(\hat{b} \hat{m}^X\right)}; \quad u > |Y_{max}| \qquad \text{Logistic} \tag{9.20}$$

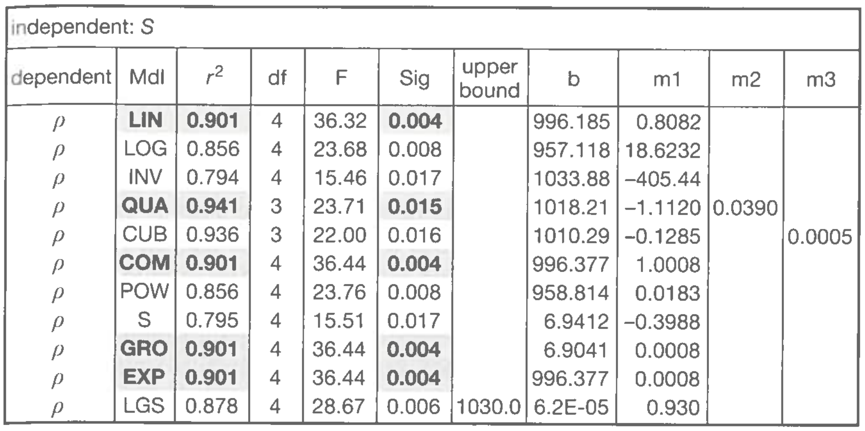
$$Y' = e^{\hat{b} + \hat{m} X} \qquad \text{Growth} \tag{9.21}$$

$$Y' = \hat{b} \cdot e^{\hat{m} X} \qquad \text{Exponential} \tag{9.22}$$

Curve estimation simply involves entering the independent and dependent variables into the statistical curve-fitting routine and selecting from the available models (see Equations 9.13–9.22). Depending on the statistical software being used, the results of curve estimation will be produced in an output file similar to **Figure 9.9**.

In this example, the salinity S was once again analyzed for its capacity to accurately predict the density ρ. According to this curve-fit analysis, all of the nonlinear models exhibit $p < 0.05$; therefore, they are all suitable as a predictive model for ρ. However, the model with the highest r^2 value should be chosen, as it represents the model that most accurately estimates the dependent variable ρ.

Recall from our earlier example that the simple linear regression between salinity and density had $r^2 = 0.901$ (see Figure 9.5). Our curve-fit analysis would allow us to improve the early model, particularly if we choose the quadratic model with $r^2 = 0.941$. If this were the case, we would use the



Independent: *S*										
dependent	Mdl	r^2	df	F	Sig	upper bound	b	m1	m2	m3
ρ	**LIN**	**0.901**	4	36.32	**0.004**		996.185	0.8082		
ρ	LOG	0.856	4	23.68	0.008		957.118	18.6232		
ρ	INV	0.794	4	15.46	0.017		1033.88	-405.44		
ρ	**QUA**	**0.941**	3	23.71	**0.015**		1018.21	-1.1120	0.0390	
ρ	CUB	0.936	3	22.00	0.016		1010.29	-0.1285		0.0005
ρ	**COM**	**0.901**	4	36.44	**0.004**		996.377	1.0008		
ρ	POW	0.856	4	23.76	0.008		958.814	0.0183		
ρ	S	0.795	4	15.51	0.017		6.9412	-0.3988		
ρ	**GRO**	**0.901**	4	36.44	**0.004**		6.9041	0.0008		
ρ	**EXP**	**0.901**	4	36.44	**0.004**		996.377	0.0008		
ρ	LGS	0.878	4	28.67	0.006	1030.0	6.2E-05	0.930		

Figure 9.9 Typical output from a statistical software program, indicating the results from a variety of curve-fitting models. The model with a significant *p*-value and the highest r^2 represents the best possible curve fit for the data. In this example, the LIN, COM, GRO, and EXP models all share the lowest *p*-value (0.004), but the QUA model has the best *r*-squared (r^2) value (0.941). Since the QUA model has the best r^2 value and still has a strong *p*-value (0.015), the QUA model is most likely to produce the best results as a predictive model of ρ when using *S* as the independent variable.

fundamental quadratic model (as defined in Equation 9.15) and refit the model with our own statistical estimates of $\hat{b}$ and $\hat{m}_i$ from Figure 9.9 to yield Equation 9.23:

$$\rho' = 1018.21 - 1.112S + 0.039S^2 \qquad (9.23)$$

Although the quadratic model in Equation 9.23 is certainly an improvement from the simple linear model we derived as Equation 9.7, we have already demonstrated that the multiple regression equation that describes ρ as a function of both salinity and water temperature (see Equation 9.12) is an improvement beyond all of the curve-fit models. Thus, we should stick with Equation 9.12 as our preferred model for estimating ρ.

Correlation Between Two Dependent Variables

In our previous examples, we have focused our attention on the correlation methods used to investigate the correlation between one or more independent variables (a set of predictors) and a single dependent variable. But what do we do if we have two dependent variables we'd like to examine for correlation? That's when we must invoke more complicated **multidimensional scaling** methods in order to analyze our data. And although that sounds impressive (or scary), we already have practice in multidimensional scaling. Recall the different slopes we derived for *S* and *T* in our multiple regression of ρ in Equation 9.12. Each of those slopes was essentially a scaling factor, applied to each of the independent variables in our equation in order to provide the best possible prediction of ρ. The more variables we have to work with, the more dimensions we have to scale.

Canonical Correlation (CC) Is Used to Explore Correlations Between Two Sets of Dependent Variables

Consider the situation we have just examined using multiple regression analysis, where multiple independent variables like salinity *S* and temperature *T* are used to predict a single dependent variable, such as ρ. If we wished to expand our oceanographic analyses to include a new multiple regression, we might be also be interested to predict phytoplankton biomass (as *chl*, a single dependent variable) as a function of several independent variables represented as dissolved nutrients (like NH_4^+, NO_3^-, and PO_4^{3-}). In much the same fashion, we would simply use the general multiple regression model (see Equation 9.9) and adapt it for our specific interests, becoming Equation 9.24:

$$\rho' = \hat{m}_S S + \hat{m}_T T + \hat{b} \;\; \& \;\; chl' = \hat{m}_{NH_4} NH_4^+ + \hat{m}_{NO_3} NO_3^- + \hat{m}_{PO_4} PO_4^{3-} + \hat{b} \qquad (9.24)$$

Here we would have two sets of dependent variables: one set describes ρ and the other set describes *chl*, each set defined by its own independent variables. In this case, if we wanted to explore both sets of dependent variables (ρ and *chl*) for correlation, we would have to use **canonical correlation** (CC).

In its simplest form, CC analysis is used to explore how the results of two different multiple regressions may be correlated with each other. These types of comparisons will produce *p*-values to indicate the significance of the correlation, as well as an r^2 value, but CC analyses can be far more complex than just a "head-to-head" correlation analysis of two multiple regressions. In some circumstances, each multiple regression might share the same independent variable, or the solution to one multiple regression (the dependent variable) may in fact be one of the independent variables used in the other multiple regression. The reader is invited to refer to the Further Reading section for more information about how to properly conduct CC analyses.

Principal Component Analysis (PCA) Is Used to Create Two New Sets of Dependent Variables and Explore Their Correlation

In its simplest sense, **principal component analysis** (PCA) is a statistical method used to analyze a large number (n) of different independent variables within a large dataset in an effort to reduce those n dimensions of variability into a single "synthetic" dimension that becomes the principal component defining that particular population. The same is then done to a different population to define its own principal component, and the principal components of each population are then compared and analyzed for correlation. Essentially, PCA attempts to render all of the variability witnessed in each population as a unique, n-dimensional "cloud," and then overlay those clouds to explore any coclustering (positive correlation) or cluster separation (negative correlation) patterns, relative to the principal components chosen to represent each population (**Figure 9.10**). Because PCA deals with multidimensional comparisons, it is an inherently complex method of statistical correlation. If the reader intends to employ multidimensional scaling methods for correlation analyses, then the Further

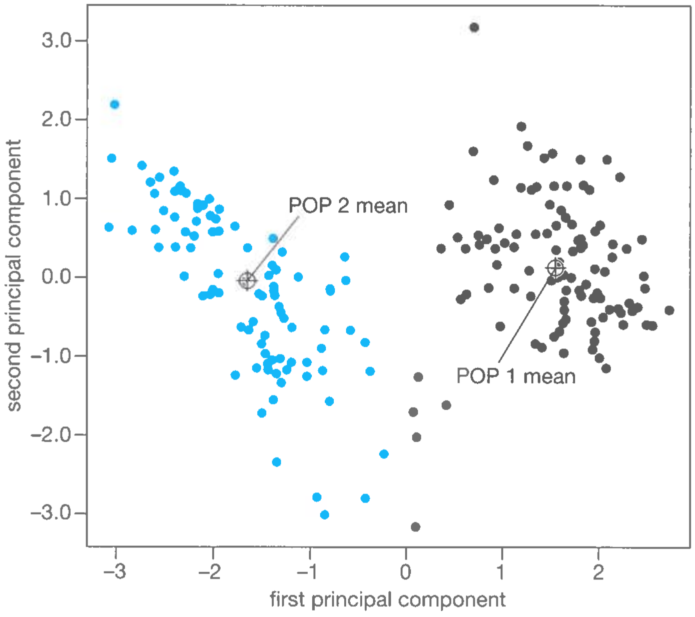


Figure 9.10 Example output from principal component analysis (PCA), where the *n*-dimensional variability among each population is reduced to a synthetic, one-dimensional principal component that represents the maximum gradient of variability within that component of analysis. By comparing (or plotting) the variability "cloud" of two different populations, across two or more principal components, the clustering patterns can be used to derive correlations. In this case, the centroid of population 1 (black) is negatively correlated with the centroid of population 2 (blue) according to the gradient of variability in principal component 1 (+1.5 and −1.7, respectively). Although there appears to be significant variability with respect to the "spread" of data along principal component 2, the centroids of both populations are virtually indistinguishable from each other (~0).

Reading section has more information about how to properly conduct principal component analyses.

From Measurements to Models

As we have seen in this chapter, correlation methods are absolutely critical to the analytical process, because rarely do we know that variables are predictive of each other. By working through the correlation and regression analysis procedures, we can investigate (and test) how multiple variables can be used simultaneously, in a predictive way.

In our earlier examples, we demonstrated how ocean density could be modeled (estimated) by using measurements of water temperature and salinity (see Equation 9.12). In this manner, it would be a relatively simple task to simulate how the density of the ocean would respond to changing temperatures and salinities. Of course, the data we used to develop our regression equations were only meant to serve as an example; if we truly wished to develop a realistic regression equation, we would be required to incorporate a vastly larger dataset, perhaps with even more variables to consider.

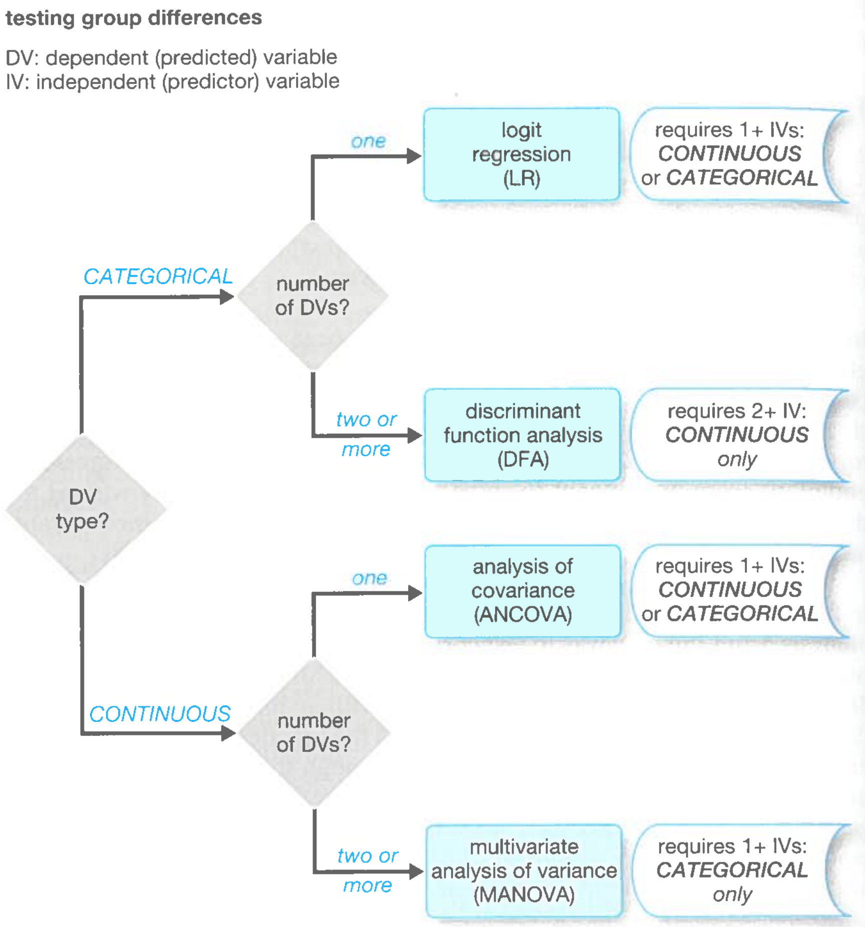
In fact, this will very likely be the case. It is important to note that regression analysis is not the only way to test correlation. In fact, considering the multitude of correlation analyses that exist, we have only scratched the surface. This chapter is meant to provide the reader with an introduction to correlation analysis as it relates more specifically to the practice of numerical modeling. If the reader is in need of a more thorough treatment of the several other correlation methods that exist beyond those included here, the suggested readings are an excellent start. Of course, it may be necessary for the reader to consult more advanced statistical texts in order to explore the CC and PCA methods referenced in Figure 9.3.

It is also possible to use complex correlation analyses to test group differences using either continuous data or categorical data, as a kind of multivariate *t* test or ANOVA. Some of these more advanced methods test for association, but include sources of variation within the independent variables that were not controlled in (or removed from) the experiment but were still expected to somehow affect the value of the dependent variable(s). Although there are several different correlation methods for testing group differences (**Figure 9.11**), the reader is invited to refer to the Further Reading at the end of this chapter for more information on these and other correlation methods.

Regardless of the multivariate method chosen, correlation analyses enable researchers to formulate meaningful associations among and between their measured variables, allowing them to transition from simply measuring variables in the field to creating predictive models of complex natural phenomena. The careful consideration of which variables to measure comes first; then, we must execute the research plan and actually take those critical measurements. Ultimately, our efforts will yield a robust dataset that will allow a variety of regression analyses to tease out the correlations between those variables and how we can use those relationships to make mathematical predictions.

You may not have realized it, but the simple act of working through our examples and developing the several different regression equations in this chapter was actually an exercise in numerical modeling. In our examples, we were modeling (simulating) ocean density by using the salinity and temperature measurements from our dataset. You just modeled an important physical variable of ocean dynamics! So let's build on that and see what other models we can develop from our measurements.

Figure 9.11 A simplified flowchart of the multivariate analyses most commonly used in the natural sciences to test group differences. These methods can be used to analyze either continuous data or categorical data, but the most appropriate method is determined by the number of dependent variables (DVs) being predicted by correlation, as well as the number of independent variables (IVs) being used as predictors.


testing group differences
DV: dependent (predicted) variable
IV: independent (predictor) variable
DV type?
CATEGORICAL
number of DVs?
one
logit regression (LR)
requires 1+ IVs: CONTINUOUS or CATEGORICAL
two or more
discriminant function analysis (DFA)
requires 2+ IV: CONTINUOUS only
CONTINUOUS
number of DVs?
one
analysis of covariance (ANCOVA)
requires 1+ IVs: CONTINUOUS or CATEGORICAL
two or more
multivariate analysis of variance (MANOVA)
requires 1+ IVs: CATEGORICAL only

References

Jones ER (1996) Statistical Methods in Research. Edward R. Jones.

Kanji GK (1999) 100 Statistical Tests. SAGE Publications.

Keeping ES (1995) Introduction to Statistical Inference. Dover Publications.

Keller DK (2006) The Tao of Statistics. SAGE Publications.

Kenny DA (1979) Correlation and Causality. John Wiley and Sons.

Linton M, Gallo Jr PS, & Logan CA (1975) The Practical Statistician: Simplified Handbook of Statistics. Wadsworth Publishing Company.

Mandel J (1964) The Statistical Analysis of Experimental Data. Dover Publications.

Newman I, & Newman C (1977) Conceptual Statistics for Beginners. University Press of America.

Rencher AC & Christensen WF (2012) Methods of Multivariate Analysis, 3rd ed. John Wiley & Sons.

Salkind NJ (2007) Statistics for People Who (Think They) Hate Statistics: The Excel Edition. SAGE Publications.

Steiner F (ed) (1997) Optimum Methods in Statistics. Akadémiai Kiadó.

Thompson SK (1992) Sampling. John Wiley and Sons.

Further Reading

Afifi A, May S, & Clark VA (2011) Practical Multivariate Analysis, 5th ed. CRC Press.

Everitt B & Hothorn T (2011) An Introduction to Applied Multivariate Analysis with R. Springer.

Gittins R (1985) Canonical Analysis: A Review with Applications in Ecology. Springer-Verlag.

Jackson JE (2003) A User's Guide to Principal Components. John Wiley & Sons.

Jolliffe IT (2002) Principal Component Analysis. Springer-Verlag.

Kachigan SK (1986) Statistical Analysis: An Interdisciplinary Introduction to Univariate and Multivariate Methods. Radius Books.

Mari DD & Kotz S (2001) Correlation and Dependence. Imperial College Press.

Parkhomenko E (2009) Sparse Canonical Correlation Analysis: Data Integration for Regular and High Dimensional Studies. VDM Verlag.

Timm NH & Mieczkowski TA (1997) Univariate and Multivariate General Linear Models. SAS Institute.

Wei WWS (2005) Time Series Analysis: Univariate and Multivariate Methods, 2nd ed. Pearson.