Module 3: Identifying relationships between variables using correlation and linear regression

Introduction

In this module we will explore statistical approaches for evaluating whether two variables are associated with one another (correlation) and for predicting the value of one variable using known values of another (linear regression). Like the previous module, our focus here will be on data that is distributed continuously, such as weight, height, or length. The types of problems we can tackle using these approaches include, for instance, assessing the strength of the relationship between latitude and species richness or perhaps evaluating whether population density predicts the growth rate of an infectious disease. In cases like these, it is possible to quantify the strength of the relationship between two continuous variables by calculating their ***correlation***, or to use one variable to predict the value of another through ***linear regression***.

Practical Background

Correlation:

We begin by developing the idea of correlation. The goal is to evaluate whether or not two variables, X and Y, covary with one another such that increases or decreases in one variable are associated with increases or decreases in the other. In general, correlation is used when we do not have a clear *a priori* expectation that changes in one variable cause changes in the other.

As long as two variables are drawn from a bivariate normal distribution, their correlation can be estimated using the following formula:

$$\rho = \frac{Cov[X,Y]}{\sigma_X \sigma_Y} \quad (1)$$

where $Cov[X, Y]$ is the covariance between the variables X and Y, σ_X is the standard deviation of variable X and σ_Y is the standard deviation of variable Y. The correlation ρ, must – by definition – always lie between -1 and 1 with a value of -1 indicating the two variables are perfectly correlated in a negative fashion and a value of 1 meaning the two variables are perfectly correlated in a positive fashion. Positive correlations indicate that both variables increase and decrease together whereas negative correlations indicate that increases in one variable are associated with decreases in the other and vice versa.

Estimating the correlation between two variables in a sample is straightforward using (1), but we would also like to know whether our estimate differs significantly from zero. This can be accomplished by first calculating the following test statistic:

$$t = \rho\sqrt{(n-2)/(1-\rho^2)} \quad (2)$$

where ρ is the correlation within the sample and n is the sample size. Next, simply compare the value of this test statistic to the critical value of t drawn from the student's t-distribution with n-2 degrees of freedom and the desired significance level, $\alpha/2$.

Linear regression:

Correlation allows us to study whether or not two variables are associated with one another but does not explicitly assume one variable causes changes in the other. As a consequence, estimating correlations does not allow us to predict values of one variable based on the observed values of another. To make predictions in this way, we need to make use of linear regression. Specifically, linear regression allows us to estimate the parameters a and b of the regression equation $\hat{Y} = a + bX$ from a sample of X and Y values. Although the derivation of the formulas we will use is beyond the scope of this course, the underlying reasoning is very simple and based on minimizing the error between the values of Y predicted from the regression equation and their actual value (Figure 1).

We begin by introducing a formula that allows the slope, b, of our linear regression to be estimated:

$$b = \frac{Cov[X,Y]}{\sigma_X^2} \tag{3}$$

where $Cov[X, Y]$ is the covariance between variable X and variable Y and σ_X^2 is the variance of variable X. Once the value of the slope, b, has been estimated, we can proceed to estimate the intercept, a, using the following formula:

$$a = \bar{Y} - b\bar{X} \tag{4}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means for variable X and Y, respectively. Remarkably, with just that small bit of calculation, we have the tools we need to estimate the parameters of our linear model! All we need to do now is learn how to test whether the slope of the regression, a, is significantly different from zero.

An easy and general way to evaluate the statistical significance of the slope is to place a confidence interval around the value of b; if this confidence interval includes zero, the data does not provide evidence that the slope differs significantly from zero. The first step in placing a confidence interval around our estimate of the slope, b, is to calculate its standard error:

$$S_b = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{(n-2)} \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{5}$$

where the quantity, $\hat{Y}_i$, is the value of Y predicted by our regression model for a known value of X and Y_i is the value of Y observed in the data. Thus, in order to calculate the standard error using (5), we must first use our linear model to predict what we expect the value of Y to be for each value of X. Next, we multiply the standard error given by (5) by the critical value of t drawn from the table of t statistic values with n-2 degrees of freedom and the desired significance level, $\alpha/2$. This value is then added to our estimate of b to find the high end of the confidence interval and subtracted from our estimate of b to find the low end of the confidence interval:

$$b - S_b t_{\alpha/2,n-2} < b < b + S_b t_{\alpha/2,n-2} \quad (6)$$

Assumptions and Limitations

- Our calculations assume a single value of Y for each value of X
- The variance of Y should be equal across X
- The variable under study should be normally distributed within both populations

Worked examples:

Correlations

Problem: As part of your graduate research you are investigating the Red Queen hypothesis for the evolution of sex. Specifically, you are interested in establishing whether or not a correlation exists between the rate of sexual reproduction and intensity of parasitism in a species of plant you study. To this end, you have collected data on the proportion of reproduction that occurs sexually (as opposed to clonally) and the intensity of infection by a pathogenic rust fungus from five different populations:

	Population				
	1	2	3	4	5
Infection intensity	0.34	0.14	0.76	0.54	0.61
Proportion sexual	0.13	0.04	0.24	0.19	0.22

Solution: In order to estimate the correlation, we need to calculate the covariance between the variables and the standard deviation of each variable. The values of these quantities are:

$$Cov(X,Y) = 0.015408$$
$$\sigma_X = 0.21637$$
$$\sigma_Y = 0.07228$$

Next, we just plug these quantities into (1), to arrive at our estimate for the correlation:

.

$$\rho = \frac{0.015408}{(0.21637)(0.07228)} = 0.9853$$

Now, in order to evaluate whether our estimated correlation differs significantly from zero, we need to use equation (2) to calculate the test statistic t:

$$t = 0.9853\sqrt{(5-2)/(1-(0.9853)^2)} = 9.98981$$

Next, compare the value of this test statistic to the critical value of t drawn from the student's t-distribution with n-2 = 3 degrees of freedom and the desired significance level, $\alpha/2$ = 0.025. According to the table of t values, this critical value is t = 3.182. Because the value we calculated for our test statistic (9.98981) greatly exceeds the critical value of t from the table (3.182), we can say with confidence that infection intensity and sexual reproduction are correlated within one another.

Linear regression

Problem: As part of your work with USFWS you have been studying the interaction between coyotes and wolves within the Greater Yellowstone Ecosystem (GYE). Your project revolves around determining the extent to which the density of wolves within a specific region predicts the abundance of coyotes within that region. Your hypothesis is that because wolves aggressively remove coyotes from their territory, coyote density will be low where wolf density is high. In addition to testing this hypothesis in a general sense, you also hope to use your study as a basis for predicting coyote densities based on information on wolf density alone. To this end, you have determined the density of wolves within five different river drainages using radio telemetry and also estimated the number of coyotes using arial surveys. Your data is shown below in the following table:

	River drainage				
	1	2	3	4	5
Wolf density	1.23	0	2.21	0.67	4.56
Coyote density	3.13	4.04	2.24	3.46	0.78

Solution: The first step in our analysis is to use equation (3) to estimate the slope of the relationship between wolf density (X) and coyote density (Y)

$$b = \frac{-1.79876}{2.521144} = -0.71347$$

Next, we can use equation (4) and our estimate of the slope to estimate the intercept

$$a = 2.73 - (-.71347)1.734 = 3.9672$$

With estimates for the slope and intercept in hand, we now have sufficient information to write down a simple expression that predicts the density of coyotes (Y) given information on the density of wolves (X):

$$\hat{Y} = 3.9672 - .71347X$$

In addition to generating this predictive relationship, however, we should also test to be certain the slope of our relationship, b, really is significantly different from zero. One way we can accomplish this goal is by placing a confidence interval around the slope. The first step in developing this confidence interval is to calculate the value of Y our model predict for the observed values of X:

	River drainage				
	1	2	3	4	5
Observed wolf density (X)	1.23	0	2.21	0.67	4.56
Observed coyote density (Y)	3.13	4.04	2.24	3.46	0.78
Predicted coyote density ($\hat{Y}$)	3.09	3.97	2.39	3.49	0.71

Next, we can use our predicted values to calculate the standard error of the slope using equation (5):

$$S_b = \sqrt{\frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{(n-2)} \Big/ \sum_{i=1}^{n}(X_i - \bar{X})^2} = \sqrt{\frac{0.0348}{(3)} / 12.61} = 0.0303$$

With the standard error in hand, we can now use equation (6) and the critical value of t for the case of n-2=3 degrees of freedom and $\alpha/2 = 0.025$, to develop our confidence interval:

$$-0.71347 - (0.0303)3.182 < b < -0.71347 + (0.0303)3.182$$

which is equal to:

$$-0.81 < b < -0.62$$

Because the confidence interval for the slope does not include zero, we can reject the null hypothesis that wolf density does not influence coyote density. Instead, we accept the alternative hypothesis that coyote density decreases with wolf density.