Regression

Garrett M Fitzmaurice

Abstract

Regression models are widely used for addressing scientific questions of interest regarding the associations among a set of variables. In particular, linear regression models describe how part of the natural individual-to-individual variation in a continuous response variable can be explained by one or more explanatory variables. In this article we provide a general overview of regression concepts, emphasizing the two most common goals of regression analysis: explanation and prediction. We discuss various aspects of interpretation of regression coefficients. We also consider the notions of confounding and interaction within regression analyses. Finally, we consider important generalizations of linear regression to handle the case where the response variable is binary (logistic regression) and also settings with correlated responses (e.g., repeated measurements on individuals over time). We conclude by discussing how linear and logistic regression are special cases of a broad and useful collection of regression models known as generalized linear models.

Keywords confounding; generalized linear models; interaction; linear regression; logistic regression; prediction

Introduction

In many experiments as well as observational studies, multiple variables are measured on individuals to address specific scientific questions of interest regarding their association. For example, we may have some histopathological variable of interest, say a quantitative measure of brain densities of senile plaques (number per square millimeter), that varies naturally from one Alzheimer's patient to another. The question of scientific interest might be whether some of that variability can be explained by other variables, such as, for example, age, gender, years of education, and scores on a measure of episodic memory such as the Rey Auditory Verbal Learning Test (RAVLT).[1,2] In this example, the density of senile plaques is referred to as the *response variable* and age, gender, years of education and RAVLT scores are referred to as *explanatory variables*; the latter variables potentially explain some of the variation in the former. Scientific questions of this nature can be answered by analysing the data using regression models.

Regression models are widely used across a range of scientific applications and provide a very general and versatile approach for describing the dependence of a response variable on a set of explanatory variables. Broadly speaking, we can distinguish two main reasons for the use of regression models. The first reason is for the purpose of *explanation*. That is, the regression model is used to estimate the effect of an explanatory variable on the response, while controlling or adjusting for the effects of many other variables that are included in the model. In many instances, the inclusion of the latter variables is to ensure that we obtain an estimate of the *unconfounded* effect of the explanatory variable of interest. For example, we may be interested in estimating the effect of a measure of episodic memory (e.g., RAVLT scores) on brain densities of senile plaques in Alzheimer's patients. However, the apparent association may well be due to the fact that patients in our sample with low scores on the RAVLT also happen to be older and with less years of education. So, in using a regression model for explanatory purposes, the regression coefficients yield an estimate of the effect of the explanatory variable of interest (e.g., episodic memory), controlling for any other factors that have been included in the regression model (e.g., age, years of education), thereby enhancing our understanding of the true relationship between the response and explanatory variable. Although this type of use of a regression model cannot by itself establish a causal explanation of the relationship between the response and an explanatory variable (the latter would be justified if, for example, individuals were randomized to levels of the explanatory variable), it may often be used to at least provide partial support of explanations that are potentially causal. The second main reason to use a regression model is for the purpose of *prediction*, e.g., prediction of the unobserved responses for new individuals or prediction of future values of the response on the basis of present values of the explanatory variables (the latter is referred to as forecasting). When used for prediction, the regression model provides an estimate of the expected or predicted values of the response as a function of the explanatory variables (we note in passing that the terms "explanatory variables" and "predictors" are often used interchangeably when discussing regression, regardless of the reason for its use). In the context of prediction, a regression model that is well calibrated should yield predicted values of the response that closely agree with the actual or realized values of the response. For example, there may be a gold-standard histopathologic measure of interest, e.g., brain densities of senile plaques in Alzheimer's patients, that can only be obtained post-mortem; the goal of the study is to find a set of biomarkers (e.g., biomarkers in saliva, blood, or cerebrospinal fluid) predictive of this post-mortem response variable that can be routinely assessed. In this example, the goal is not to obtain an estimate of the unconfounded effect of each of the biomarkers on the response variable; there is relatively little interest in interpreting their effects on the response variable. Instead, the main emphasis is on obtaining predictions of the response, as a function of the biomarkers, that closely agree with the actual values of the response variable. It is worth mentioning that these two different reasons for using a regression model, *explanation* and *prediction*, can have important implications for the decision about what variables to include and exclude from the regression model; however, that is a topic beyond the scope of this article.

Finally, our use of the term "regression model" in this article is not strictly limited to the standard linear regression model for a continuous response variable. Instead, we use this term more broadly to refer to any model that describes the dependence of the mean of a response variable on a set of explanatory variables in terms of some form of regression equation. We will begin our discussion of regression models with the simplest case: the linear

Garrett M Fitzmaurice *Sc.D Professor of Biostatistics and Director of the Laboratory for Psychiatric Biostatistics, McLean Hospital, Belmont, MA; Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. Conflict of interest: none.*

regression model for a continuous response variable; later, we briefly consider some of the many possible generalizations. With the exception of continuous responses, binary data taking on only two values (perhaps denoting absence or presence of disease) are by far the most commonly encountered data type in medical studies. For a binary response variable, linear logistic regression is widely used in many applications. Another important generalization is to observations that cannot be assumed to be statistically independent of one another, that is, regression models for correlated responses. Correlated responses might arise when multiple measurements are obtained on a single tumour from a patient or when repeated measurements are obtained over time. In later sections of this article we consider both kinds of generalizations of the standard linear regression model.

Linear regression

Linear regression describes how the mean values of a response variable (denoted by Y) vary as a linear function of a set of explanatory variables. For ease of exposition, we will first consider the case where there is only a single explanatory variable, denoted by X; when restricted to a single explanatory variable the model is often referred to as *simple* linear regression. The generalizations to more than one explanatory variable will be considered later.

Assuming there are observations on n individuals ($i = 1, \ldots, n$) in the study, the simple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_i + e_i,$$

where the β's are the regression coefficients and e_i is a random component that is assumed to be independently normally distributed with zero mean and variance σ^2. The e_i term in the model represents natural variation of Y_i among individuals around the mean (or expected value) of the response in the population,

$$E(Y_i) = \beta_0 + \beta_1 X_i.$$

The regression coefficients, β_0 and β_1, express the linear dependence of the mean response on the explanatory variable. Specifically, the *intercept* β_0 has interpretation as the mean value of the response variable when the explanatory variable X is equal to zero. The *slope*, β_1, is usually the parameter of most interest and has interpretation as the change in the mean of Y for a single unit increase in X. For the special case where the explanatory variable X is dichotomous, taking values of 0 and 1, the regression slope β_1 has a simple interpretation as the difference in the mean of Y when $X = 1$ versus $X = 0$.

When there is more than a single explanatory variable, an important generalization of the model is referred to as *multiple* linear regression. Although it is relatively straightforward to assess how the response variable is associated with each of the explanatory variables, when taken one at a time, in a simple linear regression, this is usually not satisfactory for the following two reasons. First, a series of separate analyses does not permit an assessment of how well the combined set of explanatory variables predict the response variable. Second, by failing to consider the simultaneous effects of the explanatory variables, we cannot estimate the unconfounded effect of any particular explanatory variable. Multiple linear regression overcomes both of these limitations by simultaneously estimating the effects of the set of explanatory variables on the response. In doing so, we can determine how closely the predicted values, based on the entire set of explanatory variables, agree with the actual values of the response; in addition, we can discern the effect of any particular explanatory variable, controlling or adjusting for the other variables included in the model.

Assuming there are p explanatory variables, say $X_1, \ldots, X_p$, obtained on n individuals ($i = 1, \ldots, n$), the multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + e_i,$$

where the e_i is a random component, assumed to be independently normally distributed with zero mean and variance σ^2, representing natural variation of Y_i around the mean of the response,

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}.$$

The β's express the dependence of the response variable on the explanatory variables. The interpretation of any particular regression coefficient, say the slope β_1, is in terms of the change in the mean of Y for a single unit increase in X_1, *while the remaining explanatory variables* $X_2, \ldots, X_p$ *are held constant*. That is, the interpretation of β_1 can best be understood as a hypothetical comparison of two groups of individuals that differ by one unit in X_1 but have the same values for all of the other explanatory variables, $X_2, \ldots, X_p$. The intercept, β_0, now has interpretation as the mean value of the response variable when all of the explanatory variables assume the value zero, i.e., when $X_1 = X_2 = \ldots = X_p = 0$.

Before we discuss some additional aspects of the model it is worth emphasizing that the term "*linear*" in multiple linear regression has a very precise meaning and refers to the fact that all models for the mean response are *linear in the regression parameters (coefficients)*. That is, the right hand side of the regression equation can always be constructed by adding together terms that are either a constant (say β_0) or the product of a regression coefficient and an explanatory variable (e.g., $\beta_1 X_{1i}$). For example, the following three models for the dependence of the mean response on X,

$$E(Y_i) = \beta_0 + \beta_1 X_i,$$

$$E(Y_i) = \beta_0 + \beta_1 \log(X_i),$$

$$E(Y_i) = \beta_0 + \beta_1 X_i + \beta_2 X_i^2,$$

are all cases where the mean is said to be *linear in the regression parameters*, even if the latter two models are non-linear in the explanatory variable X whereas the first model defines a straight line (or linear) relationship. It becomes more transparent that, for example, the third model is linear in the regression parameters when the model is re-written as

$$E(Y_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$$

where $X_{1i} = X_i$, $X_{2i} = X_i^2$. Thus, in linear regression the mean response can certainly change as a non-linear or curvilinear

