

Appendix E

Logistic Regression for Stream Stability

LOGISTIC REGRESSION FOR STREAM STABILITY

This appendix briefly describes how logistic regression is applied to stream stability issues. An interested reader is referred to *Statistical Methods in Water Resources* (Helsel D.R. and Hirsch R.M., 1992).

Introduction

One of the goals of the HMP project is to predict stream stability based on measurable variables such as velocity, shear stress, work done to the streams, and erosion potential (EP). Stream stability, categorized into stable or unstable, is a qualitative and discrete variable, whereas the measurable variables are quantitative and continuous. Logistic regression, also called **logit** regression, is an appropriate statistical tool for this situation.

Within the framework of logistic regression used in this project, stream stability is characterized as a binary response variable, expressing itself as either 0 – stable, or 1 – unstable, and the measurable variables as continuous explanatory variables. Given a value of one of the explanatory variables, logistic regression provides the probability of the stream stability being stable or unstable. Note that, although multiple logistic regression (i.e., multiple explanatory variables) is also possible, in this project, only simple logistic regression (i.e., one explanatory variables) will be considered.

Important Formulae

Let p be the probability of a stream being unstable. The valid range of value of p is between 0 and 1. Odds ratio is defined as:

$$\text{Odds ratio} = \frac{p}{1-p}$$

The natural log of the odds ratio is called **logit**. Thus,

$$\text{Logit} = \ln \left[\frac{p}{1-p} \right]$$

At this point, p , whose range is 0 to 1, has been transformed into the logit, which is continuous and unbounded, and thus applicable to linear regression. Logistic regression, or **logit** regression, is essentially simple linear regression relating a continuous explanatory variable to the stream stability logit. In other words, it seeks optimal values of the intercept (b_0) and slope coefficient (b) in the following equation:

$$\ln \left[\frac{p}{1-p} \right] = b_0 + bx$$

where x is a continuous explanatory variable (e.g., velocity, shear stress, work done to the streams, or erosion potential (EP)).

Therefore, p can be predicted for a given value of the explanatory variable x , and is expressed as:

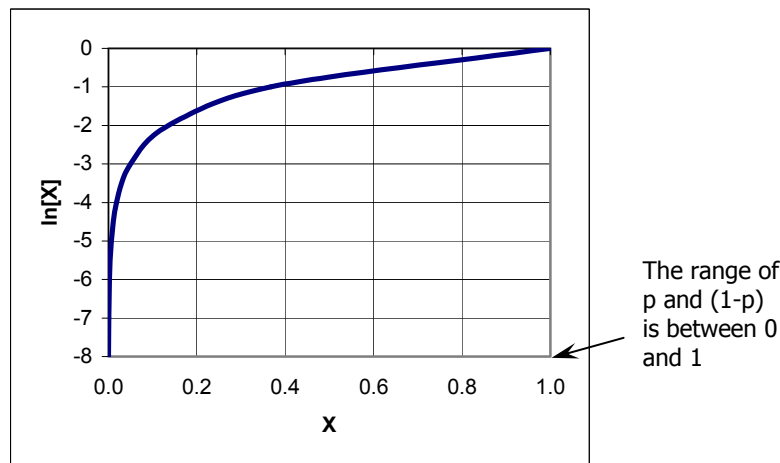
$$p = \frac{\exp(b_0 + bx)}{1 + \exp(b_0 + bx)}$$

Optimal values of b_0 and b are obtained when log likelihood (l) is maximized (see the equation below). Note that l is always negative; maximizing l is thus bringing it closet to zero.

$$l = \sum_{i=1}^n (y_i \cdot \ln[p_i] + (1 - y_i) \cdot \ln[1 - p_i])$$

The following is the rationale behind the use of this parameter as the optimality indicator. A good logistic regression model would predict a value closer to 1 for p_i when the field observation indicates that the stream is unstable ($y_i=1$), and a value closer to 0 when the stream is stable ($y_i=0$). According,

- When the stream is **unstable**, $y_i=1$, the second term becomes zero, and only the first term contributes to l . In this case, if p_i is **high** – which is **desirable**, $\ln[p_i]$ will be a **small negative number**, closer to 0; In contrary, if p_i is low – which is undesirable, $\ln[p_i]$ will be a large negative number.
- When the stream is **stable**, $y_i=0$, the first term becomes zero, and only the second term contributes to l . In this case, if p_i is high – which is undesirable, $\ln[1-p_i]$ will be a large negative number; In contrary, if p_i is **low** – which is **desirable**, $\ln[1-p_i]$ will be a **small negative number, closer to 0**.



Therefore, the best logistic regression model, associated with optimal b_0 and b , results in the value of l closet to zero, i.e., being maximized. The log likelihood (L) may be alternately reported as the positive number G^2 , $-2 \cdot \log$ likelihood, which is to be minimized to achieve the optimal b_0 and b :

$$G^2 = -2 \cdot L$$

Hypothesis Testing

There are several types of hypothesis testing applicable to logistic regression. However, as only simple (one explanatory variable) logistic regression is being considered, only the test for overall

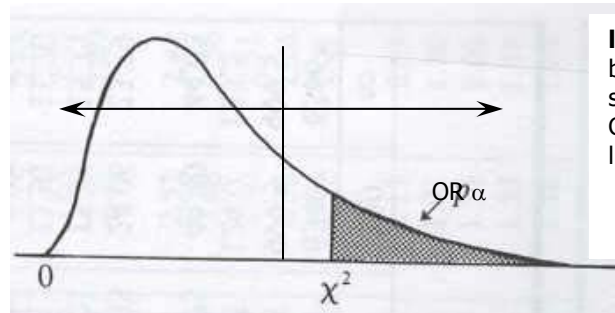
significance is relevant. In this case, the null hypothesis is that the logistic regression model with the slope coefficient b is not significantly better than an intercept-only model (where $b = 0$, i.e., constant p). The statistic parameter used in this test is overall likelihood ratio (lr_o):

$$lr_o = (G_0^2 - G^2)$$

where G^2 is $-2 \cdot \log$ likelihood of the non-zero- b model and G_0^2 is that of the intercept-only model.

The overall likelihood lr_o can be approximated by a chi-square distribution with k degrees of freedom, where k is the number of slopes estimated (Helsel and Hirsch, 1992). In this case, $k = 1$. Consequently, if $lr_o > \chi^2_{1,\alpha}$, the null hypothesis can be rejected: b is not zero. In contrary, if the null hypothesis cannot be rejected, the best estimate over all values of the explanatory variable of p is simply the proportion of the data set that $y = 1$.

$lr_o < \chi^2_{k,\alpha}$:
 b does NOT differs from zero at a significance level of α ;
 G^2 is high;
 l is low (a large negative number)



$lr_o > \chi^2_{k,\alpha}$:
 b differs from zero at a significance level of α ;
 G^2 is low;
 l is high (close to zero)