

Chapter 15

Regression for Discrete Responses

Concentrations of a volatile organic chemical are measured in numerous wells across a large study area. About 75% of the resulting samples are below the laboratory reporting limit. The likelihood of finding concentrations above this limit is suspected to be a function of several variables, including population density, industrial activity, and traffic density. What is the most appropriate way to model the probability of being above the reporting limit using a regression-like relationship?

Streams can be classified according to whether or not they meet some criteria for use set by a regulatory agency. For example, a stream may be considered "fishable" or "not fishable", depending on several concentration and esthetic standards. What is the probability that a stream reach will meet the "fishable" criteria as a function of population density, distance downstream from the nearest point source, and percentage of the basin used for crop agriculture?

The above situations involve fitting a model similar to OLS regression, in that the explanatory variables are continuous. However the response variable is discrete -- it can be designated by an integer value (see figure 4.1). Discrete (or categorical) response variables are often encountered when the measurement process is not sufficiently precise to provide a continuous scale. Instead of an estimate of concentration, for example, only whether or not a sample exceeds some threshold, such as a reporting limit or health standard, is recorded. In water resources this response is usually ordinal. Logistic regression is the most commonly-used procedure for this situation. The equation predicts the probability of being in one of the possible response groups.

Discrete response variables are commonly binary (two categories). For example, species of organism or attribute of an organism are listed as either present or absent. Analysis of binary responses using logistic regression is discussed in the following sections, beginning with 15.1. Analysis of multiple response categories is discussed in section 15.4.

15.1 Regression For Binary Response Variables.

With OLS regression, the actual magnitude of a response variable is modelled as a function of the magnitudes of one or more continuous explanatory variables. When the response is a binary categorical variable, however, it is the probability p of being in one of the two response groups that is modelled. The response variable is coded by setting the larger of the two possible responses (above or present) equal to 1, and the lower to 0. The predicted probability p is then the probability of the response being a 1, with $1-p$ as the probability of the response being a 0. The explanatory variables may be either continuous as in OLS regression, or a mixture of continuous and discrete variables similar to analysis of covariance. If all explanatory variables are discrete, logistic regression provides a multivariate alternative to the test for significance by Kendall's tau used in Chapter 14.

15.1.1 Use of Ordinary Least Squares

In the case of a binary response, the attempt to predict \hat{p} = the probability of a response of 1 could be done with OLS regression. This would be a simple but incorrect approach. There are three reasons why this is not appropriate (Judge, et al., 1985):

1. Predictions \hat{p} may fall outside of the 0 to 1 boundary.
2. The variance of \hat{p} is not constant over the range of x 's, violating one of the basic assumptions of OLS. Instead, the variance of the binary response variable equals $p \cdot (1-p)$, where p is the true probability of a 1 response for that x . Because this is not constant over x , weighted least squares must be used to obtain minimum variance and unbiased estimates of slope and intercept. See Draper and Smith (1981, pp. 108-116) for the WLS approach. WLS is still not appropriate, however, if estimates go beyond the 0 to 1 boundary.
3. Residuals from the regression cannot be normally distributed. This renders tests on the slope coefficients invalid.

OLS been used with discrete responses when multiple observations occur for all or most combinations of explanatory variables. The responses (0 or 1) are first grouped by some range of explanatory variable(s). This creates a new continuous y variable, the proportion of responses which equal 1. Even so, least-squares regression fails the three criteria above, so that more appropriate methods are warranted.

15.2 Logistic Regression

Logistic regression, also called logit regression, transforms the estimated probabilities \hat{p} into a continuous response variable with values possible from $-\infty$ to $+\infty$. The transformed response is predicted from one or more explanatory variables, and subsequently retransformed back to a value between 0 and 1. A plot of estimated probabilities has an S shape (figure 15.1). The estimates of probability change most rapidly at the center of the data. Thus logistic regression is most applicable for phenomena which change less rapidly as p approaches its limits of 0 or 1. However, when the range of predicted probabilities does not get near its extremes, the plot is one of mild curvature (figure 15.2). Thus the function is a flexible and useful one for many situations. A review of this and other categorical response models is given by Amemiya (1981).

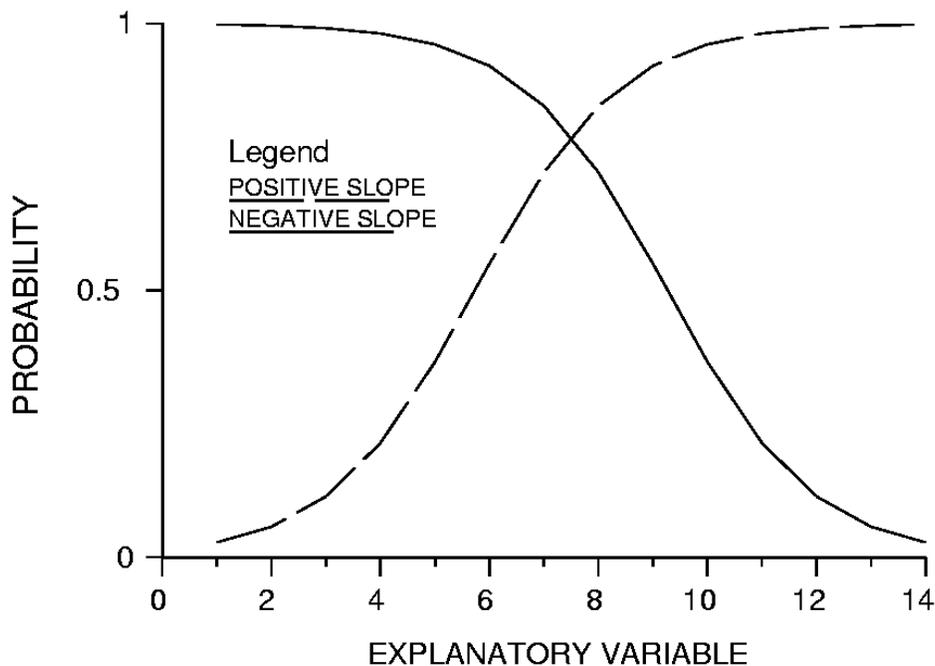


Figure 15.1 Logistic regression equations with $-$ and $+$ slopes. Note that estimates change more rapidly in the center than at the extremes.

15.2.1 Important Formulae

The **odds ratio** is defined as the ratio of the probability of obtaining a 1 divided by the probability of obtaining a 0:

$\text{odds ratio} = \left(\frac{p}{1-p} \right)$	[15.1]
--	--------

where p is the probability of a response of 1.

The log of the odds ratio or **logit** transforms a variable constrained between 0 and 1, such as a proportion, into a continuous and unbounded variable. The logit can then be modeled as a linear function of one or more explanatory variables to produce logistic regression:

$$\log\left(\frac{p}{1-p}\right) = b_0 + \mathbf{bX} \quad [15.2]$$

where b_0 is the intercept, \mathbf{X} is a vector of k explanatory variable(s), and \mathbf{bX} includes the slope coefficients for each explanatory variable so that $\mathbf{bX} = b_1X_1, b_2X_2, \dots, b_kX_k$.

Thus the odds ratio is modelled as

$$\left(\frac{p}{1-p}\right) = \exp(b_0 + \mathbf{bX}). \quad [15.3]$$

To return the predicted values of the response variable to original units, the logistic transformation (the inverse of the logit transformation) is used:

$$p = \frac{\exp(b_0 + \mathbf{bX})}{[1 + \exp(b_0 + \mathbf{bX})]} \quad [15.4]$$

For example, the multiple logistic regression equation with two explanatory variables would look like

$$p = \frac{\exp(b_0 + b_1X_1 + b_2X_2)}{[1 + \exp(b_0 + b_1X_1 + b_2X_2)]}$$

For a single x variable, the odds of obtaining a 1 response increase multiplicatively by e^{b_1} for every unit increase in X . The inflection point of the curve is at $-b_0/b_1$, which is the median of the data. The slope of the estimated probability is greatest at this point. Equations are analogous for multiple explanatory variables. Biologists call the inflection point the median lethal dose (LD_{50}) when predicting the probability of death from some concentration (dose) of toxicant. The animal has a 50% chance of survival at this dose.

15.2.2 Computation by Maximum Likelihood

Estimates b_j of the $j=1, \dots, k$ slope coefficients could physically be computed by WLS when the input data are proportions between 0 and 1 (but they should not -- see section 15.1.1).

However, the original data are most often coded only in the binary form, with replicates not available for computing proportions. A more general method for computing slope coefficients, valid for both binary and proportions as input data, is maximum likelihood estimation.

Maximum likelihood optimizes the likelihood that the observed data would be produced from a given set of slopes. It is an iterative procedure available in the more complex statistical software packages. A function called the **log likelihood** (l) of the overall regression model is written as:

$$l = \sum_{i=1}^n \left(y_i \cdot \ln[\hat{p}_i] + (1-y_i) \cdot \ln[1-\hat{p}_i] \right) \quad [15.5]$$

for the $i=1, n$ binary observations y_i and predicted probabilities \hat{p} . When $y_i = 0$, the second term inside the brackets is nonzero, and a \hat{p} is desirable which is close to 0. When $y_i = 1$, the first term is nonzero and a \hat{p} close to 1 is desirable. The log of either \hat{p} or $[1-\hat{p}]$ will be negative, and therefore l is a negative number which is maximized (brought closest to 0) by iteratively substituting estimates of p derived from estimates of slopes and intercept. The log likelihood may be alternately reported as the positive number G^2 , the **-2 log likelihood**, which is minimized by the MLE procedure:

$$-2 \log \text{likelihood } G^2 = -2 \cdot l. \quad [15.6]$$

15.2.3 Hypothesis Tests

15.2.3.1 Test for overall significance

An overall test of whether a logistic regression model fits the observed data better than an intercept-only model (where all slopes $b_j = 0$), analogous to the overall F test in multiple regression, is given by the **overall likelihood ratio** (lr_o):

$$lr_o = 2 \cdot (l - l_0) = (G^2_0 - G^2) \quad [15.7]$$

where l is the log likelihood of the full model, l_0 is the log likelihood of the intercept-only model, and G^2_0 is the $-2 \log$ likelihood of the intercept only model.

The overall likelihood ratio lr_o can be approximated by a chi-square distribution with k degrees of freedom, where k is the number of slopes estimated. If $lr_o > \chi^2_{k, \alpha}$ then the null hypothesis that all $b_j = 0$ can be rejected. Should the null hypothesis not be rejected, the best estimate over all \mathbf{X} of the probability of a 1 is simply the proportion of the entire data set which equals 1.

15.2.3.2 Testing nested models

To compare nested logistic regression models, similar to the partial F tests in OLS regression, the test statistic is the **partial likelihood ratio** lr :

$$lr = 2 \cdot (l_c - l_s) = (G^2_s - G^2_c) \quad [15.8]$$

where l_c is the log likelihood for the more complex model, and l_s is the log likelihood for the simpler model.

The partial likelihood ratio is approximated by a chi-square distribution with $(k_c - k_s)$ degrees of freedom, the number of additional coefficients in the more complex model. For the case where only one additional coefficient is added, the chi-square with 1 degree of freedom equals the

square of a t-statistic called **Wald's t**, computed from the estimated coefficient b divided by its standard error. Degrees of freedom for the t-statistic are the number of observations n minus the number of estimated slopes, or $n-k$. As with OLS regression, some computer software will report the t-statistic, while others report the $t^2 = \chi^2$ value; p-values will be essentially the same for either form of the test.

15.2.4 Amount of Uncertainty Explained, R^2

A measure of the amount of uncertainty explained by the model, actually the proportion of log-likelihood explained, is McFadden's R^2 , or the **likelihood- R^2** ,

$$R^2 = 1 - \frac{l}{l_0} \quad [15.9]$$

where l and l_0 are as before. The likelihood- R^2 is uncorrected for the number of coefficients in the model. much like R^2 in OLS regression.

A second measure of the amount of uncertainty explained by the model is the R^2 between the observed and predicted values of p , or **Efron's R^2**

$$\text{Efron's } R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{p})^2}{\sum_{i=1}^n (y_i - \bar{p})^2} \quad [15.10]$$

where $\bar{p} = \sum y_i/n$, the proportion = 1 for the entire data set. However, this version of R^2 is not as appropriate as the likelihood- R^2 because the residuals $(y_i - \hat{p})$ are heteroscedastic due to the binary nature of the y_i .

15.2.5 Comparing Non-Nested Models

To compare two or more non-nested logistic regression models, partial likelihood ratios are not appropriate. This is the situation in OLS regression where Mallows's C_p or PRESS is used. For likelihood ratio tests, a statistic related to Mallows's C_p is **Akaike's Information Criteria (AIC)**. AIC includes both a measure of model error ($-l$) and a penalty for too many variables, the number of explanatory variables k . Better models are those with small AIC. Akaike's information criteria

$$\text{AIC} = -l + k \quad [15.11]$$

AIC can also be written to expressly include the comparison of each candidate model to the full model (the model which includes all possible explanatory variables).

$$\begin{aligned} \text{AIC}^* &= 2(l_f - l) - 2 \cdot (k_f - k) \\ &= (G^2 - G^2_f) - 2 \cdot \Delta df \\ &= lr - 2 \cdot \Delta df \end{aligned}$$

where l_f is the log likelihood of the full model, k_f is the degrees of freedom of the full model, Δdf is the difference in the degrees of freedom between the model and the full model, and lr is the partial likelihood ratio comparing the candidate and full models. Either form should be minimized to find the best model.

Related to the AIC is an **adjusted R^2** which adjusts for the degrees of freedom in the model. It penalizes a model which includes too many slope parameters. The adjusted R^2 allows comparisons between models with differing number of explanatory variables:

$$\text{adjusted } R^2 = 1 - \frac{(1 - k)}{l_0} = 1 - \frac{2 \cdot \text{AIC}}{G^2_0} \quad [15.12]$$

This adjusted R^2 should be maximized.

Example 1

Eckhardt et al. (1989) reported the pattern of occurrence for several volatile organic compounds in shallow groundwaters on Long Island, NY. TCE detections for 643 samples are listed in table 15.1 below, where 1 signifies a concentration above the reporting limit of 3 ppb. Logistic regression between occurrence (1) or non-occurrence (0) as a function of population density gives the following results:

Population Density	no. 1s	no. 0s	N	%1s
1	1	148	149	0.7
2	4	80	84	4.8
3	10	88	98	10.2
5	25	86	111	22.5
6	11	33	44	25.0
8	8	24	32	25.0
9	29	14	43	67.4
11	19	31	50	38.0
13	6	5	11	54.5
14	2	11	13	15.4
17	2	5	7	28.6
19	<u>0</u>	<u>1</u>	<u>1</u>	<u>0.0</u>
overall	117	526	643	18.2

Table 15.1 TCE data in the Upper Glacial Aquifer, Long Island

The log likelihood for the intercept-only model $l_0 = -305.0$ ($G^2_0 = 610.0$). To determine the significance of population density (POPDEN) as an explanatory variable, the likelihood ratio is

computed by subtracting the log likelihood of this one-variable model from that of intercept-only model, and comparing to a chi-square distribution:

$$lr = 610.0 - 533.0 = 77.0 \quad \text{with 1 df resulting in a p-value} = 0.0001.$$

Table 15.2 gives the important statistics for the model. A plot of the logistic regression line along with bars of ± 2 standard errors are shown in figure 15.2.

$$-2 \log \text{likelihood} = 533.0$$

Explanatory variable	Estimate	Partial t-statistic	p-value
INTERCEPT	-2.80	-13.4	0.0001
POP DEN	0.226	8.33	0.0001

Table 15.2 Statistics for the popden model.

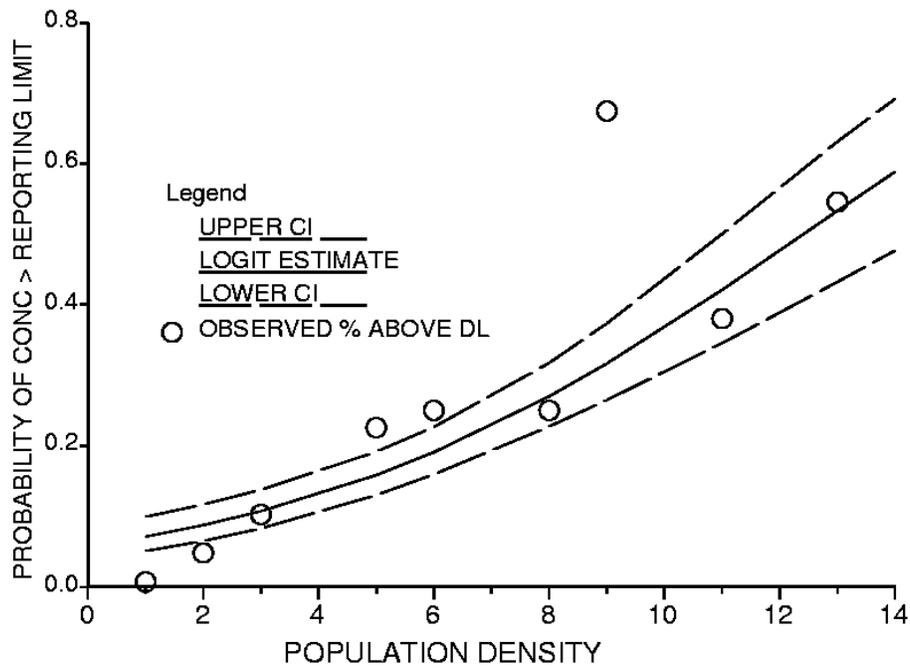


Figure 15.2 Logistic regression line for the TCE data, with percent detections observed for each population density.

The positive slope coefficient for popden means that the probability of a response = 1 (concentration above the reporting limit) increases with increasing population density. Note that the line did not fit the observed data well at popden = 9. A second variable, a binary indicator of whether or not the area around the well was sewered, was added to the model in hopes of improving the fit. Does this second variable help explain more of the variation observed? The results are presented in table 15.3.

-2 log likelihood = 506.3

<u>Explanatory variable</u>	<u>Estimate</u>	<u>Partial t-statistic</u>	<u>p-value</u>
INTERCEPT	-3.24	-12.47	0.0001
POPDEN	0.13	4.07	0.0001
SEWER	1.54	4.94	0.0001

Table 15.3 Statistics for the popden + sewer model.

The likelihood ratio test determines whether this model is better than an intercept-only model
 $lr_O = 610.0 - 506.3 = 103.7$ with 2 df resulting in a p-value = 0.0001.

Thus this logistic regression is significantly better than just estimating the proportion of data above the detection limit without regard to the two variables. The positive slope estimate for sewer means that the probability of detection of TCE increases with increasing proportion of sewerage around the well. Note that this does not prove that sewerage itself is the cause -- this could result from sewerage as a surrogate for increasing urbanization or industrialization of the area. The usefulness of sewer in comparison to the popden-only model is seen by the significance of its partial t-statistic. It may also be measured by the difference in likelihood ratios for the one and two-variable models:

$$lr = 533.0 - 506.3 = 26.7 \quad \text{with 1 df resulting in a p-value} = 0.0001.$$

Next a model with completely different explanatory variables was tried, relating TCE detections to the amount of land near the well which was classified as industrial land (indlu), and to the depth of the water below land surface. The results are given in table 15.4. As the partial t-statistics are both significant, a logical question is which of the two 2-variable models is preferable?

-2 log likelihood = 557.8

<u>Explanatory variable</u>	<u>Estimate</u>	<u>Partial t-statistic</u>	<u>p-value</u>
INTERCEPT	-1.07	-5.49	0.0001
INDLU	0.092	4.61	0.0001
DEPTH	0.008	-4.52	0.0001

Table 15.4 Statistics for the indlu + depth model.

As these models are not nested, they must be compared using AIC. Magnitudes of their partial t-statistics will not help decide which to use. As seen in table 15.5, the AIC for the population+sewer model is lower, and therefore is the preferable model between these two candidates.

Explanatory variables	-l	k (# exp. vars.)	AIC
POPDEN, SEWER	253.2	2	255.2
INDLU, DEPTH	278.9	2	280.9

Table 15.5 AIC for comparing two 2-variable logit models.

15.3 Alternatives to Logistic Regression

Two other methods have been used to relate one or more continuous variables to a binary variable.-- discriminant function analysis (parametric), and the nonparametric rank-sum test. In the following sections these methods are compared to logistic regression.

15.3.1 Discriminant Function Analysis

Discriminant function analysis is used as a multivariate classification tool, to decide in which of several groups a response is most likely to belong (Johnson and Wichern, 1982). Probabilities of being in each of the groups is computed as a function of one or more continuous variables. The group having the highest probability is selected as the group most likely to contain that observation. An equation (the discriminant function) is computed from data classified into known groups, and used to classify additional observations whose group affiliation is unknown. As each group is assigned an integer value, these objectives are identical to those of logistic regression.

The primary drawback of discriminant analysis is that it makes two assumptions:

1) multivariate normality, and 2) that the variance of data within each group is identical for all groups. Thus it requires the same assumptions as does a t-test or analysis of variance, but in multiple dimensions when multiple explanatory variables are employed. It will be slightly more efficient than logistic regression if these assumptions are true, but is much less robust (Press and Wilson, 1978). Therefore logistic regression should be preferred when multivariate normality and equality of variances cannot be assumed, as is the case for most of the data found in water resources.

15.3.2 Rank-Sum Test

Dietz (1985) has shown that the rank-sum test is a powerful alternative to the more complicated likelihood-ratio test for determining whether a binary response variable is significantly related to one continuous explanatory variable. The responses of 0 and 1 are treated as two separate groups, and the ranks of the continuous variable are tested for differences among the two response groups. When the probabilities of a 0 or 1 differ as a function of x, the ranks of x will differ between the two response variable groups. A slight modification to the rank-sum test is necessary for small sample sizes (see Dietz, 1985). The rank-sum test is equivalent to the significance test for Kendall's tau between the binary y variable and a continuous x.

When software is not available to perform likelihood-ratio tests, the rank-sum test can be used with little loss in power. However, it only considers the influence of one explanatory variable. There also is no slope estimate or equation associated with the rank-sum test when the responses are recorded as 0 or 1. When the responses are proportions between 0 and 1, Kendall's robust line may be used to linearly relate logits to the explanatory variable, though estimates below 0 or above 1 may result.

15.4 Logistic Regression for More Than Two Response Categories

In water resources applications, response variables may often be discretized into more than two response categories. Extensions of logistic regression for binary responses are available to analyze these situations. The method of analysis should differ depending on whether the response variable is ordinal or simply nominal. Ordinal responses such as low, medium and high are the most common situation in water resources. Here a common logit slope is computed, with multiple thresholds differing by offset intercepts in logit units. When responses are not ordinal, the possible response contrasts -- such as the probabilities of being in group 1 versus group 2 and in group 2 versus 3 -- are independent. In this case independent logit models are fit for each threshold.

15.4.1 Ordered Response Categories

Categorical response variables sometimes represent an underlying continuous variable which cannot be measured with precision sufficient to provide a continuous scale. For example, concentration data may be discretized into above and below a detection limit, or into three categories based on two thresholds (see below). Biologic activity may be categorized as not affected, slightly affected or severely affected by pollution. The resulting multiple responses y_i , $i=1$ to m are ordinal, so that $y_1 < y_2 < \dots < y_m$.

For example, suppose 3 responses are possible:

- 0: concentrations are below the reporting limit,
- 1: concentrations are above the reporting limit but below a health standard, and
- 2: concentrations are above the health standard.

This corresponds to two thresholds, one below versus above the reporting limit (0 versus not 0) and the second below versus above the health standard (not 2 versus 2). Figure 15.3 shows that for $y=2$, a transformation of the underlying continuous concentration Y^* can be developed such that $y=2$ only when $X > Y^*$ for one explanatory variable X . Similarly, $y > 0$ (above the reporting limit) only when $X > Y^* - \delta$, where δ is the difference between the two thresholds in the transformed scale. Therefore the upper threshold can be modeled as:

$$\log \left(\frac{\text{Prob}(y=2)}{\text{Prob}(y=0)+\text{Prob}(y=1)} \right) = \text{Prob}(X > Y^*) = b_0 + b_1 X, \quad [15.13]$$

where b_0 is the estimate of intercept and b_1 the estimate of slope. This is a standard logistic regression identical to the binary case of not 2 versus 2. The probability of being above the lower threshold (reporting limit) is modelled using

$$\begin{aligned} \log \left(\frac{\text{Prob}(y=1)+\text{Prob}(y=2)}{\text{Prob}(y=0)} \right) &= \text{Prob}(X > Y^* - \delta) = b_0 + b_1(X + \delta), & [15.14] \\ &= b_0' + b_1 X \\ &= b_0 + \lambda + b_1 X \end{aligned}$$

where $\lambda = b_1 \delta$ is a shift parameter that must be estimated (McCullagh, 1980). Because the responses are ordered, the slope b_1 is common to all thresholds, and represents the proportional effect of X on the underlying and unobserved Y^* . The resulting s-shaped curves for each threshold are simply offset (figure 15.4). Unfortunately the method for efficiently estimating these parameters is not available on many commercial statistics packages. McCullagh (1980) discusses the mathematics. As an alternative, separate logistic regressions can be estimated for each threshold (see below). This procedure is less efficient for the case of ordered responses, being appropriate for nominal responses. Unfortunately, it is the best that is available to most practitioners.

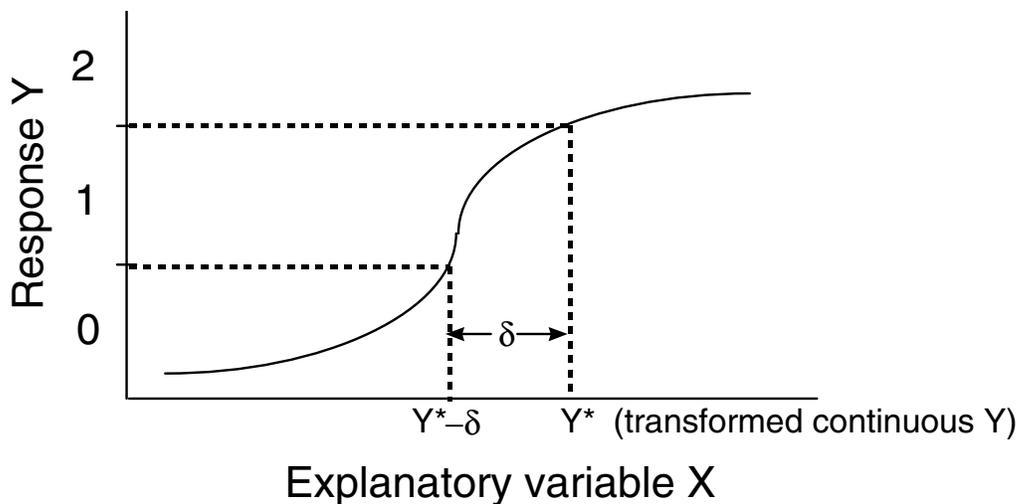


Figure 15.3 Diagram of continuous variable Y^* underlying a discrete response variable

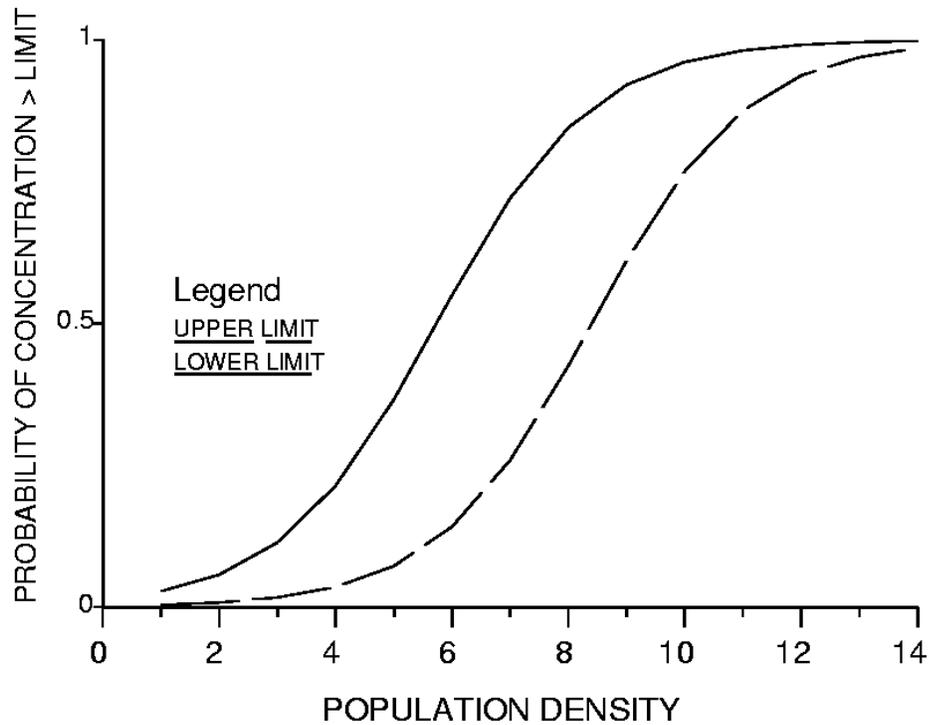


Figure 15.4 Offset logistic curves for an ordered response variable.

15.4.2 Nominal Response Categories

For the situation where there is no natural ordering of the possible response categories, an independent logistic regression must be performed for each possible contrast. Thus if there are m response categories there must be $m-1$ logistic regressions performed. Coefficients of intercept and slope are estimated independently for each. The econometrics literature has treated this situation in depth -- see for example Maddala (1983). Econometrics categories are often ones of choice -- to purchase one product or another, etc. Examples of unordered variables for water resources applications are not as obvious. However, an understanding of the equations appropriate for nominal responses is important, because these are used when most commercial software is employed to perform logistic regression of ordinal responses.

When independent logistic regressions are computed to determine the likelihood of being below versus above adjacent pairs of categories, no requirement of constant slope across thresholds is made. The probabilities employed may take several forms, but the easiest to interpret are logits of the cumulative probabilities of being below versus above each of the $m-1$ thresholds

$$\log \left(\frac{\sum \text{Prob}(y > i)}{\sum \text{Prob}(y \leq i)} \right), \quad i = 1 \text{ to } m-1. \quad [15.15]$$

These are called **cumulative logits**, as discussed by Agresti (1984) and Christensen (1990).

For the situation of $m=3$ ordered responses (0, 1, and 2) corresponding to two thresholds (reporting limit and health standard), $m-1$ or two logistic regressions must be performed. One equation determines the probability of being at least 1 -- the probability of being above the reporting limit:

$$L_1 = \log \left(\frac{\text{prob}(y=1) + \text{prob}(y=2)}{\text{prob}(y=0)} \right) = b_0 + b_1 X . \quad [15.16]$$

A second equation describes the probability of being at least 2 -- the probability of being above the health standard:

$$L_2 = \log \left(\frac{\text{prob}(y=2)}{\text{prob}(y=0) + \text{prob}(y=1)} \right) = b_0' + b_2 X . \quad [15.17]$$

Together, these two equations completely define the three probabilities as a function of the k explanatory variables X .

Example 1, cont.

Suppose a second threshold at 10 $\mu\text{g/L}$ were important for the TCE data of Eckhardt et al. (1989). This could represent an action limit, above which remedial efforts must be taken to clean up the water before use. Separate logistic regressions were performed for the probabilities of being above the 3 $\mu\text{g/L}$ reporting limit and the 10 $\mu\text{g/L}$ action limit. A new binary response variable, 0 if TCE concentrations were below 10 and 1 if above, was regressed against population density. The results are reported in table 15.6, and the curves plotted in figure 15.5. Note that the two curves are not simply offsets of one another, but have differing slopes. This situation could be viewed as an interaction, where the rate of increase in probability with unit X is not the same for the two thresholds.

Response category	b_0	b_1	lr_0
Above 3 $\mu\text{g/L}$ report. limit	-2.80	0.226	77.0
Above 10 $\mu\text{g/L}$ action limit	-3.37	0.164	23.9

Table 15.6 Independent logistic regressions for two TCE thresholds.

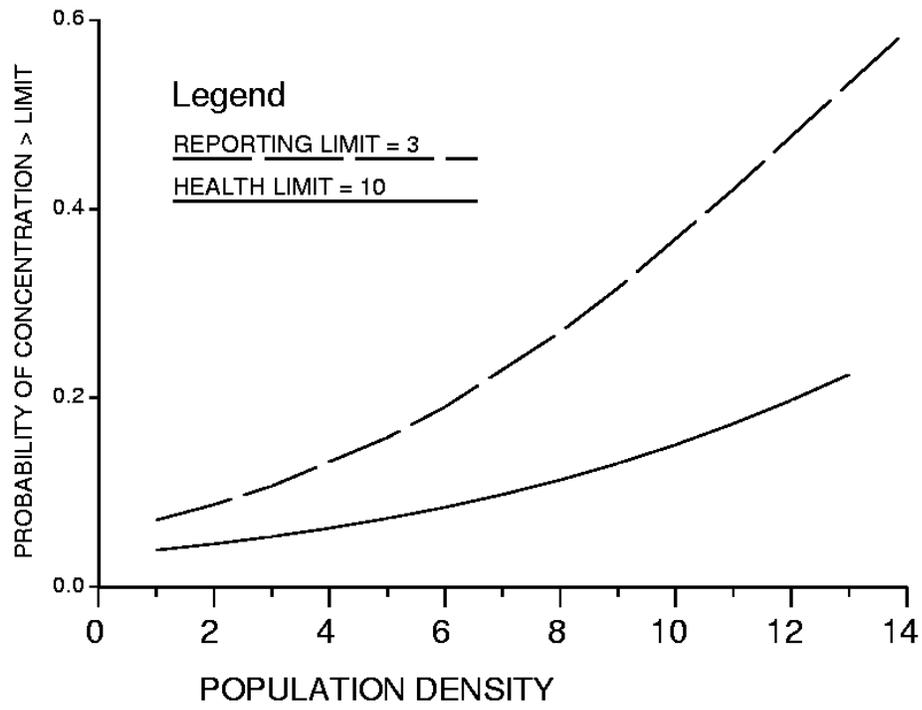


Figure 15.5 Independent logistic curves for two TCE thresholds.

Exercises

- 15.1 Person and others (1983) evaluated the ability of four factors to predict whether a surface impoundment was contaminated or not. Of particular interest was which of the four factors, information for which must be collected in other areas in the future, showed ability to predict contamination. The factors were:

<u>Factor</u>	<u>Possible scores</u>
Unsaturated Thickness	0 (favorable) to 9 (unfavorable)
Yields: aquifer properties	0 (poor) to 6 (good)
Groundwater Quality	0 (poor) to 5 (excellent)
Hazard Rating for Source	1 (low) to 9 (high)

Each impoundment was rated as contaminated or uncontaminated. Using the data in Appendix C20, compute a logistic regression to determine which of the four explanatory variables significantly affects the probability of contamination. What is the best regression equation using one or more of these variables?