

Classification of Different Models for Categorical Variables:

- The **dependent** variable is categorical (contrast to OLS, which assumes a continuous distribution).

The **independent** variables can be any mix of metric variables and categorical variables (factors) as well as their interaction terms.

- Classification based on
 - [a] the **number of categories** of the response variable and
 - [b] whether each observation i constitutes a **single records** (binary response) with $n_i = 1$ or several observations are grouped together into **aggregates** (rates based on group counts n_g):

	Two Categories (dichotomous)	Multiple Categories (polytomous)
Individual Observations	<i>Binary distribution</i>	<i>Binary multinomial distribution</i>
Grouped Observations	<i>Binomial distribution</i>	<i>Multinomial distribution</i>

- Basic logistic regression focuses on the simple case with **individual** observations and each observation can fall into just one of **two mutually exclusive categories**.

The category status is coded binary:

$$Y_i = \begin{cases} 1 & \text{observation } i \text{ in first category} \\ 0 & \text{observation } i \text{ in second category} \end{cases}$$

Problems of Modeling Dichotomous Data by Linear Regression

Problem 1: Linear Predictions outside the feasible range:

- The predicted value cannot be interpreted as probability

$$\hat{\pi}_i = \Pr(Y_i = 1 | x_i) = b_0 + b_1 \cdot x_i \text{ for a given exogenous } x_i.$$

- The predicted linear probability value can fall *outside* the feasible range of probabilities $[0,1]$.
- For example: see Fig 7.4 when years lived in town is greater than 70.

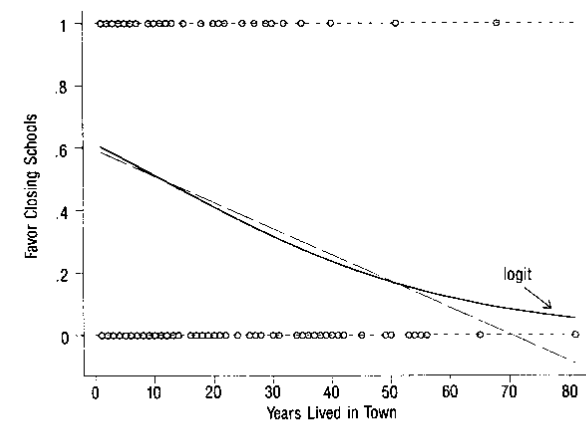


Figure 7.4 Logit regression of school-closing opinion on years lived in town, also showing linear regression line.

Solution to Problem 1: Infeasible Range of Predictions

- Calculate **odds** $\mathcal{O}_i = \frac{\pi_i}{1-\pi_i}$ with value range $\mathcal{O}_i \in [0, \infty]$.
- Transform odds into **logits** $\mathcal{L}_i = \ln(\mathcal{O}_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$ with the value range $\mathcal{L}_i \in [-\infty, +\infty]$.
 \Rightarrow the logits can be modeled by a linear function!


- The *linear function* in the logits becomes (here for the i -th observation):

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathcal{L}_i = \beta_0 + \beta_1 \cdot x_{i,1} + \dots + \beta_{K-1} \cdot x_{i,K-1}$$

- The estimated probabilities are given by the **inverse** of the logit function:

$$\begin{aligned}\ln\left(\frac{\pi_i}{1-\pi_i}\right) &= \mathcal{L}_i \\ \Rightarrow \pi &= \frac{1}{1 + \exp(-\mathcal{L}_i)} \\ \Rightarrow \pi &= \frac{\exp(\mathcal{L}_i)}{1 + \exp(\mathcal{L}_i)} \quad (\text{equivalent expression})\end{aligned}$$

- The inverse logit function gives the logistic curve.

- For bivariate logistic regression this curve is monotonically increasing for a positive “slope” parameter β_1 and monotonically decreasing for a negative “slope” parameter β_1 .
- See also the -script `FunctionalProbForms.r`.

Problem 2: Heteroscedasticity of disturbances

- Residuals for each observation Y_i can take only two distinct values:
$$e_i = \begin{cases} 1 - \hat{\pi}_i & \text{if } Y_i = 1 \\ 0 - \hat{\pi}_i & \text{if } Y_i = 0 \end{cases}$$
- The variance of the disturbance is given by: $Var(\varepsilon_i) = (1 - \pi_i) \cdot \pi_i$
- The spread of the disturbances depends on the varying estimated probabilities of the individual observations i and, therefore, their variances become **heteroscedastic**.

Solution to the heteroscedasticity problem:

- Maximum likelihood estimation automatically accounts for the heteroscedasticity.
- Alternatively, one could use **iteratively re-weighted** least squares.

Estimation of logistic regression by Maximum Likelihood:

- The maximum likelihood method asks the question “Given the observed sample data, what set of hypothetical population parameter values **most likely** has generated the data?”
- Discuss highlighted text block:

Let X_i stand for the i th combination of X values. Based on a logit model, the conditional probability that $Y_i = 1$ is

$$P_i = \frac{1}{1 + e^{-L_i}} \quad [7.6]$$

where

$$L_i = \beta_0 + \sum_{k=1}^{K-1} \beta_k X_{ik} \quad [7.7]$$

The contribution of the i th case to the likelihood function equals P_i if $Y_i = 1$, and it equals $1 - P_i$ if $Y_i = 0$. We could write this contribution as

$$P_i^{Y_i} (1 - P_i)^{1 - Y_i}$$

Assuming that the cases are independent (no autocorrelation), the likelihood function itself is the product of these individual contributions:

$$\mathcal{L} = \prod \{P_i^{Y_i} (1 - P_i)^{1 - Y_i}\} \quad [7.8]$$

Π is a *multiplication operator*, analogous to the summation operator Σ .

- The individual case probabilities are either

$$\begin{cases} \hat{\pi}_i^{y_i} \cdot \underbrace{(1 - \hat{\pi}_i)^{1-y_i}}_{=1} = \hat{\pi}_i & \text{for } y_i = 1 \text{ and} \\ \hat{\pi}_i^{y_i} \cdot \underbrace{(1 - \hat{\pi}_i)^{1-y_i}}_{=1} = 1 - \hat{\pi}_i & \text{for } y_i = 0 \end{cases}$$
- The specification of $\Pr(Y_i = 1)$ is inserted into equation 7.8.
- Under the independence assumption, individual probabilities can be linked multiplicatively.

We seek estimates of the β parameters that yield the highest possible values for the likelihood function, Equation [7.8]. Equivalently, we maximize the logarithm of [7.8], called the *log likelihood*:

$$\log_e \mathcal{L} = \sum \{ Y_i \log_e P_i + (1 - Y_i) \log_e (1 - P_i) \} \quad [7.9]$$

Logarithms convert multiplication into addition, making the log likelihood easier to work with.

To find maximum likelihood estimates, take first derivatives of the log likelihood with respect to each of the estimated parameters, and then set these derivatives equal to zero. This results in simultaneous equations:

$$\sum (Y_i - P_i) = 0 \quad [7.10]$$

and

$$\sum (Y_i - P_i) X_{ik} = 0 \quad \text{for } k = 1, 2, 3, \dots, K - 1 \quad [7.11]$$

These equations are nonlinear in the parameters and cannot be solved directly (unlike the normal equations for OLS). Instead, we resort to an iterative procedure, in which the computer finds successively better approximations for β_k values that satisfy [7.10]–[7.11].

- Since these equations are non-linear they can only be solved iteratively for the unknown parameters $\{\beta_0, \beta_1, \dots, \beta_{K-1}\}$
- The predicted probabilities, which are estimated by maximum likelihood, satisfy the constraints:

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{\pi}_i$$
$$\sum_{i=1}^n x_{ik} \cdot Y_i = \sum_{i=1}^n x_{ik} \cdot \hat{\pi}_i$$

- The first equation guarantees that the **number** of the observed “successes” matches the **sum of the estimated probabilities** for “success” in a sample. Therefore, the estimated probabilities provide unbiased predictions.
- The second equation constraints the estimated probabilities weighted by x_{ik} . If x_{ik} is an indicator variable then the estimated probabilities in the associated group are equal to the observed number of “successes” in that group.

Parameter Interpretation

- As consequence of the **non-linearity** in the relationship between the dependent variable $\Pr(Y_i = 1)$ and the independent variables, all interpretations of the single estimated logit

regression parameters must be conducted **conditionally** to the given values of the other variables.

For instance, the **average** of the remaining independent variables can be used.

- See the -function `effects::allEffects()`.

[a] Approach in terms of the logit

- Recall that the logit-transformation is a monotone function (the larger the logit the larger the underlying probability).

Thus the estimated parameters in $\hat{\mathcal{L}}_i = b_0 + b_1 \cdot x_{i,1} + \dots + b_{K-1} \cdot x_{i,K-1}$ can be interpreted as:

for **positive** coefficients b_k , the **greater X the larger the expected probability**. Analog for negative b_k s.

- Interpretation of the model in terms of the logits is only the starting point, because it does provide only an indirect link to the more meaningful probabilities.

[b] Approach in terms of the odds

- The odds can be written as

$$\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = e^{b_0} \cdot e^{b_1 \cdot x_{i1}} \dots e^{b_{K-1} \cdot x_{i(K-1)}}$$

$$= e^{b_0} \cdot (e^{b_1})^{x_{i1}} \dots (e^{b_{K-1}})^{x_{i(K-1)}}$$

- Assuming all **variables remain at a preset level** except for the k -th one, then one unit change in x_{ik} changes the odds for observing $Y_i = 1$ by the factor e^{b_k} (**multiplicative link**).
- If the estimated parameter is zero, i.e., $b_k = 0$, then the odds won't change because $(e^0)^{x_{ik}} = 1$ (one is the neutral factor in multiplication)
- The percentage change in the odds for observing $Y_i = 1$ is given by $100 \cdot (e^{b_k} - 1)$.
- For a dummy variable X this has a clear interpretation by the **odds-ratios**: $OR = \frac{\hat{\pi}_{X=1}}{\hat{\pi}_{X=0}} = e^{b_k}$
because the effects of the **remaining variables** cancels out irrespectively of their observed

levels:

$$OR = \frac{e^{b_0} \cdot e^{b_1 \cdot x_{i1}} \cdot \dots \cdot e^{b_k \cdot 1} \cdot \dots \cdot e^{b_{K-1} \cdot x_{i(K-1)}}}{\underbrace{e^{b_0} \cdot e^{b_1 \cdot x_{i1}} \cdot \dots \cdot e^{b_k \cdot 0} \cdot \dots \cdot e^{b_{K-1} \cdot x_{i(K-1)}}}_{=1}} = e^{b_k}$$

[c] Approach in terms of the probabilities

- The estimated probability for observation i can be expressed by:

$$\hat{\pi}_i = \frac{1}{1 + \exp(-L_i)} = \frac{1}{1 + \exp(-\mathbf{x}_i^T \cdot \mathbf{b})}$$

- The interpretation that one unit change in X causes b units changes in the probability (either additively or multiplicatively) is no longer valid.

It depends on the relative value of the predicted probabilities $\hat{\pi}_i$ (see **HAM Fig 7.3**):

- The slope of the logistic curve for a given probability $\hat{\pi}_i$ with respect to one-unit change in an independent variable is $b_k \cdot \hat{\pi}_i \cdot (1 - \hat{\pi}_i)$

$\hat{\pi}_i$	$slope = b_k \cdot \hat{\pi}_i \cdot (1 - \hat{\pi}_i)$
0.01	$b_k \times 0.0099$
0.05	$b_k \times 0.04575$
0.10	$b_k \times 0.09$

0.50	$b_k \times 0.25$
0.90	$b_k \times 0.09$
0.95	$b_k \times 0.04575$
0.99	$b_k \times 0.0099$

- It has its maximum at $\hat{\pi}_i = 0.5$ and is ***fairly linear*** in the interval $\hat{\pi}_i \in [0.25, 0.75]$.
- For metric variables only ***conditional effect plots***, where the remaining variables are fixed at a ***specific*** level, can be given:

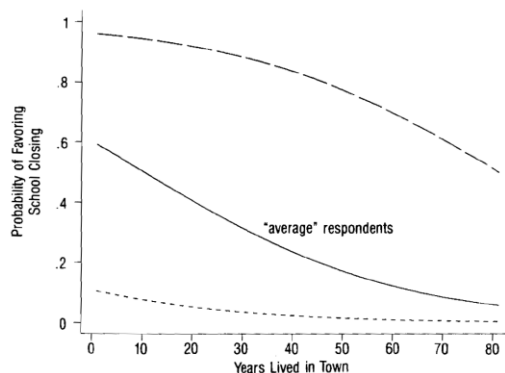


Figure 7.5 Conditional effects of years lived in town, at proclosing (top), average, and anticlosing levels of other X variables.

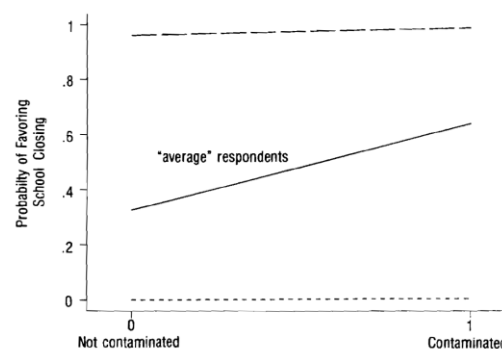


Figure 7.6 Conditional effects of contamination, at proclosing, average, and anticlosing levels of other X variables.

- Note the identity of the logistic curve and the logistic curve of the "average" respondent in Fig 7.5.

- This allows investigating the effects of one variable on the probability for different population segments (for example *risk-takers* against *cautious individuals*).