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).
  Here each group member shares the same exogenous group attributes:

	Two Categories (*dichotomous*)	**Multiple Categories (*polytomous*)**
Individual Observations	*Binary distribution*	*Binary multinomial distribution*
Grouped Observations	*Binomial distribution*	*Multinomial distribution*

- There are specialized specifications of ordinal scaled dependent variables available.

- 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

- Logistic regression aims at predicting the probabilities of belonging each category

Problem 1: Linear Predictions outside the feasible range:

- The predicted value cannot be interpreted as probability
  $\hat{\pi}_i = \Pr(Y_i = 1 | x_i) = b_0 + b_1 \cdot x_i$ for a given exogenous x_i.
- The predicted linear probability value can fall *outside* the feasible range of probabilities $[0,1]$.
- For example: see HAM Fig 7.4 when years lived in town is greater than 70.

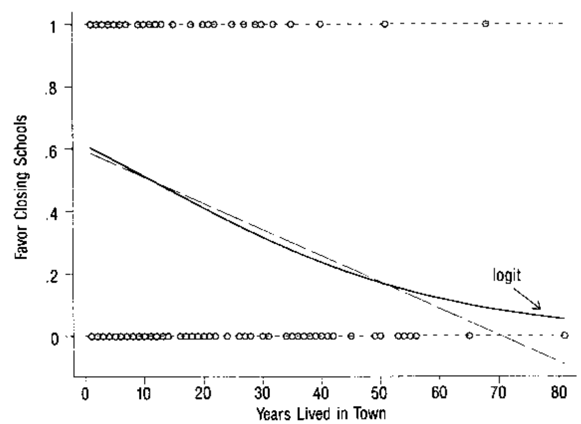


Figure 7.4 Logit regression of school-closing opinion on years lived in town, also showing linear regression line.

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!

Probability π	Odds $\frac{\pi}{1-\pi}$	Logit $\log_e \frac{\pi}{1-\pi}$
.01	1/99 = 0.0101	−4.60
.05	5/95 = 0.0526	−2.94
.10	1/9 = 0.1111	−2.20
.30	3/7 = 0.4286	−0.85
.50	5/5 = 1	0.00
.70	7/3 = 2.333	0.85
.90	9/1 = 9	2.20
.95	95/5 = 19	2.94
.99	99/1 = 99	4.60

$$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$

$$\frac{\pi_i}{1-\pi_i} = \exp(\alpha + \beta X_i) = \exp(\alpha)\exp(\beta X_i) = \exp(\alpha)[\exp(\beta)]^{X_i}$$

- <u>Note:</u> This specification only works for probabilities ***excluding*** the certain events of 0 and 1. This means, it is impossible to transform revealed preferences $Y_i = 1$ or $Y_i = 0$ directly into odds or logits.
- The *linear function* in the logits becomes (here for the *i*-th observation):

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

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

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

- The inverse logit function gives the logistic curve displayed in HAM Fig 7.3.
- 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 (HAM Fig 7.4)
- See also the R-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 population 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 us ***iteratively re-weighted*** least squares (see Hamilton note 10 on page 238 and page 247).
  However, the weights must be given in terms of logits $\mathcal{L}_i$ for which the variance is at the k^{th} iteration
  $$Var\left(\ln \frac{Y_i}{1-Y_i}\right) = \frac{1}{n_i \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)}$$
- Note: In contrast, the variance of the proportion estimate $Y_i = \frac{\text{\# of successes}}{n_i}$ is
  $$Var(Y_i) = \frac{\hat{\pi}_i \cdot (1-\hat{\pi}_i)}{n_i}$$

Estimation of logistic regression:

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

- Discuss highlighted text block in **HAM** p 223-224:

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

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

where

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

The contribution of the ith 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:

$$\mathscr{L} = \Pi\{P_i^{Y_i}(1 - P_i)^{1 - Y_i}\} \qquad [7.8]$$

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

- The individual case probabilities are either $$\begin{cases} \hat{\pi}_i^{y_i} \cdot \underbrace{(1-\hat{\pi}_i)^{1-y_i}}_{=1} = \hat{\pi}_\mathrm{i} & \text{for } y_i = 1 \text{ and} \\ \underbrace{\hat{\pi}_i^{y_i}}_{=1} \cdot (1-\hat{\pi}_i)^{1-y_i} = 1 - \hat{\pi}_i & \text{for } y_i = 0 \end{cases}$$

- 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*:

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

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:

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

and

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

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 $\{\beta_0, \beta_1, \ldots, \beta_{K-1}\}$

- The predicted probabilities, which are estimated by maximum likelihood, satisfy the constraints (**HAM eqs 7.10** and **7.11**):

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

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

  - 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 sum of estimated probabilities in the associated group are equal to the observed sum of "successes" in that group.

- Discussion of **HAM** Table 7.1:

Table 7.1 Logit regression of school-closing opinion on years lived in town.

Iteration 0: Log Likelihood = –104.60578
Iteration 1: Log Likelihood = –97.80942
Iteration 2: Log Likelihood = 97.634236
Iteration 3: Log Likelihood = –97.633571

Logit Estimates Number of obs = 153
chi2(1) = 13.94
Log Likelihood = –97.633571 Prob > chi2 = 0.0002

Variable	Coefficient	Std. Error	*t*	Prob > |*t*|	Mean
close					.4313725
lived	–.0409876	.01214	–3.376	0.001	19.26797
_cons	.4599786	.2625656	1.752	0.082	1

  - An initial guess for $\beta_1^{0-step} = 0$ (i.e., the observed x_i's have no effect on $\Pr(Y_i = 1)$) and for β_0 it is based on $\frac{\sum_{i=1}^{n} Y_i}{n} = \frac{\sum_{i=1}^{n} \frac{1}{1+e^{-\beta_0}}}{n} = \frac{1}{1+e^{-\beta_0}}$, which can be solved for β_0 .
  - The bivariate logistic regression in **HAM** Table 7.1 uses $\hat{\beta}_0^{0-step}$ and $\hat{\beta}_1^{0-step} = 0$ as starting value for the iteration 0. This gives a ***log-likelihood*** of –104.6.

At the third iteration the parameter estimates are $\hat{\beta}_0^{3-step} = 0.46$ and $\hat{\beta}_1^{3-step} = -0.041$ at an improved ***log-likelihood*** value of –97.6

- A test equivalent to the global *F*-test $H_0 : \beta_1 = \ldots = \beta_{K-1} = 0$ is given by the likelihood ratio test:
  $\chi^2_{df=K-1} = -2 \cdot \left[\text{-104.60-(-97.63)}\right] = 13.94$ with $df = K - 1$ for $K = \{\beta_0, \beta_1\}$.

Excursion 1: Alternative derivations of logistic regressions

- Assume an underlying ***cumulative distribution*** function.
  Link the observation Y_i by ***an unobserved variable formulation*** ξ_i to the independent variable.

  A prominent example of this perspective is the economic *discrete choice theory* (Nobel Prize in Economics to McFadden in 2002).

$$Y_i = \begin{cases} 0 & \text{when } \xi_i \le 0 \\ 1 & \text{when } \xi_i > 0 \end{cases}$$

$$\xi_i = \alpha + \beta X_i - \varepsilon_i$$

$$\pi_i \equiv \Pr(Y_i = 1) = \Pr(\xi_i > 0) = \Pr(\alpha + \beta X_i - \varepsilon_i > 0)$$
$$= \Pr(\varepsilon_i < \alpha + \beta X_i)$$

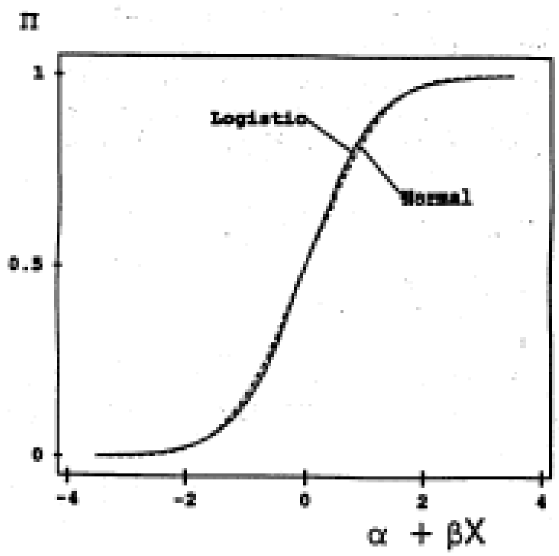
If the errors are independently distributed according to the unit-normal distribution, $\varepsilon_i \sim N(0, 1)$, then

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Phi(\alpha + \beta X_i)$$

which is the probit model.[11] Alternatively, if the ε_i follow the similar logistic distribution, then we get the logit model

$$\pi_i = \Pr(\varepsilon_i < \alpha + \beta X_i) = \Lambda(\alpha + \beta X_i)$$

- Two cumulative distribution functions are:
  [a] the standard normal distribution leads to the *probit*-model
  [b] the logistic distribution leads to the *logit*-model
- The ***logit*** model is mathematically easier to implement with because its cumulative distribution function is known analytically and therefore it avoids numerical integration as in the case of the probits.



- Using the normal distribution $\Phi(\cdot)$ yields the *linear probit model*:

$$\pi_i = \Phi(\alpha + \beta X_i)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} \exp\left(-\frac{1}{2}Z^2\right) dZ$$

- Using the logistic distribution $\Lambda(\cdot)$ produces the *linear logistic-regression* or *linear logit model*:

$$\pi_i = \Lambda(\alpha + \beta X_i) \qquad [15.8]$$
$$= \frac{1}{1 + \exp[-(\alpha + \beta X_i)]}$$

Figure 15.3. Once their variances are equated, the cumulative logistic and cumulative normal distributions—used here to transform $\alpha + \beta X$ to the unit interval—are virtually indistinguishable.

Excursion B: Iterative estimation of non-linear systems

- Recall that the first derivative of the log-likelihood function (i.e., its slope) has to be zero at the optimal value of its argument β.

- The Newton-Raphson algorithm iteratively approximates the curved first derivative of the log-likelihood by a linear function

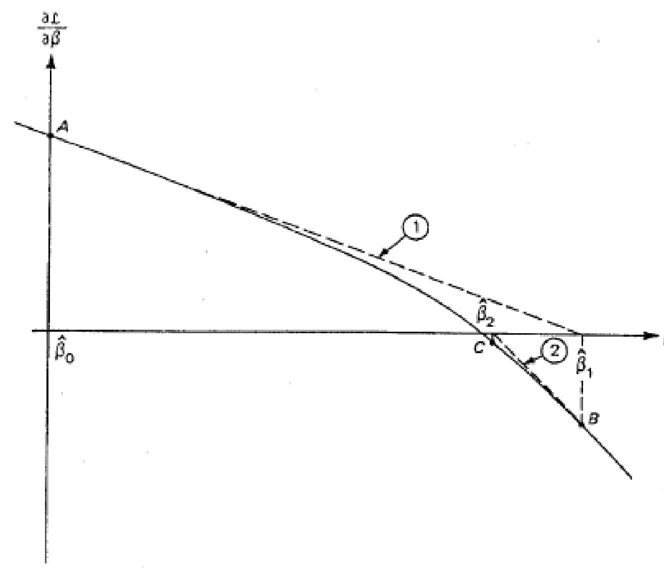


Figure 4.7
Illustration of Newton-Raphson algorithm

Source: Ben-Akiva, Lerman (1985). Discete Choice Analysis. Theory and Application to Travel Demand. MIT-Press

- The zeros of the linear function are easily calculated.
- The linear approximation continues at this tentative zero value until the algorithm converges, i.e., the tentative zeros do not change any more.

Global and partial tests via the likelihood ratio test:

- Different models produce different likelihood values: Prominent models are:
  - $\hat{\pi}_i = \frac{\sum_{i=1}^{n} Y_i}{n} \; \forall \, i \;\; \Longrightarrow \; L_0$
    This is a model with ***constant probabilities*** using only the intercept estimate b_0.
  - $\hat{\pi}_i = Y_i \; \forall \, i \;\; \Longrightarrow \; L_S$
    This is the ***saturated model***. It fits the observed data perfectly because it has ***one estimated parameter per observation***, which leads to a predicted probability $\hat{\pi}_i$ that is equal to the revealed outcome Y_i.
    For a binary logistic model the likelihood of the saturated model becomes one:
    $L_S = \prod_{i=1}^{n} \underbrace{\hat{Y}_i}_{=\hat{\pi}_i}{}^{Y_i} \cdot \underbrace{(1 - \hat{Y}_i)}_{=1-\hat{\pi}_i}{}^{1-Y_i} = 1.$
- The likelihoods are ordered in a sequence $L_0 \leq L_{K-H} \leq L_K \leq L_S$ ranging from the least fitting model with just the intercept to the fully fitting model having n estimated parameters.
  The intermediate models comprise of $K - H$ and K parameters where H is the number of constraint parameters (i.e., set to the assumed value under the null hypothesis).
- As with the partial *F*-test the estimated parameters need to be in a ***nested sequence***.
- See **Ham Tables 7.1-7.4** for the nested test strategy.

- Testing is conducted in terms of the log-likelihoods l_K of the models in a nested-modeling strategy
  - Let $l_K = \ln(L_K)$ be the log likelihood of the full model with *K* parameters
  - Let $l_{K-H} = \ln(L_{K-H})$ be the log likelihood the restricted model with *H* less parameters
- The ***likelihood ratio*** test statistic $\chi_H^2 = -2 \cdot (l_{K-H} - l_K)$ can be used.
  - It follows approximately a χ^2-distribution with *H* degrees of freedom. If both models do not differ then the χ^2-value would be zero. Thus large χ^2-value indicates significant differences and we need to apply a one-sided test.
  - In order to perform the likelihood ratio test we need to estimate the full and the restricted model.
- Several software packages report the ***deviance*** rather than likelihoods or log-likelihoods.
  - The deviance compares the estimated model against the saturated model: $D = -2 \cdot (l_K - l_S)$.
  - Since $l_S = \log(\underbrace{L_S}_{=1}) = 0$, the deviance is related to the log-likelihood of the model by $D_K = -2 \cdot \log(L_K)$.

  - The deviance has a similar interpretation than residual sum of squares in OLS. That is, smaller deviances are better.
  - A likelihood ratio test can be performed in terms of differences in deviances: $\chi_H^2 = D_{K-H} - D_K$

- Individual ***Wald*** *t*-tests $H_0 : \beta_k = 0$ on the parameters can be conducted by the *t*-stastistic: $t = \dfrac{b_k}{SE_{b_k}}$.
- Remember: The *likelihood ratio* test as well as the *Wald t*-test are only asymptotically valid.
- The Akaike's information criterion evaluates the log-likelihood of a model against the number of estimated parameters in the model: $AIC = -2 \cdot l_K + 2 \cdot K = D_K + 2 \cdot K$.
  Smaller *AIC* indicate a better fitting model.

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 R-function `car::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 $\hat{\mathcal{L}}_i = b_0 + b_1 \cdot x_{i,1} + \cdots + b_{K-1} \cdot x_{i,K-1}$ can be interpreted as:

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

- Individual confidence intervals can be calculated for the logit specification:

  $b_k - t(SE_{b_k}) \leq \beta_k \leq b_k + t(SE_{b_k})$

  We can the reject the null hypothesis $H_0 : \beta_k = 0$ if the interval does not include zero.

- 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-specification can be written as

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

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

- Assuming all ***variables remain at a preset level*** except for the *k*-th variable, then one unit change in x_{ik} changes the odds of 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 $\left(e^0\right)^{x_{ik}} = 1$ (one is the neutral factor in multiplication)
- The confidence interval in terms of the exponential function becomes

$$\exp\left(b_k - t(SE_{b_k})\right) \leq \exp\left(\beta_k\right) \leq \exp\left(b_k + t(SE_{b_k})\right)$$

If it does not include 1 the null hypothesis then $H_0\colon \exp(\beta_k) = 1$ can be rejected.

- The percentage change in the odds for observing $Y_i = 1$ is given by $100 \cdot \left(e^{b_k} - 1\right)$.

- For a dummy variable *X* this has a clear interpretation by the ***odds-ratios***: $OR = \frac{\hat{O}_{X=1}}{\hat{O}_{X=0}} = e^{b_k}$

  because the effects of the ***remaining variables*** cancels out irrespectively of their observed levels:

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

[c] Approach in terms of the probabilities

- The estimated probability for observation *i* can is 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 dependents 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 given by $b_k \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)$

$\hat{\pi}_i$	$slope = b_k \cdot \hat{\pi}_i \cdot (1-\hat{\pi}_i)$
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 it maximum at $\hat{\pi}_i = 0.5$
- The logistic curve ***fairly linear*** in the short intervals of $\pi \in [0 < \pi^{lower}, \pi^{upper} < 1]$. Consequently, *weighted* linear regression can be used for observed rates, as long as they don't vary too much.

- For metric variables only ***conditional effect plots***, where the remaining variables are set at a ***specific fixed*** level, can be given (see **HAM** Figs 7.5 and 7.6)

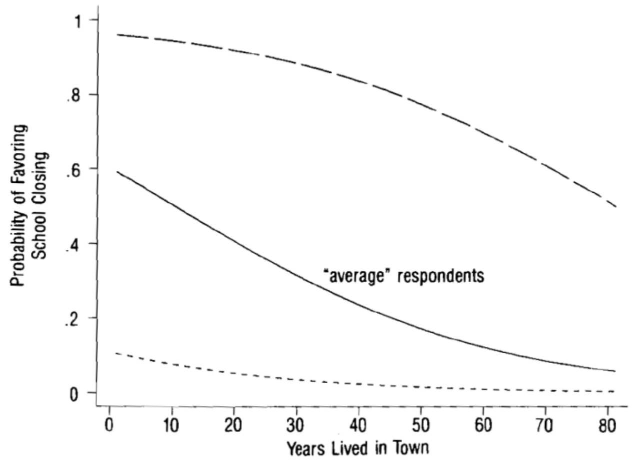


Figure 7.5 Conditional effects of years lived in town, at proclosing (top), average, and anticlosing levels of other X variables.

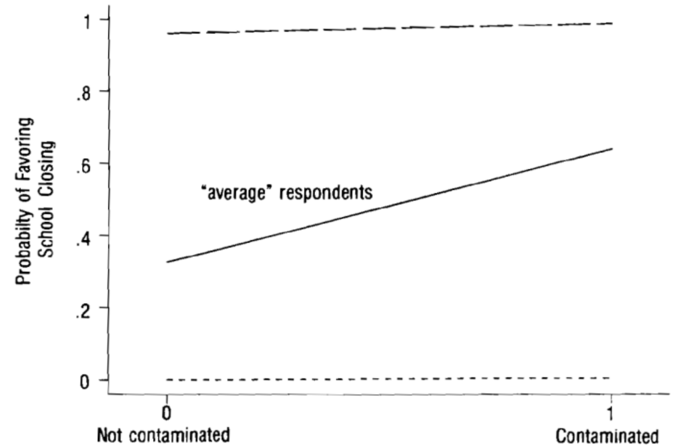


Figure 7.6 Conditional effects of contamination, at proclosing, average, and anticlosing levels of other X variables.

- Note the identity of the logistic curve in Fig 7.4 (model with the only variable **"lived"**) 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 scenarios (in the example **HAM p 231** : population segment of *risk-takers* against *cautious people*).

Statistical Problems

- Examine multicollinearity: [a] correlation among variables, [b] correlation matrix among estimate coefficients (preferred approach), [c] VIFs.
- High ***discrimination*** does not allow estimating parameters (see example **HAM** Tab 7.5) because the measured level of an independent variable pre-determines the outcome of the dependent variable. I.e., for all $x_k = c$ the observed Y is either 0 or 1.

Excursion: Residual Analysis

- Detailed information on how to perform residual analysis and diagnostics for generalized linear models can be found in section 8.6 “Diagnostics for Generalized Linear Models” in Fox & Weissberg’s *R Companion to Regression Analysis*
- Discuss notation of grouped independent variables **HAM** Tab. 7.6:
  - Y_j is the sum of “successes” in a grouped pattern j of the independent variables.
  - n_j is the number of cases with a grouped pattern j.
- Two different kinds of residuals:

- Pearson residuals (analog to the χ^2-test for independence in contingency tables):

$$r_j = \frac{Y_j - n_j \cdot \hat{\pi}_j}{\sqrt{n_j \cdot \hat{\pi}_j \cdot (1-\hat{\pi}_j)}}$$

- Deviance residuals are based on the log-likelihood of individual cases (common grouped patterns n_j):

$$d_j = \pm\left\{2 \cdot \left[Y_j \cdot \log\left(\frac{Y_j}{n_j \cdot \hat{\pi}_j}\right) + \left(n_j - Y_j\right) \cdot \log\left(\frac{n_j - Y_j}{n_j \cdot \left(1-\hat{\pi}_j\right)}\right)\right]\right\}^{1/2}$$

  where the sign depends on the sign of $Y_j - n_j \cdot \hat{\pi}_j$.

  Deviance residuals measure in how far a predicted probability of an individual observation differs from that of its saturated equivalent, which is zero.

- Both statistics follow a χ^2-distribution. A low χ^2-value indicates that the model is not much different from a model with a perfect fit for Y_j.

Excursion: The Saturated Model Perspective

- Using the ***saturated model*** as baseline switches the testing perspective:

a. Instead of looking for a model with a good fit starting from a basic model with only the intercept (=> constant predicted probability for all observations), we test against a ***perfect*** model that ***overfits*** the data by having as many parameters as we have observations.

b. We are looking for a ***parsimonious*** model and thus reduce the number of parameters as much as possible but keeping the χ^2-test statistics ***barely insignificant***.
That is, the final model with fewer parameters does not differ significantly from a perfectly fitting model with many parameters.