Sociology 740 John Fox

Lecture Notes

9. Logit and Probit Models For Dichotomous Data

1. Goals:

- To show how models similar to linear models can be developed for qualitative response variables.
- To introduce logit and probit models for dichotomous response variables.

2. An Example of Dichotomous Data

- To understand why logit and probit models for qualitative data are required, let us begin by examining a representative problem, attempting to apply linear regression to it:
  - In September of 1988, 15 years after the coup of 1973, the people of Chile voted in a plebiscite to decide the future of the military government. A 'yes' vote would represent eight more years of military rule; a 'no' vote would return the country to civilian government. The no side won the plebiscite, by a clear if not overwhelming margin.
  - Six months before the plebiscite, FLACSO/Chile conducted a national survey of 2,700 randomly selected Chilean voters.
    - Of these individuals, 868 said that they were planning to vote yes, and 889 said that they were planning to vote no.
    - Of the remainder, 558 said that they were undecided, 187 said that they planned to abstain, and 168 did not answer the question.

    - I will look only at those who expressed a preference.
  - Figure 1 plots voting intention against a measure of support for the status quo.
    - Voting intention appears as a dummy variable, coded 1 for yes, 0 for no.
    - Support for the status quo is a scale formed from a number of questions about political, social, and economic policies: High scores represent general support for the policies of the miliary regime.
  - Does it make sense to think of regression as a conditional average when the response variable is dichotomous?
    - An average between 0 and 1 represents a 'score' for the dummy response variable that cannot be realized by any individual.

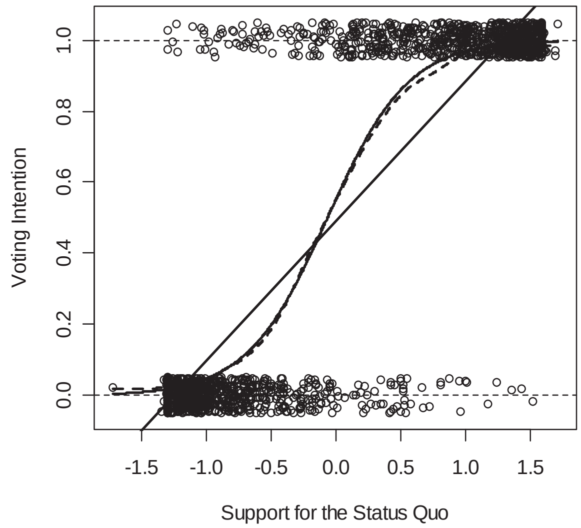

Figure 1. The Chilean plebiscite data: The solid straight line is a linear least-squares fit; the solid curved line is a logistic-regression fit; and the broken line is from a nonparametric kernel regression with a span of .4.The individual observations are all at 0 or 1 and are vertically jittered.

– In the population, the conditional average $E(Y|x_i)$ is the proportion of 1's among those individuals who share the value x_i for the explanatory variable — the conditional probability π_i of sampling a 'yes' in this group:

$$\pi_i \equiv \Pr(Y_i) \equiv \Pr(Y = 1|X = x_i)$$

and thus,

$$E(Y|x_i) = \pi_i(1) + (1 - \pi_i)(0) = \pi_i$$

- If X is discrete, then in a sample we can calculate the conditional proportion for Y at each value of X.
  – The collection of these conditional proportions represents the sample nonparametric regression of the dichotomous Y on X.
  – In the present example, X is continuous, but we can nevertheless resort to strategies such as local averaging, as illustrated in the figure.

3. The Linear-Probability Model

- Although non-parametric regression works here, it would be useful to capture the dependency of Y on X as a simple function, particularly when there are several explanatory variables.
- Let us first try linear regression with the usual assumptions:
$$Y_i = \alpha + \beta X_i + \varepsilon_i$$
where $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$, and ε_i and ε_j are independent for $i \neq j$.
  - If X is random, then we assume that it is independent of ε.
- Under this model, $E(Y_i) = \alpha + \beta X_i$, and so
$$\pi_i = \alpha + \beta X_i$$
  - For this reason, the linear-regression model applied to a dummy response variable is called the *linear probability model*.
- This model is untenable, but its failure points the way towards more adequate specifications:

- *Non-normality:* Because Y_i can take on only the values of 0 and 1, the error ε_i is dichotomous as well — not normally distributed:
  - If $Y_i = 1$, which occurs with probability π_i, then
$$\begin{aligned}\varepsilon_i &= 1 - E(Y_i)\\ &= 1 - (\alpha + \beta X_i)\\ &= 1 - \pi_i\end{aligned}$$
  - Alternatively, if $Y_i = 0$, which occurs with probability $1 - \pi_i$, then
$$\begin{aligned}\varepsilon_i &= 0 - E(Y_i)\\ &= 0 - (\alpha + \beta X_i)\\ &= 0 - \pi_i\\ &= -\pi_i\end{aligned}$$
  - Because of the central-limit theorem, however, the assumption of normality is not critical to least-squares estimation of the normal-probability model.

- *Non-constant error variance:* If the assumption of linearity holds over the range of the data, then $E(\varepsilon_i) = 0$.
  - Using the relations just noted,
$$\begin{aligned} V(\varepsilon_i) &= \pi_i(1-\pi_i)^2 + (1-\pi_i)(-\pi_i)^2 \\ &= \pi_i(1-\pi_i) \end{aligned}$$
  - The heteroscedasticity of the errors bodes ill for ordinary-least-squares estimation of the linear probability model, but only if the probabilities π_i get close to 0 or 1.
- *Nonlinearity:* Most seriously, the assumption that $E(\varepsilon_i) = 0$ — that is, the assumption of linearity — is only tenable over a limited range of X-values.
  - If the range of the X's is sufficiently broad, then the linear specification cannot confine π to the unit interval $[0, 1]$.
  - It makes no sense, of course, to interpret a number outside of the unit interval as a probability.

  - This difficulty is illustrated in the plot of the Chilean plebiscite data, in which the least-squares line produces fitted probabilities below 0 at low levels and above 1 at high levels of support for the status-quo.

▶ Dummy *regressor* variables do not cause comparable difficulties because the general linear model makes no distributional assumptions about the X's.

▶ Nevertheless, if π doesn't get too close to 0 or 1, the linear-probability model estimated by least-squares frequently provides results similar to those produced by more generally adequate methods.

▶ One solution — though not a good one — is simply to constrain π to the unit interval:
$$\pi = \begin{cases} 0 & \text{for } 0 > \alpha + \beta X \\ \alpha + \beta X & \text{for } 0 \leq \alpha + \beta X \leq 1 \\ 1 & \text{for } \alpha + \beta X > 1 \end{cases}$$

- The *constrained linear-probability* model fit to the Chilean plebiscite data by maximum likelihood is shown in Figure 2. Although it cannot be dismissed on logical grounds, this model has certain unattractive features:
  - *Instability:* The critical issue in estimating the linear-probability model is identifying the X-values at which π reaches 0 and 1, since the line $\pi = \alpha + \beta X$ is determined by these two points. As a consequence, estimation of the model is inherently unstable.
  - *Impracticality:* It is much more difficult to estimate the constrained linear-probability model when there are several X's.
  - *Unreasonableness:* Most fundamentally, the abrupt changes in slope at $\pi = 0$ and $\pi = 1$ are unreasonable. A smoother relationship between π and X, is more generally sensible.

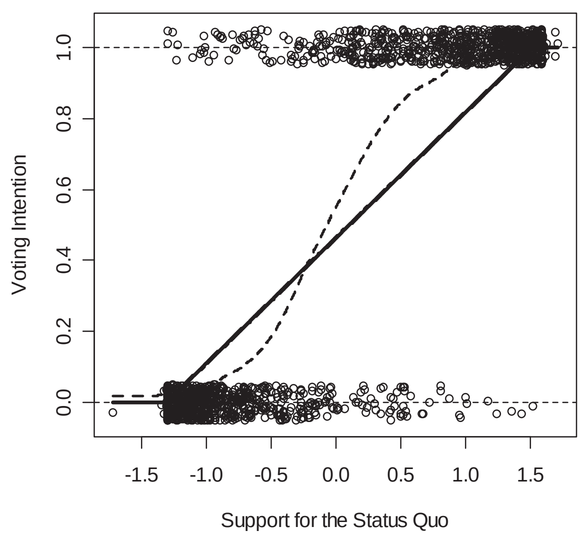

Figure 2. The solid line shows the linear-probability model fit by maximum likelihood to the Chilean plebiscite data; the broken line is for a nonparametric kernel regression.

4. Transformations of π: Logit and Probit Models

- To insure that π stays between 0 and 1, we require a positive monotone (i.e., non-decreasing) function that maps the 'linear predictor' $\eta = \alpha + \beta X$ into the unit interval.
  - A transformation of this type will retain the fundamentally linear structure of the model while avoiding probabilities below 0 or above 1.
  - Any cumulative probability distribution function meets this requirement:
$$\pi_i = P(\eta_i) = P(\alpha + \beta X_i)$$
where the CDF $P(\cdot)$ is selected in advance, and α and β are then parameters to be estimated.
  - If we choose $P(\cdot)$ as the cumulative rectangular distribution then we obtain the constrained linear-probability model.
  - An *a priori* reasonable $P(\cdot)$ should be both smooth and symmetric, and should approach $\pi = 0$ and $\pi = 1$ as asymptotes.

  - Moreover, it is advantageous if $P(\cdot)$ is strictly increasing, permitting us to rewrite the model as
$$P^{-1}(\pi_i) = \eta_i = \alpha + \beta X_i$$
where $P^{-1}(\cdot)$ is the inverse of the CDF $P(\cdot)$, i.e., the quantile function.
    - Thus, we have a linear model for a transformation of π, or — equivalently — a nonlinear model for π itself.

- The transformation $P(\cdot)$ is often chosen as the CDF of the unit-normal distribution
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{1}{2}Z^2} dZ$$
or, even more commonly, of the *logistic distribution*
$$\Lambda(z) = \frac{1}{1 + e^{-z}}$$
where $\pi \approx 3.141$ and $e \approx 2.718$ are the familiar constants.

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

$$\begin{aligned}\pi_i &= \Phi(\alpha + \beta X_i) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta X_i} e^{-\frac{1}{2}Z^2} dZ\end{aligned}$$

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

$$\begin{aligned}\pi_i &= \Lambda(\alpha + \beta X_i) \\ &= \frac{1}{1 + e^{-(\alpha+\beta X_i)}}\end{aligned}$$

- Once their variances are equated, the logit and probit transformations are so similar that it is not possible in practice to distinguish between them, as is apparent in Figure 3.
- Both functions are nearly linear between about $\pi = .2$ and $\pi = .8$. This is why the linear probability model produces results similar to the logit and probit models, except when there are extreme values of π_i.

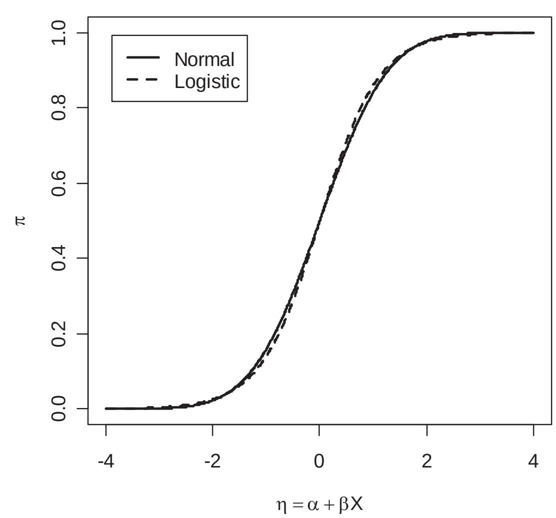

Figure 3. The normal and logistic cumulative distribution functions (as a function of the linear predictor and with variances equated).

▶ Despite their similarity, there are two practical advantages of the logit model:

1. *Simplicity:* The equation of the logistic CDF is very simple, while the normal CDF involves an unevaluated integral.
   - This difference is trivial for dichotomous data, but for polytomous data, where we will require the *multivariate* logistic or normal distribution, the disadvantage of the probit model is more acute.
2. *Interpretability:* The inverse linearizing transformation for the logit model, $\Lambda^{-1}(\pi)$, is directly interpretable as a *log-odds*, while the inverse transformation $\Phi^{-1}(\pi)$ does not have a direct interpretation.
   - Rearranging the equation for the logit model,
   $$\frac{\pi_i}{1-\pi_i} = e^{\alpha+\beta X_i}$$
   - The ratio $\pi_i/(1-\pi_i)$ is the *odds* that $Y_i = 1$, an expression of relative chances familiar to gamblers.

- Taking the log of both sides of this equation,
$$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta X_i$$
- The inverse transformation $\Lambda^{-1}(\pi) = \log_e[\pi/(1-\pi)]$, called the *logit* of π, is therefore the log of the odds that Y is 1 rather than 0.
- The logit is symmetric around 0, and unbounded both above and below, making the logit a good candidate for the response-variable side of a linear model:

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

- The logit model is also a multiplicative model for the odds:

$$\begin{aligned}\frac{\pi_i}{1-\pi_i} &= e^{\alpha+\beta X_i} = e^{\alpha}e^{\beta X_i}\\ &= e^{\alpha}\left(e^{\beta}\right)^{X_i}\end{aligned}$$

  – So, increasing X by 1 changes the logit by β and multiplies the odds by e^{β}.
  – For example, if $\beta = 2$, then increasing X by 1 increases the odds by a factor of $e^2 \approx 2.718^2 = 7.389$.
- Still another way of understanding the parameter β in the logit model is to consider the slope of the relationship between π and X.
  – Since this relationship is nonlinear, the slope is not constant; the slope is $\beta\pi(1-\pi)$, and hence is at a maximum when $\pi = 1/2$, where the slope is $\beta/4$:

π	$\beta\pi(1-\pi)$
$.01$	$\beta \times .0099$
$.05$	$\beta \times .0475$
$.10$	$\beta \times .09$
$.20$	$\beta \times .16$
$.50$	$\beta \times .25$
$.80$	$\beta \times .16$
$.90$	$\beta \times .09$
$.95$	$\beta \times .0475$
$.99$	$\beta \times .0099$

– The slope does not change very much between $\pi = .2$ and $\pi = .8$, reflecting the near linearity of the logistic curve in this range.

▶ The least-squares line fit to the Chilean plebescite data has the equation

$$\widehat{\pi}_{\text{yes}} = 0.492 + 0.394 \times \text{ Status-Quo}$$

- This line is a poor summary of the data.

▶ The logistic-regression model, fit by the method of maximum-likelihood, has the equation

$$\log_e \frac{\widehat{\pi}_{\text{yes}}}{\widehat{\pi}_{\text{no}}} = 0.215 + 3.21 \times \text{ Status-Quo}$$

- The logit model produces a much more adequate summary of the data, one that is very close to the nonparametric regression.
- Increasing support for the status-quo by one unit multiplies the odds of voting yes by $e^{3.21} = 24.8$.
- Put alternatively, the slope of the relationship between the fitted probability of voting yes and support for the status-quo at $\widehat{\pi}_{\text{yes}} = .5$ is $3.21/4 = 0.80$.

4.1 An Unobserved-Variable Formulation

- An alternative derivation posits an underlying regression for a continuous but unobservable response variable ξ (representing, e.g., the 'propensity' to vote yes), scaled so that

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

  - That is, when ξ crosses 0, the observed discrete response Y changes from 'no' to 'yes.'
  - The latent variable ξ is assumed to be a linear function of the explanatory variable X and the unobservable error variable ε:

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

- We want to estimate α and β, but cannot proceed by least-squares regression of ξ on X because the latent response variable is not directly observed.

- Using these equations,

$$\begin{aligned} \pi_i &\equiv \Pr(Y_i = 1) = \Pr(\xi_i > 0) = \Pr(\alpha + \beta X_i - \varepsilon_i > 0) \\ &= \Pr(\varepsilon_i < \alpha + \beta X_i) \end{aligned}$$

  - 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.
  - Alternatively, if the ε_i follow the similar logistic distribution, then we get the logit model

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

- We will return to the unobserved-variable formulation when we consider models for ordinal categorical data.

5. Logit and Probit Models for Multiple Regression

▶ To generalize the logit and probit models to several explanatory variables we require a linear predictor that is a function of several regressors.

- For the logit model,

$$\begin{aligned}\pi_i &= \Lambda(\eta_i) = \Lambda(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}) \\ &= \frac{1}{1 + e^{-(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})}}\end{aligned}$$

  or, equivalently,

$$\log_e \frac{\pi_i}{1 - \pi_i} = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$$

- For the probit model,

$$\pi_i = \Phi(\eta_i) = \Phi(\alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik})$$

▶ The X's in the linear predictor can be as general as in the general linear model, including, for example:

- quantitative explanatory variables;
- transformations of quantitative explanatory variables;
- polynomial regressors formed from quantitative explanatory variables;
- dummy regressors representing qualitative explanatory variables; and
- interaction regressors.

▶ Interpretation of the partial regression coefficients in the general logit model is similar to the interpretation of the slope in the logit simple-regression model, with the additional provision of holding other explanatory variables in the model constant.

- Expressing the model in terms of odds,

$$\begin{aligned}\frac{\pi_i}{1 - \pi_i} &= e^{(\alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik})} \\ &= e^{\alpha} \left(e^{\beta_1}\right)^{X_{i1}} \cdots \left(e^{\beta_k}\right)^{X_{ik}}\end{aligned}$$

- Thus, e^{β_j} is the multiplicative effect on the odds of increasing X_j by 1, holding the other X's constant.

- Similarly, $\beta_j/4$ is the slope of the logistic regression surface in the direction of X_j at $\pi = .5$.

▶ The general linear logit and probit models can be fit to data by the method of maximum likelihood.

▶ Hypothesis tests and confidence intervals follow from general procedures for statistical inference in maximum-likelihood estimation.

- For an individual coefficient, it is most convenient to test the hypothesis H_0: $\beta_j = \beta_j^{(0)}$ by calculating the Wald statistic

$$Z_0 = \frac{B_j - \beta_j^{(0)}}{\text{SE}(B_j)}$$

where $\text{SE}(B_j)$ is the asymptotic standard error of B_j.

  – The test statistic Z_0 follows an asymptotic unit-normal distribution under the null hypothesis.

- Similarly, an asymptotic $100(1-a)$-percent confidence interval for β_j is given by

$$\beta_j = B_j \pm z_{a/2}\text{SE}(B_j)$$

where $z_{a/2}$ is the value from $Z \sim N(0,1)$ with a probability of $a/2$ to the right.

- Wald tests for several coefficients can be formulated from the estimated asymptotic variances and covariances of the coefficients.

- It is also possible to formulate a likelihood-ratio test for the hypothesis that several coefficients are simultaneously zero, H_0: $\beta_1 = \cdots = \beta_q = 0$. We proceed, as in least-squares regression, by fitting two models to the data:

  – The full model (model 1)

$$\text{logit}(\pi) = \alpha + \beta_1 X_1 + \cdots + \beta_q X_q + \beta_{q+1} X_{q+1} + \cdots + \beta_k X_k$$

– and the null model (model 0)

$$\begin{aligned}\text{logit}(\pi) &= \alpha + 0X_1 + \cdots + 0X_q + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k \\ &= \alpha + \beta_{q+1}X_{q+1} + \cdots + \beta_k X_k\end{aligned}$$

– Each model produces a maximized likelihood: L_1 for the full model, L_0 for the null model.

– Because the null model is a specialization of the full model, $L_1 \geq L_0$.

– The generalized likelihood-ratio test statistic for the null hypothesis is

$$G_0^2 = 2(\log_e L_1 - \log_e L_0)$$

– Under the null hypothesis, this test statistic has an asymptotic chisquare distribution with q degrees of freedom.

- A test of the omnibus null hypothesis H_0: $\beta_1 = \cdots = \beta_k = 0$ is obtained by specifying a null model that includes only the constant, logit$(\pi) = \alpha$.

- The likelihood-ratio test can be inverted to produce confidence intervals for coefficients.
- The likelihood-ratio test is less prone to breaking down than the Wald test.

▶ An analog to the multiple-correlation coefficient can also be obtained from the log-likelihood.

- By comparing $\log_e L_0$ for the model containing only the constant with $\log_e L_1$ for the full model, we can measure the degree to which using the explanatory variables improves the predictability of Y.
- The quantity $G^2 \equiv -2 \log_e L$, called the *residual deviance* under the model, is a generalization of the residual sum of squares for a linear model.
- Thus,

$$\begin{aligned} R^2 &= 1 - \frac{G_1^2}{G_0^2} \\ &= 1 - \frac{\log_e L_1}{\log_e L_0} \end{aligned}$$

  is analogous to R^2 for a linear model.

▶ To illustrate logistic regression, I will use data from the 1994 wave of the Statistics Canada Survey of Labour and Income Dynamics (the “SLID”).

- Confining attention to married women between the ages of 20 and 35, I examine how the labor-force participation of these women is related to several explanatory variables:
  - the region of the country in which the woman resides;
  - the presence of children between zero and four years of age in the household, coded as absent or present;
  - the presence of children between five and nine years of age;
  - the presence of children between ten and fourteen years of age
  - family after-tax income, excluding the woman’s own income (if any);
  - education, defined as number of years of schooling.
- The SLID data set includes 1936 women with valid data on these variables.

- Some information about the distribution of the variables:

Variable	*Summary*
Labor-Force Participation	Yes, 79 percent
Region (R)	Atlantic, 23 percent; Quebec, 13; Ontario, 30; Prairies, 26; BC, 8
Children 0–4 (K04)	Yes, 53 percent
Children 5–9 (K59)	Yes, 44 percent
Children 10–14 (K1014)	Yes, 22 percent
Family Income (I, $1000s)	5-number summary: 0, 18.6, 26.7, 35.1, 131.1
Education (E, years)	5-number summary: 0, 12, 13, 15, 20

- To produce “Type-II” likelihood-ratio tests for the terms in the model, I fit the following models to the data:

Model	*Terms in the Model*	*Number of Parameters*	*Residual Deviance*
1	R, K04, K59, K1014, I, E	10	1810.125
2	K04, K59, K1014, I, E	6	1827.423
3	R, K59, K1014, I, E	9	1870.355
4	R, K04, K1014, I, E	9	1820.729
5	R, K04, K59, I, E	9	1810.444
6	R, K04, K59, K1014, E	9	1819.186
7	R, K04, K59, K1014, I	9	1890.480

- Contrasting pairs of these models produces the following likelihood-ratio tests, arrayed in an *analysis of deviance table*:

Term	*Models Contrasted*	df	G_0^2	p
Region (R)	$2-1$	4	17.298	.0017
Children 0–4 (K04)	$3-1$	1	60.230	$\ll .0001$
Children 5–9 (K59)	$4-1$	1	10.604	.0011
Children 10–14 (K1014)	$5-1$	1	0.319	.57
Family Income (I)	$6-1$	1	9.061	.0026
Education (E)	$7-1$	1	80.355	$\ll .0001$

- Retaining the statistically significant terms in the model (all but children 10–14) produces the following final model:

Coefficient	*Estimate* (B_j)	*Standard Error*	e^{B_j}
Constant	-0.3763	0.3398	
Region: Quebec	-0.5469	0.1899	0.579
Region: Ontario	0.1038	0.1670	1.109
Region: Prairies	0.0742	0.1695	1.077
Region: BC	0.3760	0.2577	1.456
Children 0–4	-0.9702	0.1254	0.379
Children 5–9	-0.3971	0.1187	0.672
Family Income ($1000s)	-0.0127	0.0041	0.987
Education (years)	0.2197	0.0250	1.246
Residual Deviance	1810.444		

- This model is summarized in the effect plots in Figure 4.

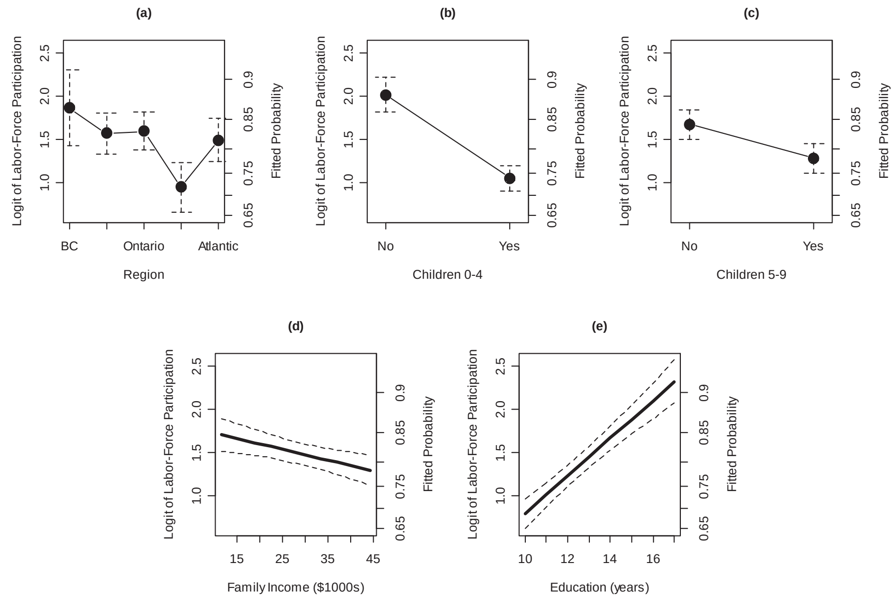

Figure 4. Effect plots for the final model fit to the SLID women's labor-force participation data.

6. Summary

- It is problematic to apply least-squares linear regression to a dichotomous response variable:
  - The errors cannot be normally distributed and cannot have constant variance.
  - Even more fundamentally, the linear specification does not confine the probability for the response to the unit interval.
- More adequate specifications transform the linear predictor $\eta_i = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$ smoothly to the unit interval, using a cumulative probability distribution function $P(\cdot)$.
  - Two such specifications are the probit and the logit models, which use the normal and logistic CDFs, respectively.

- Although these models are very similar, the logit model is simpler to interpret, since it can be written as a linear model for the log-odds:

$$\log_e \frac{\pi_i}{1-\pi_i} = \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}$$

▶ The dichotomous logit model can be fit to data by the method of maximum likelihood.

- Wald tests and likelihood-ratio tests for the coefficients of the model parallel t-tests and F-tests for the general linear model.
- The deviance for the model, defined as $G^2 = -2\times$ the maximized log-likelihood, is analogous to the residual sum of squares for a linear model.