Quantitative Methods in Political Science: Logit and Probit Models

UNIVERSITY OF MANNHEIM
School of Social Sciences

Thomas Gschwend | Oliver Rittmann | Viktoriia Semenova | David M. Grundmanns

Week 11 - 17 November 2021

The Course

Roadmap

- Understand and model stochastic processes
- Understand statistical inference
- Implement it mathematically and learn how to estimate it
  - OLS
  - Maximum Likelihood
- Implement it using software
  - R
  - Basic programming skills

Overview: Week 11

- Modeling Dichotomous Dependent Variables
  - Motivation
  - Limited Dependent Variable Models
  - The Generalized Linear Model Approach
- Estimation
- Interpretation
  - Example: Determinants of Civil War
- Assessing Model Fit

Modeling Dichotomous Dependent Variables

Motivating Binary Dependent Variable Models

- Often our dependent variable is not continuous but binary.
- There are many examples in the social sciences:
  - A voter's choice to go to the polls.
  - A politician's choice to vote "yes" or "no" in legislation (roll call data).
  - A government's decision to implement an EU directive or not.
  - A student's response in an exam can be correct or incorrect.
- In all these cases we have observations on a binary variable, where $y_i = \{0, 1\}$ with $i = 1, ..., n$.
- The basic problem is: How do we estimate regression models when our dependent variable is a dummy?

Notation

Recall that we can write a linear regression model as

$$
\begin{aligned}
Y_i &\sim N(y_i|\mu_i, \sigma^2) && \text{stochastic} \\
\mu_i &= X_i\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots && \text{systematic}
\end{aligned}
$$

We will generalize that and write any statistical model as

$$
\begin{aligned}
Y_i &\sim f(y_i|\theta_i, \alpha) && \text{stochastic} \\
\theta_i &= g(X_i, \beta) && \text{systematic}
\end{aligned}
$$

Modeling Binary Dependent Variables

- Statistical modeling always operates through modeling stochastic processes.
- Hence, we need a probability model that generates "0" and "1" as outcomes.
- We already know the Bernoulli distribution as a discrete distribution.
- This distribution distinguishes between successes (coded as 1) and failures (coded as 0), where the probability to get a success is given as π and the probability for failure is $1 - \pi$.
- Since the Bernoulli distribution takes on only two values as does our dependent variable, we can model each single observation, y_i, as an outcome from a Bernoulli experiment. Hence, our *stochastic component* (with $\theta = \pi$) is

$$Y_i \sim Bernoulli(\pi_i).$$

- With this, we get:

$$\pi_i = P(y_i = 1) = E(y_i) \text{ with density } f(y_i \mid \pi_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

- The density function gives back the probability to either get $y_i = 1$ or $y_i = 0$ for a given success probability π_i.

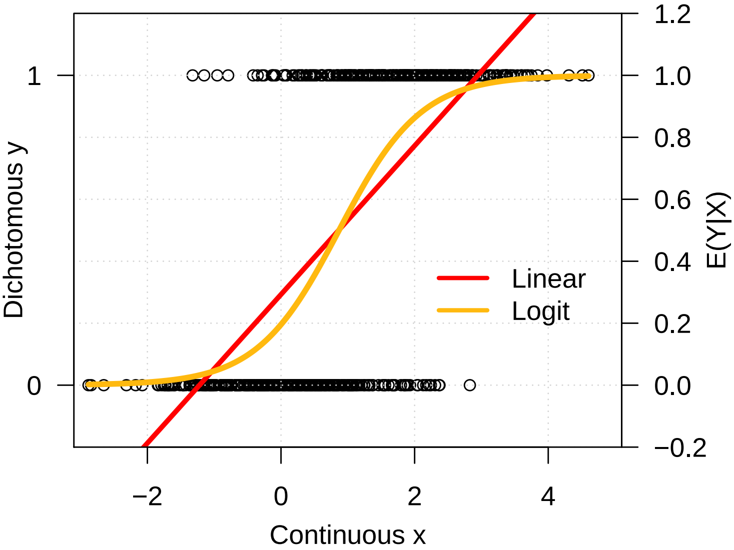
Modeling Binary Dependent Variables

- While we observe a dependent variable, y_i, with $i = 1, ..., n$, our goal is to model the (unobservable) predicted probability $\pi_i (\in [0, 1])$, the expectation of observing $y = 1$ across repeated Bernoulli trials using a function g of covariates, $X = \{x_1, ..., x_k\}$ and respective parameters, i.e. the *systematic component*.
- Thus, the statistical model looks as follows

$$\begin{array}{lcll} Y_i & \sim & \mathit{Bernoulli}(\pi_i) & \text{stochastic component} \\ \pi_i & = & g(X_i, \beta_i) & \text{systematic component} \end{array}$$

- But why not simply use OLS?
  - OLS does not guarantee that predicted probabilities fall into the unit interval.
  - OLS necessarily induces heteroskedasticity since y_i only takes on two values.
  - OLS assumes unrealistic functional form for many applications, i.e., a unit change in x_k results in a constant change in $\hat{\beta}_k$ in the probability of an event holding all other variables constant.
- Hence, we need a different type of model.

Model Predictions from OLS and Logit Model



Generalized Linear Model (GLM) Formulation

- To avoid *out-of-bounds* predictions, i.e., $\eta_i \notin [0, 1]$, we need to force the linear predictor
$$\eta_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} = \mathrm{X}_i \boldsymbol{\beta}$$
to lie inside the unit interval.
- We could do this by restricting the values of $\beta_0, \ldots, \beta_k$. Yet, this results in overly complex models.
- Instead, we choose a function, g, which maps the linear predictor, η_i, into the unit interval.
- We can write this as
$$\pi_i = g(\eta_i) = g(\mathrm{X}_i \boldsymbol{\beta}).$$
- The appealing feature is that this response function, g, "automatically" ensures that the linear predictions η_i lie inside $[0, 1]$.
- The CDF of the logistic distribution function and the normal distribution function are most often used as response functions.

Logit Model

- The *response function* $g(\eta)$ is related to η via the inverse function $h = g^{-1}$, called the link function:

$$\eta_i = h(\pi_i)$$

- If we choose the CDF of the logistic distribution function as a response function, we get

$$\pi_i = g(\eta_i) = \frac{exp(\eta_i)}{1 + exp(\eta_i)} = \frac{1}{1 + exp(-\eta_i)}$$

- The trick now is that we can use our *systematic component* $\mathrm{X}_i\boldsymbol{\beta}$ to reparameterize η_i, which allows us to write the success probability π_i as a function of our covariates and coefficients.
- With this, we get as predicted probabilities:

$$\pi_i = P(y_i = 1) = \frac{exp(\mathrm{X}_i\boldsymbol{\beta})}{1 + exp(\mathrm{X}_i\boldsymbol{\beta})}$$

- This model is referred to as logit model.

Probit Model

- If we do not choose a logistic response function, but use the CDF of the standard normal distribution function ($\mu = 0$, $\sigma = 1$), we get

$$\pi_i = \Phi(\eta_i) = \Phi(X_i\boldsymbol{\beta})$$

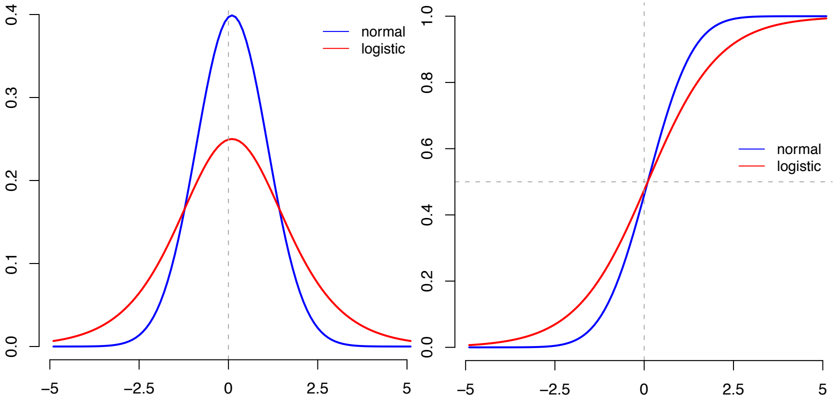
- Again, we can derive predicted probabilities which are given as

$$\pi_i = P(y_i = 1) = \Phi(X_i\boldsymbol{\beta})$$

- This model is referred to as probit model.

Logistic and Standard Normal Distribution

- The logistic distribution has fatter tails.
- Logit and probit coefficients differ by a factor of about 1.81, but produce almost identical results.



Estimation

Likelihood Function for a Binary Dependent Variable

- Consider a binary variable $y_i \sim Bernoulli(\pi_i)$ with
$$\pi_i = P(y_i = 1) = E(y_i).$$
- The density for one realization is given as
$$f(y_i \mid \pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}.$$
- Since all observations y_i are independent realizations the likelihood to observe the data that we did observe is given by the the following expression:
$$L(\boldsymbol{\pi}) = \prod_{i=1}^{n} f(y_i \mid \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$
with
$$\pi_i = E(y_i \mid x_i) = g(\mathrm{X_i}\boldsymbol{\beta}).$$
- When estimating such a model we need to maximize the likelihood L, or for convenience rather $log\ L$ to find those parameter vectors ($\hat{\pi}$ or, $\hat{\boldsymbol{\beta}}$), that most likely generated the data.

Log-Likelihood of the Logit Model

- Thus, taking the log of the likelihood function yields

$$log\ L(\boldsymbol{\pi}|y) = \sum_{i=1}^{n} (y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)) \, .$$

- Using our parameterization of $\boldsymbol{\pi_i} = g(\mathrm{X_i}\boldsymbol{\beta}) = \frac{exp(\mathrm{X}_i\boldsymbol{\beta})}{1+exp(\mathrm{X}_i\boldsymbol{\beta})}$ to include the *systematic component* of the model, the corresponding *log-likelihood function* becomes

$$log\ L(\boldsymbol{\beta}|y, \mathrm{X}) = \sum_{i=1}^{n} (y_i \cdot log(\frac{exp(\mathrm{X}_i\boldsymbol{\beta})}{1 + exp(\mathrm{X}_i\boldsymbol{\beta})}) + (1 - y_i) \cdot log(1 - \frac{exp(\mathrm{X}_i\boldsymbol{\beta})}{1 + exp(\mathrm{X}_i\boldsymbol{\beta})}))$$

- The likelihood is maximized numerically by “hill climbing” algorithms.

Log-Likelihood of the Probit Model

- The probit model has the same stochastic component as the Logit model, hence the log-likelihood function is

$$log\ L(\boldsymbol{\pi}|y) = \sum_{i=1}^{n} (y_i log(\pi_i) + (1 - y_i) log(1 - \pi_i)) \, .$$

- Using our parameterization of $\boldsymbol{\pi_i} = \Phi(\mathrm{X}_i \boldsymbol{\beta})$ to include the *systematic component* of the model, the corresponding *log-likelihood function* becomes

$$log\ L(\boldsymbol{\beta}|y, \mathrm{X}) = \sum_{i=1}^{n} (y_i log(\Phi(\mathrm{X}_i \boldsymbol{\beta})) + (1 - y_i) log(1 - \Phi(\mathrm{X}_i \boldsymbol{\beta}))) \, .$$

Interpretation

Interpreting the Logit Model

- Only the OLS model has nice linear marginal effects.
- Since OLS fits a straight line to the data, the slope of this line is the same for each value of any x_i.
- For the OLS model with $\pi_i = \mathrm{X_i}\boldsymbol{\beta}$, it holds that

$$\frac{\partial \hat{\pi}_i}{\partial x_{ij}} = \hat{\beta}_j$$

- For all other non-linear models this interpretation is however not valid.
- Assume that we have a logit model for which

$$P(y_i = 1) = \hat{\pi}_i = \frac{exp(\mathrm{X_i}\boldsymbol{\beta})}{1 + exp(\mathrm{X_i}\boldsymbol{\beta})}$$

- Clearly, the marginal effect is no longer linear, but is given as

$$\frac{\partial \hat{\pi}_i}{\partial x_{ij}} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

The Log Odds or Logits

- In the logit model it is true that

$$log \left(\frac{P(y_i = 1 \mid \mathrm{x_i})}{P(y_i = 0 \mid \mathrm{x_i})} \right) = log \left(\frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right) = \mathrm{X_i}\boldsymbol{\beta}.$$

- Hence, an estimated coefficient $\hat{\beta}_2 = 2$ in a logit model can be interpreted such that, for a one unit change in x_2, the log ratio of the probability to observe a “1” relative to observing a “0” doubles.
- What does this tell us?
- Let us rather calculate predicted probabilities or other meaningful quantities of interest.

Quantities of Interest

- Instead of directly interpreting coefficients, we usually want to calculate quantities of interest and the uncertainty around them.
- Predicted probabilities (aka expected values) describe the probability of observing an outcome ($\in [0,1]$).
- Predicted values in contrast are on the scale of the dependent variable, i.e., they are either 0 or 1.
- As before, a first-difference is the difference of the expected values (predicted probabilities) of two scenarios.
- Using our simulation techniques we can estimate confidence intervals for predicted probabilities (expected values), first-differences, predicted values and the like.

Quantities of Interest: Predicted Probabilities

- Once more, predicted probabilities for the logit model are given as

$$\hat{\pi}_i = P(y_i = 1) = \frac{exp(X_i\boldsymbol{\beta})}{1 + exp(X_i\boldsymbol{\beta})}$$

- For the probit model, we get predicted probabilities as:

$$\hat{\pi}_i = P(y_i = 1) = \Phi(X_i\boldsymbol{\beta})$$

Example: Determinants of Civil War

- Suppose you are interested in the following research question: Why have some countries had civil wars while others have not?
- To address this questions you could compile a large dataset that contains information if in a certain country-year a civil war took place.
- These so called onset of civil war could be modeled using logit or probit models.
- Take a look at the regression table from Fearon and Latin's 2003 APSR article: Ethnicity, Insurgency, and Civil War. How can you interpret the coefficients?

TABLE 1. Logit Analyses of Determinants of Civil War Onset, 1945–99

	Model				
	(1) Civil War	(2) "Ethnic" War	(3) Civil War	(4) Civil War (Plus Empires)	(5) Civil War (COW)
Prior war	−0.954**	−0.849*	−0.916**	−0.688**	−0.551
	(0.314)	(0.388)	(0.312)	(0.264)	(0.374)
Per capita income[a,b]	−0.344***	−0.379***	−0.318***	−0.305***	−0.309***
	(0.072)	(0.100)	(0.071)	(0.063)	(0.079)
log(population)[a,b]	0.263***	0.389***	0.272***	0.267***	0.223**
	(0.073)	(0.110)	(0.074)	(0.069)	(0.079)
log(% mountainous)	0.219**	0.120	0.199*	0.192*	0.418***
	(0.085)	(0.106)	(0.085)	(0.082)	(0.103)
Noncontiguous state	0.443	0.481	0.426	0.798**	−0.171
	(0.274)	(0.398)	(0.272)	(0.241)	(0.328)
Oil exporter	0.858**	0.809*	0.751**	0.548*	1.269***
	(0.279)	(0.352)	(0.278)	(0.262)	(0.297)
New state	1.709***	1.777***	1.658***	1.523***	1.147**
	(0.339)	(0.415)	(0.342)	(0.332)	(0.413)
Instability[a]	0.618**	0.385	0.513*	0.548*	0.584*
	(0.235)	(0.316)	(0.242)	(0.225)	(0.268)
Democracy[a,c]	0.021	0.013			
	(0.017)	(0.022)			
Ethnic fractionalization	0.166	0.146	0.164	0.490	−0.119
	(0.373)	(0.584)	(0.368)	(0.345)	(0.396)
Religious fractionalization	0.285	1.533*	0.326		1.176*
	(0.509)	(0.724)	(0.506)		(0.563)
Anocracy[a]			0.521*		0.597*
			(0.237)		(0.261)
Democracy[a,d]			0.127		0.219
			(0.304)		(0.354)
Constant	−6.731***	−8.450***	−7.019***	−6.801***	−7.503***
	(0.736)	(1.092)	(0.751)	(0.681)	(0.854)
N	6327	5186	6327	6360	5378

Note: The dependent variable is coded "1" for country years in which a civil war began and "0" in all others. Standard errors are in parentheses. Estimations performed using Stata 7.0. *$p < .05$; **$p < .01$; ***$p < .001$.

[a] Lagged one year.

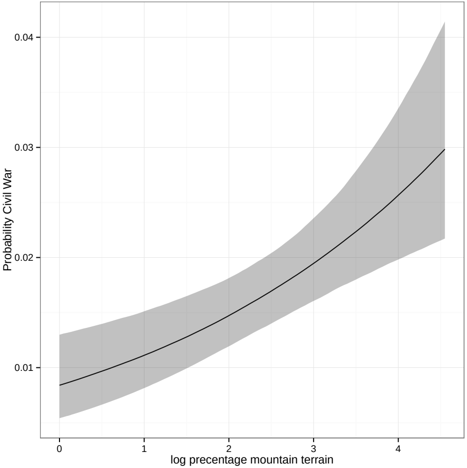
[b] In 1000's.

[c] Polity IV; varies from −10 to 10.

[d] Dichotomous.

The effect of mountain terrain of probability civil war

- Better to use simulated probabilities.



Hypothesis Testing in Non-Linear Models: Wald Test

- Assume, we want to test

$$H_0: \ \beta^* = \tilde{\beta} \text{ against } H_1: \beta^* \neq \tilde{\beta}.$$

- The Wald test is a generalization of the standard t-test that we know from linear models.
- The Wald test statistics is calculated as

$$\mathcal{W}^2 = \frac{(\hat{\beta} - \beta^*)^2}{Var(\hat{\beta})}, \text{ with } \mathcal{W} \sim \chi^2_{df=1}.$$

- For significance tests against $\beta^* = 0$, the Wald statistic becomes

$$\mathcal{W} = \frac{\hat{\beta}}{SE(\hat{\beta})}, \text{ with } \mathcal{W} \sim \mathcal{N}(\mu = 0, \sigma = 1),$$

  which is now distributed standard normal.
- In the “OLS world” the Wald test and the t-test are conceptually equivalent.

Hypothesis Testing in Non-Linear Models: LR Test

- The likelihood ratio test allows to test two *nested* models against each other, which have some common covariates (but one model is a special case of the other model).
- The likelihood ratio test statistic is constructed as

$$LR = -2\ log(\frac{L(\boldsymbol{\beta})^R}{L(\boldsymbol{\beta})^U}) = -2\left(log\ L(\boldsymbol{\beta})^R - log\ L(\boldsymbol{\beta})^U\right),\ \text{with } LR \sim \chi^2_{df=u-r},$$

  where $log\ L(\boldsymbol{\beta})^R$ denotes the log likelihood of the restricted model (the special case), $log\ L(\boldsymbol{\beta})^U$ denotes the log-likelihood of the unrestricted model.
- The test statistic, LR, is distributed χ^2 with the difference in model parameters between the unrestricted and the restricted model, i.e., the number of restrictions $u - r$, as degrees of freedom. H_0 is no difference between models.
- In the "OLS world", the likelihood ratio test and the F-test are conceptually equivalent.

Assessing Model Fit

Classification Table for Logit and Probit Models

- Simply running a model without testing for model fit is dangerous.
- An easy test is to classify predicted probabilities as either "0" or "1" depending on some cut-point c. Usually, the cut-point is chosen to be .5.
- Given this, we can tabulate predicted and observed observations in a 2x2 classification table (aka confusion matrix).

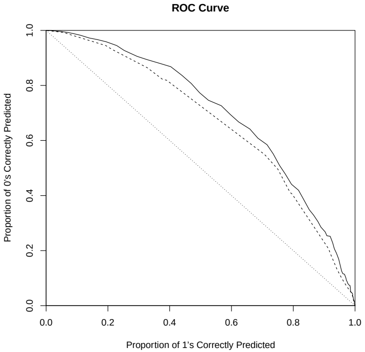
	Predicted ($\hat{y}_i$)	
Observed (y_i)	0	1
0	n_{00}	n_{01}
1	n_{10}	n_{11}

- From this, we can construct a measure for the percentage of correctly predicted cases (PCP):

$$PCP = \frac{n_{00} + n_{11}}{n_{00} + n_{01} + n_{10} + n_{11}}$$

- If, say, the DV is distributed 70 : 30, then just fixing the prediction to one would predict 70% of the cases correctly. Thus, your model should do better than that.

ROC Curves



- Plot percentage of correctly predicted “1s” and “0s”against each other.
- The further the curve is shifted to the northeast corner, the better the model fit.
- This method is insensitive to the exact choice of the cutoff.
- The area under the curve is often used as a measure of fit.