3

Maximum Likelihood for Binary Outcomes

3.1 BINARY RESPONSES

The normal distribution is the starting place for many analyses for a variety of good reasons: it arises naturally in repeated sampling situations; it is flexible; and it crops up frequently in nature. But many concepts in social science and public policy are inherently categorical or bounded. Voting is often a choice between two (frequently terrible) discrete alternatives. By and large, survey respondents are in or out of the labor force; individuals are married or not married. Countries experience a civil war or they do not. In some cases, there may be a plausibly continuous range of values that are interesting, such as the level of conflict between two countries, but we may only observe that quantity if it is sufficiently large to be recognizable as a militarized dispute or a war. In these examples, observations contain information about social phenomena measured at a binary level. Often, this is recorded as 0 for no/not/absent and 1 for yes/present, but these numbers are a convenience, not a necessary structure. Votes in the US Senate are recorded as *Yea* and *Nay*, for example, and, by convention, translated into 1*s* and 0*s*. But the translation could be to any two arbitrary digits or symbols because it is not the numerical characteristics we can use. The 1s are not one unit larger than the 0s. Rather, these two digits are used in a set theoretic fashion: cases are in the set of 1s or in the set of 0s, exclusively and exhaustively.

Binary data can be organized in a variety of ways, among which two are dominant. The social sciences traditionally use a case-based orientation, in which each row is an individual case and each column represents a variable on which each case is observed. It is common in other fields (e.g., medicine) to organize the data by the covariate class or stratum, with the binary variable reported as the number of "successes" among the total number of cases in each class. Table 3.1 illustrates these two data structures.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.1 *Alternative data structures for binary data.*

Case	Covariates		Response	Covariates	Class Size	Response
	$x_{i,1}$	$x_{i,2}$	y_i	$(x_1, x_2)_k$	m_k	y_k
i	Gender	Employed	Voted	(Gender, Employed)		Voted
1	M	yes	no	(M,yes)$_1$	2	1
2	M	no	yes	(M,no)$_2$	4	3
3	M	no	no	(F,yes)$_3$	1	0
4	F	yes	no	(F,no)$_4$	2	1
5	F	no	no			
6	F	no	yes			
7	M	no	yes			
8	M	no	yes			
9	M	yes	yes			

In the left panel of Table 3.1 we see the traditional case-based orientation. On the right, we display the same data using the grouped binary data orientation. Responses are reported in the form of the number of successes (y_k) in each of k covariate classes, out of m_k possible successes, such that $0 \leq y_k \leq m_k$.

Grouped versus ungrouped is an important distinction even though the underlying binary data may be identical. Ungrouped (case-based) data has a natural connection with the Bernoulli distribution, whereas grouped data are easily described using the closely related binomial distribution. With grouped data, the normal approximation is more readily useful; asymptotic assumptions can be based on imagining that m or n approach ∞, although grouped data become increasingly cumbersome as the number of covariates (and therefore combinations thereof) increase, especially when covariates are continuously valued.

This simple example holds another tiny point that will reemerge often in the analysis of nominal data: more often than not, the data are not actually recorded as numbers but as discrete categories of labels, often referred to as *factors*. Some software packages deal with this nuance automatically, but often the analyst must translate these words into integers. For the record, the integer 1 is almost always the code for the occurrence of the event in probability space.

3.2 BINARY DATA

As a running example, we consider the Mroz (1987) study of female labor force participation. Although the example is a bit dated, the data have several attributes that make them worth revisiting here. Mroz analyzed data taken from the Panel Study of Income Dynamics (PSID) on 753 women and their experience in the labor market in 1975. Mroz was interested in female labor

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.2 *Select variables from Long's (1997) reexamination of Mroz's (1987) female labor force participation study.*

Variable Name	Description
LFP	1 if respondent is in paid labor force; else 0
young kids	Number of children younger than 6
school kids	Number of children ages 6 to 18
age	Age in years
college	1 if attended college; else 0
wage	Woman's expected after tax wage rate (logged)

force participation, i.e., the decision to seek employment in the formal paid labor market. For this example we use five of the variables described in Long (1997, p. 37), as shown in Table 3.2.

3.2.1 Odds, Odds Ratios, and Relative Risk

Odds are one way of expressing probabilities. If you flip a (fair) coin twice, what is the probability of getting two heads? The probability is $\frac{1}{2} \times \frac{1}{2} = 0.25$, but we might also say that the *odds* are one-to-three. In other words we should expect one pair of heads for every three "failures." Formally,

$$\omega_i \equiv \text{Odds}(y_i = 1) = \frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)}.$$

Notice that probabilities and odds are both bounded by zero at the lower end. But odds are unbounded at the upper end, while probabilities are bounded by 1; this means that the odds are not symmetric, but skewed. As a result, the odds is frequently made even more odd by taking the natural logarithm:

$$\log \omega_i = \log \left[\frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)} \right].$$

Another less commonly used but arguably more intuitive quantity is *relative risk*, also called the *risk ratio*. Relative risk is simply the probability of one event divided by the probability of another. In the coin example the probability of two heads is the same as the probability of two tails, so the relative risk is 1.

The Mroz data consists of 753 married white women between the ages of 30 and 60,428 of whom were in the paid labor force at some time during 1975. The probability of a white American woman being in the labor force in 1975, p, is estimated by the proportion of women in the sample who participated in the labor force during that year, $\hat{p} = \frac{428}{753} = 0.568$. By extension, the probability that a woman is not employed would be estimated as $1 - \hat{p} = 0.432$. The ratio

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.3 *Labor force participation and the number of young children.*

	Young Children			
In Labor Force?	**0**	**1**	**2**	**3**
No	231	72	19	3
Yes	375	46	7	0
Odds of LFP	1.62	0.64	0.37	0.00
Risk, relative to 0 young children	1.00	0.63	0.44	0.00

of these two numbers is the odds, ω. The odds of a woman being employed in these data is:

$$\hat{\omega} = \frac{\hat{p}}{1-\hat{p}} = \frac{0.568}{1-0.568} = 1.3,$$

which means that it is about 30% more likely that a woman is employed than not.

Consider the following data from the Mroz study, given in Table 3.3. How does the probability that a woman has young children influence her odds of being a paid participant in the labor force? Compare the odds for being in the labor for women with no young children, to those with one young child. Such calculations are given as:

- The probability of labor force participation (LFP) for a woman with *zero* young children is $\hat{p}_0 = \frac{375}{(375+231)} = 0.619$, which implies the odds of $\hat{\omega}_0 = \frac{0.619}{1-0.619} = 1.62$.
- The probability of LFP for a woman with *one* young child is $\hat{p}_1 = \frac{46}{(46+72)} = 0.39$, which implies the odds of $\hat{\omega}_0 = \frac{0.39}{1-0.39} = 0.64$. Her risk of LFP relative to a woman with no young children is $\frac{0.39}{0.619} = 0.63$.
- The probability of LFP for a woman with *two* young children is $\hat{p}_2 = \frac{7}{(7+19)} = 0.27$, which implies the odds of $\hat{\omega}_0 = \frac{0.27}{1-0.27} = 0.37$. Her risk of LFP relative to a woman with no young children is $\frac{0.27}{0.619} = 0.44$.
- The probability of employment for a woman with *three* young children is $\hat{p}_3 = \frac{0}{(3)} = 0.0$, which implies the odds of $\hat{\omega}_0 = \frac{0}{1-0.0} = 0.0$. Her risk of LFP relative to a woman with no young children is $\frac{0}{0.619} = 0$.

This illustrates that the odds of being employed are about 1.6 (to 1) for women with no young children, whereas these odds fall to about 0.6 for those with just one child younger than six years of age, and further fall to approximately 0.4 if there are two young children. A woman with one young

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

child has a 37% lower probability of being in the labor force than a woman with no young children.

In some fields, especially medicine, people are interested in the *odds-ratio*. For example, the odds-ratio for being employed given you have zero versus one young child is simply $\frac{1.62}{0.64} = 2.56$, which means that the odds a woman is employed if she has no young children is about two and one-half times the odds of her being employed if she has a single child under five. Similarly, a woman with one young child is not quite twice as likely to be employed as a woman with two young children ($\frac{0.64}{0.37} = 1.73$). But all these women are *enormously* more likely to be employed than a woman with three young children under five.

The odds are notoriously difficult to describe and interpret. There is extensive empirical work documenting how people's interpretation of risk assessments, including the odds and odds ratios, do not coincide with probability theory.

3.3 THE LINEAR PROBABILITY MODEL

The *linear probability model* fits a linear regression model to a binary response variable, often using OLS. Mroz (1987) is a famous application of the linear probability model. In this section we follow Long (1997), showing what can go wrong with pressing the OLS button when your dependent variable cannot be considered unbounded, much less normally distributed.

We fit a linear regression model with LFP as the dependent variable and the four other variables as predictors. While recognizing that the dependent variable is binary, the OLS estimates produce what is called a linear probability model, which is specified as

$$\Pr(Y_i = 1 \mid \mathbf{x}_i) = \mathbf{x}_i^{\top} \boldsymbol{\beta}$$

where $\mathbf{X} = (\mathbf{1}, \text{young kids, school kids, age, college})$, i.e., a matrix with a row for each observation and a column for each of the four independent variables along with a vector of ones for the intercept. The term $\mathbf{x}_i$ is a vector of length 5 for observation i. The vector $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$ contains the intercept and slope parameters to be estimated. Using OLS to fit the model produces the results presented in Table 3.4.

This table shows that families with more children under age five are less likely to have a female in the paid labor force, at least in 1975. The estimated coefficient for this relationship is -0.3. One might be tempted to state that each child under five will reduce by about 0.3 the probability that a woman will be in the paid labor force.

A negative relationship seems reasonable, but the model predicts that a mother of 3-year-old quintuplets will have a negative probability of being in the labor force. While holding a full-time paid job seems difficult in such a situation, can the probability be less than zero? In the observed data the

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.4 *Linear probability model of labor force participation (LFP) as a function of five independent variables plus an intercept.*

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	*t*-Ratio	*p*-Value
intercept	1.10	0.13	16.53	0.00
young kids	−0.30	0.04	−8.37	0.00
school kids	−0.02	0.01	−1.17	0.24
age	−0.01	0.00	−5.60	0.00
college	0.13	0.04	3.19	0.00
wage	0.12	0.03	3.82	0.00
n	753			
$\hat{\sigma}$	0.46			
$\overline{R}^2$	0.13			

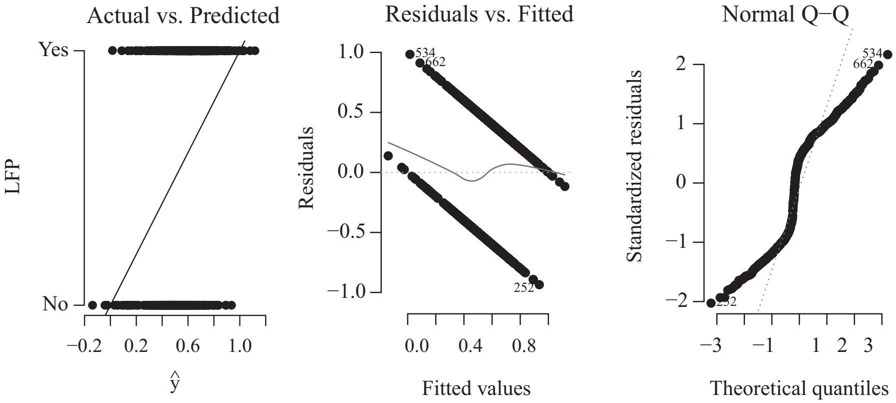

FIGURE 3.1 Diagnostic plots of the linear probability model described in Table 3.4.

model produces many predictions outside the [0,1] interval. Three predicted probabilities are below zero, with the largest at −0.14, indicating that one woman has a negative fourteen percent chance of being employed, clearly nonsensical. Similarly, there are five predictions that are greater than one, including the largest at 1.12.

It is clear that the linear probability model can and will produce results that are nonsensical and difficult to explain to the secretary of labor should she ask you how you arrived at a negative probability. Other problems arising with the LPM appear in the diagnostic plots in Figure 3.1. The left panel shows the dependent variable plotted against the predicted values of the dependent variable, along with the line representing the regression line. All the true values of the dependent variable are at the top and the bottom of this plot, but the prediction line takes on continuous, linear values both inside the range $[0, 1]$ and outside. The middle panel shows two clusters of

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

residuals, one representing the values when the predicted values are subtracted from 1, another when predicted values are subtracted from 0. These residuals clearly cannot be normally distributed, nor do several other conditions for the linear model remain credible. In particular, the variance of the residuals around the regression line is not constant but is rather grouped in two clumps that correspond to the 1*s* and 0*s* in the data. The model is heteroskedastic by construction since the calculated variance of the dependent variable as estimated depends not only on the values of the independent variables but also on the estimated values of the parameters ($\hat{\beta}$). The bottom line is that the model produces results that are nonsensical in the domain of prediction, even if the coefficients may remain unbiased.

In short, the LPM:

- runs the risk of making nonsensical predictions. If any of the predicted probabilities from the LPM are outside the $[0, 1]$ interval, then the OLS estimate of the LPM is biased and inconsistent (Horrace and Oaxaca, 2006). The more nonsensical predictions the model makes, the more biased the estimates are;
- assumes that the relationship between a covariate and the probability of a success is constant across all values of the covariate;
- will generate residuals with inherent heteroskedasticity that give biased confidence intervals for parameter estimates.

With these weaknesses, why do we still see the LPM employed, especially in some subfields of economics? There are several arguments, some more convincing than others:

- In the situation where we actually have a randomly assigned treatment and all we care about is the average treatment effect, then OLS and logit/probit will give effectively identical answers. Moreover, under binary randomized treatment, the functional form problems do not affect the ability to estimate treatment effects. So why bother with the complications? (Angrist and Pischke, 2009, p. 107)
- Recent emphasis on causal identification has led to a reliance on so-called fixed effects in short panels and a myopic focus on "marginal effects." These are arguably better implemented in an OLS framework. Similarly, *instrumental variables* strategies are currently easier to implement and have some desirable properties under a LPM. We do not discuss these topics in this volume.
- Heckman and Snyder, Jr. (1997) provide a rigorous justification for the LPM as an exact representation of a specific class of random-utility models which impose asymmetric random shocks on decision makers' utilities. In other words, the LPM is sometimes an attractive procedure for theoretical or aesthetic reasons.
- Some argue that the LPM is a simple approach that is feasible when there are too many observations to maximize a likelihood, even with modern

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

computing power. One colleague ran a logistic regression on a database with 50 million observations, using SAS. So if you have more than 50,000,000 cases, maybe this is justifiable.

- The LPM will not necessarily give nonsensical predictions in all situations, so in cases where this does not happen it may be useful. But how can you know this unless you actually check the alternative?
- Standard errors can be "corrected" ex post for the heteroskedasticity (see Section 2.4.1), so the biased standard error problem goes away. However, you can transform virtually anything post-estimation.
- Some argue for the LPM because of the apparent arbitrariness of the distributional assumptions of logit and probit models, relying on the fact that classical OLS makes weaker distributional assumptions about Y and the error process.
- Some argue that LPM coefficients are "easy" to interpret, whereas regression coefficients for logit or probit are "hard" or require more effort to communicate to audiences.

Our position is that statistical models are fit for a variety of reasons, only some of which have causal identification as their objective. Ultimately any statistical model is a description of a DGP that combines data and a set of assumptions. As such, we favor evaluating statistical models using comparative, predictive heuristics as we discuss in detail in Chapter 5. This emphasis on model comparison and prediction generally leads us to prefer versions of logit and probit in most applied circumstances. Logit and probit models may have drawbacks in some situations, but, in our experience the LPM is virtually never a superior *predictive* model for binary data.

3.4 THE LOGIT MODEL, A.K.A. LOGISTIC REGRESSION

What if we made different assumptions about the process generating our (binary) data? The logit model is one way of doing so.

The introduction of logistic regression is typically attributed to David Cox (1958). Cox credits the introduction of the term to Joseph Berkson. The *logit* transformation relies on the odds, as developed above, to map a probability into the unbounded real line. A logit transformation of the variable $p \in [0,1]$ is given as the log of the odds:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

The logistic transformation, also known as the inverse logit, accomplishes the goal of mapping real-valued variable, x, into the $[0,1]$ interval appropriate for probabilities:

$$\text{logit}^{-1}(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}.$$

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

In the next section, this transformation is used to develop a standard, nonlinear approach to modeling binary dependent variables.

3.4.1 The Logit Model

Binary outcomes are modeled as Bernoulli trials. A Bernoulli random variable (e.g., a coin flip) is defined by a single parameter, θ, which describes the probability of a "success."

Using our standard setup, we can express such a Bernoulli model in terms of two components:

$$\text{stochastic: } Y_i \sim f_B(y_i; \theta_i)$$

or, alternatively,

$$\Pr(Y_i = y_i) = \theta_i^{y_i}(1-\theta_i)^{1-y_i} = \begin{cases} \theta_i & \text{for} \quad y_i = 1 \\ 1-\theta_i & \text{for} \quad y_i = 0 \end{cases}$$

and

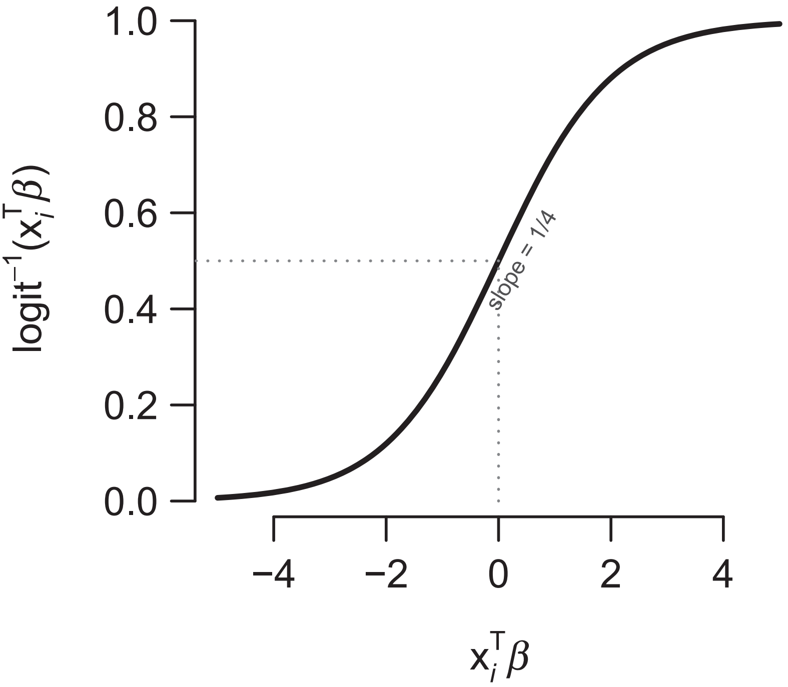

FIGURE 3.2 The logistic function is nonlinear with range bounded by 0 and 1. It is nearly linear in the midrange, but is highly nonlinear towards extrema.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

$$\text{systematic: } \theta_i \equiv \text{logit}^{-1}\left(\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}\right) = \frac{1}{1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}},$$

where for each observation i, Y_i is the binary dependent variable, $\mathbf{x}_i$ is the k-vector of $k-1$ independent variables and a constant, and $\boldsymbol{\beta}$ is a vector of k regression parameters to be estimated, as always. Since the mean of a Bernoulli variable must be between 0 and 1, the inverse logit function is one way to link the systematic component to the outcome domain. For reference, some presentations will highlight the inverse logit, while others will present material in terms of the logit.

The inverse logit transformation is shown graphically in Figure 3.2. The x-axis represents the value of $\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}$, called the *linear predictor*, while the curve portrays the logistic mapping of these linear predictor values into the $[0,1]$ probability interval on the y-axis. The midpoint of the line is at a probability of 0.5 on the y-axis. The logistic curve is nearly linear around this point with a slope of approximately $\frac{1}{4}$.

Given the systematic and stochastic components of this kind of model, the joint probability of the data is straightforward:

$$\Pr(\mathbf{y} \mid \boldsymbol{\theta}) = \prod_{i=1}^{n} \theta_i^{y_i}(1-\theta_i)^{1-y_i}$$

which gives rise to the log-likelihood,

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{y}) &= \sum_{i=1}^{n}\left[y_i \log\theta_i + (1-y_i)\log(1-\theta_i)\right] \\
&= \sum_{i=1}^{n}\left[y_i \log\left(\frac{1}{1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right) + (1-y_i)\log\left(1-\frac{1}{1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)\right] \\
&= \sum_{i=1}^{n}\left[-y_i \log(1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}) + (1-y_i)\log\left(\frac{e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}{1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right)\right] \\
&= \sum_{i=1}^{n}\log\left(\frac{\left(e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}\right)^{(1-y_i)}}{\left(1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}\right)^{(1-y_i)}\left(1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}\right)^{y_i}}\right) \\
&= \sum_{i=1}^{n}\log\left(\frac{e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}(1-y_i)}}{1+e^{-\mathbf{x}_i^\mathsf{T}\boldsymbol{\beta}}}\right).
\end{aligned}$$

3.4.2 Estimation

This particular class of problems is easy to solve by the computation of numerical derivatives, but it may be helpful to know that the partial derivative of the log-likelihood with respect to a particular slope parameter, β_j, is given by

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 3.1 *MLE for Logit*

```
# getting data
colnames(lfp)<-c("LFP","young.kids","school.kids","age",
  "college.woman", "college.man", "wage", "income")
lfp$lfpbin<-rep(0,length(lfp$LFP))
lfp$lfpbin[lfp$LFP=="inLF"]<-1
attach(lfp)
x<-cbind(young.kids, school.kids, age,
  as.numeric(college.woman)-1, wage)
y<-lfpbin

# simple optimization of logit for Mroz data
binreg <- function(X,y){
   X <- cbind(1,X)                 #1s for intercept
   negLL <- function(b,X,y){       #the log-likelihood
       p<-as.vector(1/(1+exp(-X %*% b)))
       - sum(y*log(p) + (1-y)*log(1-p))
       }
   #pass the likelihood to the optimizer
   results <- optim(rep(0,ncol(X)), negLL, hessian=T, method="BFGS", X=X,y=y)
   list(coefficients=results$par,varcovariance=solve(results$hessian),
         deviance=2*results$value,
         converged=results$convergence==0) #output an convergence check.
   }

mlebin.fit<-binreg(x,y)
# some results...
round(mlebin.fit$coefficients,2)
>[1]  2.88  -1.45  -0.09  -0.07   0.61   0.56
```

$$\frac{\partial \log \mathcal{L}}{\partial \beta_j} = \sum_{i=1}^{n} (y_i - \theta_i) x_{ij}.$$

The gradient of the logit log-likelihood is then a vector of k such partial derivatives.

In Example 3.1 we present a simple $\mathcal{R}$ program that implements the logistic regression log-likelihood and finds the MLE. The matrix **X** contains the independent variables from the Mroz study and **y** holds the 0s and 1s reflecting whether or not each of the 753 women surveyed is an active participant in the wage labor force.

These results can also be obtained more readily via the `glm()` procedure in $\mathcal{R}$. We go in to more of the details of the $\mathcal{R}$ syntax for `glm` in Section 7.3.1 but we introduce it here because the function – and $\mathcal{R}$'s formula notation – is used in subsequent chapters. The `glm` function handles categorical or "factor" variables without forcing the user to translate them into numbers. Estimation and output appear in Example 3.2.

At the top of the $\mathcal{R}$ output in Example 3.2, we see some basic information on the model. "Call" simply reproduces the formula that was used to fit the model, for ease of reference. The next set of numerical results describes the residual

$\mathcal{R}$ Code Example 3.2 *Using `glm` to estimate logit*

```
fit.glm<−glm(lfpbin ~ young.kids + school.kids + age +
              college.woman + wage,
              family = binomial(link = "logit"), data = lfpdata)
summary(fit.glm)
#which results in the following output
Call:
glm(formula = lfpbin ~ young.kids + school.kids + age + college.woman +
    wage, family = binomial(link = "logit"), data = lfp)

Deviance Residuals:
    Min       1Q    Median      3Q      Max
−2.0970  −1.1148   0.6463   0.9821   2.2055

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)            2.87807    0.62291   4.620 3.83e−06
young.kids            −1.44670    0.19363  −7.471 7.94e−14
school.kids           −0.08883    0.06676  −1.331 0.183336
age                   −0.06796    0.01246  −5.452 4.99e−08
college.womanCollege   0.61112    0.19372   3.155 0.001607
wage                   0.55867    0.14893   3.751 0.000176

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1029.75  on 752  degrees of freedom
Residual deviance:  925.31  on 747  degrees of freedom
AIC: 937.31

Number of Fisher Scoring iterations: 4
```

distribution. The default is for `glm` to return *deviance residuals*, which is each observation's contribution to the residual deviance as defined in Section 2.3. We can see here that the deviance residuals are not centered on 0 and do not appear to be symmetric.

The main output is the description of the regression coefficients and their estimated standard errors. Below that $\mathcal{R}$ tells us that it fit a canonical Generalized Linear Model with an assumed dispersion parameter; we take this up in Chapter 7. The bottom section of the output displays basic model fit information, including the AIC. Finally, the summary output tells us how many steps the optimizer needed to arrive at its answer. In this case, it was four; we go into details in the next chapter.

3.4.3 Output and Interpretation

The standard output from a logistic regression is the same as for most regression models, a BUTON, shown in Table 3.5 and reproducing (in a cleaned-up form) the information returned by $\mathcal{R}$'s `summary` command.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.5 *Logit estimation of female labor force participation (LFP).*

	$\hat{\beta}$	$\sigma_{\hat{\beta}}$	z-Ratio	p-Value
intercept	2.88	0.62	4.62	0.00
young kids	−1.45	0.19	−7.47	0.00
school kids	−0.09	0.07	−1.33	0.18
age	−0.07	0.01	−5.45	0.00
college	0.61	0.19	3.16	0.00
wage	0.56	0.15	3.75	0.00
n	753			
AIC	937			

While the signs and p-values of these estimates are similar to those found using the LPM, the estimates themselves are quite different. These estimates, as presented in the table, require more care in interpretation for two basic reasons. First, the underlying model is nonlinear, so, unlike OLS, the effect of a particular covariate on the response is not constant across all levels of the independent variable. To see this, we calculate the *marginal effect*, or the rate of change of the outcome variable with respect to a particular independent variable, x_k. We take the derivative of the systematic component with respect to the independent variable of interest:

$$\frac{\partial \mathrm{E}[Y_i]}{\partial x_{ki}} = \frac{\partial \theta_i}{\partial x_{ki}} = \beta_k \frac{\exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})}{\left(1 + \exp(\mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta})\right)^2}.$$

This equation shows that the change in the predicted outcome induced by a change in x_k is reflected not just in the regression coefficient β_k. The marginal effect also depends on the value of x_k *and the values of all the other covariates in the model*. In other words, a variable's marginal effect in a logit model is not constant; it depends on the covariate values at which it is evaluated.

As a first-order approximation, we can use the fact that the logistic curve is steepest in the middle. Since the slope of the inverse logit is 0.25 at that point, dividing $\hat{\beta}_k$ by 4 gives an estimate of the maximum difference a one-unit change in x_k can induce in the probability of a success (Gelman and Hill, 2007, p. 82). Thus, an additional young child reduces the probability of labor force participation by about 40%, i.e., $\frac{-1.45}{4} = -0.4$.

Second, the logit model is a linear regression on the log odds. As a result, the exponentiated regression coefficients are odds ratios; a coefficient greater than 1 represents an increase in the relative probability of obtaining a 1 in the dependent variable. An exponentiated coefficient less than 1 represents a decrease in the probability of, in this case, being employed.

This exponentiation trick can be useful for analysts, but if you want to confuse someone – say a student, client, or journalist – try describing your

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

results in terms of either log odds or odds ratios. In some fields, such as medicine, these are routinely reported, and scholars in these fields seem to have a firm handle on their meaning, but in general it is better to interpret your results in terms of the original scales on which the data were measured. To this end, two different approaches have evolved for interpreting logistic regression results on the probability scale.

The Method of "First Differences"

The method of first differences estimates the "effect" of x_k on E[Y] by calculating the change in the predicted Y for different values of x_k. The method is generally useful if the analyst is interested in interpreting and evaluating the regression coefficients near the central tendencies of the independent variables. In calculating central tendencies we typically employ the median for ordered or continuous variables and the mode for categorical variables, although other choices are certainly possible. The central tendencies for all independent variables in the Mroz data are displayed in the first column of Table 3.6. The abbreviations in parentheses are used in some equations below.

Using these values in combination with the estimated coefficients facilitates calculation of a predicted probability for a "typical" respondent. That respondent is represented as a set of specific values for the covariates: a 42 and one-half year old woman, with no young children, but one child between 6 and 18. She has not attended college, and her log wage rate is 1.1. This vector of covariate values represents a specific *scenario* on which to calculate the model's implications.

In the Mroz example, we can evaluate the probability of being in the labor force depending on whether the woman attended college, holding all the other independent variables at their respective central tendencies. We let $\bar{\mathbf{x}}_{\neg\text{coll}}$ represent the vector of central tendencies for all variables except college education. If we are interested in the implied consequence of having attended college, we difference the following two equations:

TABLE 3.6 *The central tendencies (medians and modes) for the variables included in the analysis of labor force participation in Table 3.5.*

Variable	Central Tendency (full)	No College	College
intercept	1.00	1.00	1.00
young kids (k5)	0.00	0.00	0.00
school kids (k618)	1.00	1.00	1.00
age	42.54	43.00	41.50
college (coll)	0.00	0.00	1.00
wage	1.10	0.98	1.40

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

$$
\begin{aligned}
&\Pr\left(y = 1 \mid \text{coll} = 0, \overline{\mathbf{x}}_{\neg\text{coll}}\right) \\
&\quad = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\overline{\text{k5}} + \hat{\beta}_2\overline{\text{k618}} + \hat{\beta}_3\overline{\text{age}} + \hat{\beta}_4 * 0 + \hat{\beta}_5\overline{\text{wage}})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\overline{\text{k5}} + \hat{\beta}_2\overline{\text{k618}} + \hat{\beta}_3\overline{\text{age}} + \hat{\beta}_4 * 0 + \hat{\beta}_5\overline{\text{wage}})} \\
&\Pr\left(y = 1 \mid \text{coll} = 1, \overline{\mathbf{x}}_{\neg\text{coll}}\right) \\
&\quad = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1\overline{\text{k5}} + \hat{\beta}_2\overline{\text{k618}} + \hat{\beta}_3\overline{\text{age}} + \hat{\beta}_4 * 1 + \hat{\beta}_6\overline{\text{wage}})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1\overline{\text{k5}} + \hat{\beta}_2\overline{\text{k618}} + \hat{\beta}_3\overline{\text{age}} + \hat{\beta}_4 * 1 + \hat{\beta}_6\overline{\text{wage}})},
\end{aligned}
$$

which more simply (and generally) is

$$
\begin{aligned}
\Pr\left(y = 1 \mid \text{coll} = 0, \overline{\mathbf{x}}_{\neg\text{coll}}\right) &= \left.\frac{\exp\left(\overline{\mathbf{x}}^{\mathsf{T}}_{\neg\text{coll}}\hat{\boldsymbol{\beta}}\right)}{1 + \exp\left(\overline{\mathbf{x}}^{\mathsf{T}}_{\neg\text{coll}}\hat{\boldsymbol{\beta}}\right)}\right|_{\text{coll}=0} \\
\Pr\left(y = 1 \mid \text{coll} = 1, \overline{\mathbf{x}}_{\neg\text{coll}}\right) &= \left.\frac{\exp\left(\overline{\mathbf{x}}^{\mathsf{T}}_{\neg\text{coll}}\hat{\boldsymbol{\beta}}\right)}{1 + \exp\left(\overline{\mathbf{x}}^{\mathsf{T}}_{\neg\text{coll}}\hat{\boldsymbol{\beta}}\right)}\right|_{\text{coll}=1} .
\end{aligned}
$$

Substituting these values into the linear predictor, $\mathbf{x}_i^{\mathsf{T}}\hat{\boldsymbol{\beta}}$, and then using the inverse logit transformation, we calculate $\hat{\theta}_i$ for this "typical" respondent as 0.63. A similar calculation in which the woman attended college produces 0.75. The difference between these two scenarios is 0.13; the model predicts that having attended college increases the probability of being in the labor force by 20%, holding all the other measured attributes of the woman at "typical" levels.

But suppose we want to take account of the fact that women in the sample who have attended college look different than those who did not on a variety of dimensions. For example, Table 3.6 shows that college-attending women have a median wage rate of 1.4, against a median of 0.98 for women who did not attend. To make a comparison in this case we construct two more scenarios. In the first the values of the covariates take on those for a "typical" woman who never attended college, as shown in the second column of Table 3.6. The model predicts that such a woman has a 0.60 probability of being in the labor force. The second scenario uses the typical covariate values for college-attending women, as displayed in the third column of the table. This woman is predicted to have a 0.79 probability of being in the labor force, for a difference of 0.19 between these scenarios. The typical women who attended college in 1973 was 24% more likely to be in the labor force than the typical woman who did not.

These examples illustrate how the interpretation of nonlinear models requires the analyst to decide what types of comparisons best illustrate the model's implications for the purpose at hand. The structure of the model reinforces a more general point: there is no single-number summary that communicates all the model's interesting insights.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

The basic approach to first differences is accomplished in a small number of computationally tedious steps:

1. Construct two (or more) scenarios, each embodied in a different vector of values for the covariates. Let $\mathbf{x}_\alpha$ represent the vector of the independent values for the first scenario and $\mathbf{x}_\omega$ be the second scenario.
2. If the model function is defined as m, the first difference is $m\left(\mathbf{x}_\alpha \mid \hat{\boldsymbol{\beta}}\right) - m\left(\mathbf{x}_\omega \mid \hat{\boldsymbol{\beta}}\right)$. For a logistic regression this is simply $\frac{1}{1+e^{-\mathbf{x}_\alpha^\top \hat{\boldsymbol{\beta}}}} - \frac{1}{1+e^{-\mathbf{x}_\omega^\top \hat{\boldsymbol{\beta}}}}$.
3. A simple table can be arranged showing the first difference for each variable, under two different conditions.

The method of first differences is almost always a better way of presenting and interpreting logistic regression results than trying to calculate odds, odds ratios, or logged odds ratios. The method requires the construction of alternative scenarios, not just unit changes in independent variables. For example, compare a 30-year-old woman with three young children and no college education with a 50-year-old college-educated woman with no children. Who is more likely to be in the labor force? These kinds of policy scenarios are well suited to the difference method for interpretation of logistic regression.

The method of first differences has one large disadvantage: it returns single numbers – *point estimates* – for each comparison, with no accompanying estimate of our uncertainty around this estimate. Based on simple point estimates, the implications of a good model may play out the same as those from a bad one. For instance, what should we conclude if the standard errors in the example above were all 10? Should that not be reflected in our interpretation of the results?

3.4.4 Estimation Uncertainty around the MLE

In Chapter 2 we showed that the MLE is (asymptotically) consistent and normally distributed. As a practical matter this means that $\hat{\boldsymbol{\theta}} \dot{\sim} \mathcal{N}\left(\hat{\boldsymbol{\theta}}, -H(\hat{\boldsymbol{\theta}})^{-1}\right)$, that is the covariance of the MLE is the negative of the inverse of the Hessian, evaluated at the MLE. The standard error is the square root of the diagonal of the covariance matrix. Importantly, the standard errors alone do not describe how the various parameters are correlated.

It is easy to extract, look at, and use the variance-covariance matrices. Table 3.7 reports just such a matrix for the logit model of labor force participation reported in Table 3.5. It is straight forward to confirm that the square root of the diagonal elements of this matrix are identical to the standard error estimates reported in Table 3.5.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.7 *The matrix $cov(\hat{\beta})$ from the Mroz estimation in Table 3.5.*

	Intercept	**Young Kids**	**School Kids**	**Age**	**College**	**Wage**
intercept	$\sigma^2_{\hat{\beta}_0}$ =0.388	−0.061	−0.023	−0.007	0.001	−0.020
young kids	−0.061	$\sigma^2_{\hat{\beta}_1}$ =0.037	0.002	0.001	−0.005	−0.001
school kids	−0.023	0.002	$\sigma^2_{\hat{\beta}_2}$ =0.004	0.000	0.000	0.001
age	−0.007	0.001	0.000	$\sigma^2_{\hat{\beta}_3}$ =0.000	−0.000	−0.000
college	0.001	−0.005	0.000	−0.000	$\sigma^2_{\hat{\beta}_4}$ =0.038	−0.007
wage	−0.020	−0.001	0.001	−0.000	−0.007	$\sigma^2_{\hat{\beta}_5}$ =0.022

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

3.4.5 Graphical Presentation of Effects in Logistic Regression

Rather than just having point predictions, we would like to interpret models in light of our uncertainty. Similarly, we would like to have an understanding of a good part of the response surface, not just at one or two values. The first differences approach can be informative, though it typically ignores information about the estimation uncertainty, generally contained in the variance-covariance matrix.

Fortunately, with modern computing power combined with likelihood theory we can directly simulate the quantities of interest, incorporating our uncertainty rather than just relying on standard errors of coefficients. Essentially, we simulate *expected values* and *predicted values* by drawing parameter vectors from their asymptotic distribution:

1. Estimate the model by maximizing the likelihood function, storing the coefficient point estimates and the variance-covariance matrix. Let the former be denoted $\hat{\boldsymbol{\beta}}$ and the latter $\text{cov}(\hat{\boldsymbol{\beta}})$.
2. Create a set of values for the independent variables that represents a scenario of analytic or substantive interest. For a single scenario, this might consist of a single value for each of the independent variables. Often, all but a single independent variable of interest are set to their central tendencies. The special variable is set to some specified value. Denote this vector of independent variables representing a scenario of interest $\mathbf{x}_\alpha$.
3. Draw a new vector of parameter values from the multivariate $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \text{cov}(\hat{\boldsymbol{\beta}})\right)$. Denote this vector of k elements $\tilde{\boldsymbol{\beta}}$. This draw uses the covariance (off-diagonal) of the different parameters, whereas conventional BUTON reporting only provides the (square root of) diagonal elements.
4. Given $\tilde{\boldsymbol{\beta}}$ and $\mathbf{x}_\alpha$, calculate $\tilde{\theta} = \text{logit}^{-1}\left(\mathbf{x}_\alpha^\top \tilde{\boldsymbol{\beta}}\right)$. We now have one draw of the *expected value* of the outcome variable under the scenario described in $\mathbf{x}_\alpha$, namely $\text{E}\left[Y \mid \mathbf{x}_\alpha, \tilde{\beta}\right]$. This value incorporates one realization of the estimation uncertainty.
5. To calculate *predicted values*, use the expected values just calculated to simulate the outcome variable $\tilde{Y}$ by a random draw from the stochastic component of the model: $f_B(\tilde{\theta})$.
6. *Important:* Don't do these steps just once; do them hundreds of times. Suppose we draw c different coefficient vectors and construct several different scenarios, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_s$. Let $\tilde{B}_{c \times k}$ be the matrix of c draws from $\mathcal{N}\left(\hat{\boldsymbol{\beta}}, \text{cov}(\hat{\boldsymbol{\beta}})\right)$, and let $\mathbf{X}_{k \times s}$ be the matrix containing the s scenarios. $\tilde{B}\mathbf{X}$ will be the $c \times s$ matrix, with each column representing c different draws of the linear predictor value for a particular scenario.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

We can use these columns to summarize both the central tendency (mean) and uncertainty (standard deviation or quantile interval) surrounding the model's predictions.

Using this approach, we can construct and display many scenarios simultaneously, leading to dense and informative interpretations of the model. For example, we may wish to look at the predicted "effect" of the wage rate on the probability of being in the labor force, comparing women with no young children to those with one young child over the range of observed wage rates. Figure 3.3 illustrates exactly this comparison. We display the 95% confidence bands around the predicted employment probability for each wage rate, holding all other covariates at central tendencies and varying the number of young children in the home. Code for generating this figure follows in Example 3.3.

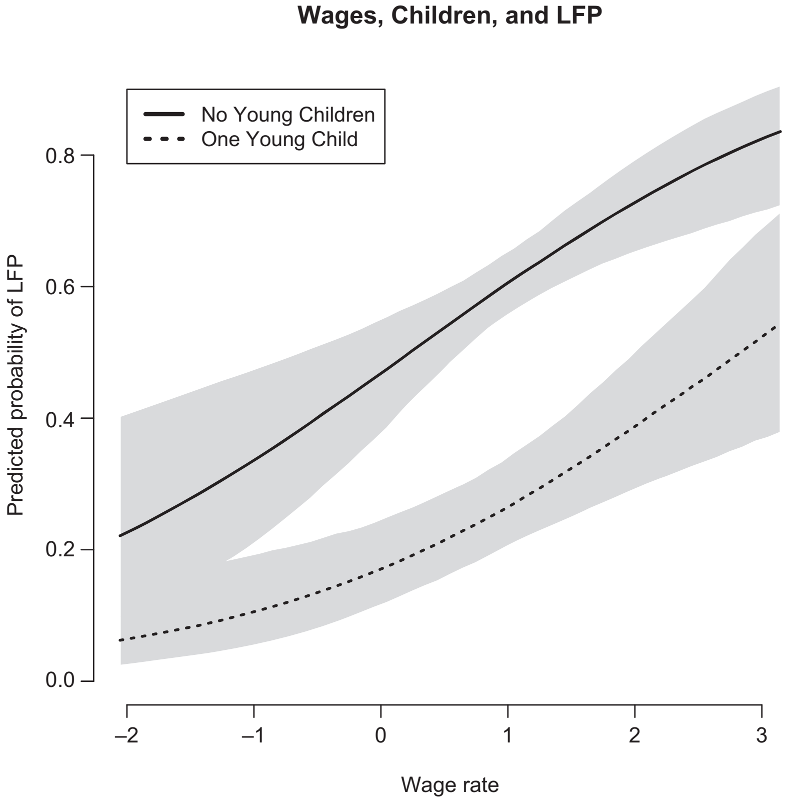

FIGURE 3.3 Plot displaying the 95% confidence bands for the predicted probability of LFP across different wage rates for women with and without young children. The estimated relationship between wages and employment probability differs between the two groups of women.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

3.5 AN EXAMPLE: US SENATE VOTING

In October 2009, Tom Coburn – a Republican US senator from Oklahoma – introduced an amendment to House Resolution 2847 that would strip all funding for political science research from the National Science Foundation (NSF) budget. The amendment was defeated in a roll call vote on November 2, 2009. Uscinski and Klofstad (2010) used this vote as a way of probing the impact of senators' individual, constituency, and institutional characteristics on their vote. They treated these Senate votes as independent binary data; their tool of choice was logistic regression.

Votes are not exactly binary, as some senators will abstain, but on the Coburn Amendment there were 98 recorded votes out of 100 senators. Thirty-six senators voted in favor of the amendment, while 64 voted against 2. As is typical, votes in favor are typically coded as a numeric 1 and votes against as 0. We fit two models. In the first we include only a single predictor: the senators' partisan affiliations (Democrat or not). In the second we include all of Uscinski and Klofstad's covariates. They grouped these variables into three categories: individual, constituency, and institutional. In terms of individual characteristics, two variables are pertinent: the gender of the senator and whether they majored in political science in college. Constituency effects basically focus on whether the senator's state has a large number of prominent doctoral programs in political science, gets substantial awards from the National Science Foundation, and has voters who have petitioned the subcommittee concerning the amendment. Finally, Uscinski and Klofstad measure how long until each senator's next election and whether the senator was a member of the Subcommittee on Labor, Health, and Human Services.

Maximum likelihood estimation of this model produces the standard BUTON, given as Table 3.8. The results indicate that partisanship was a strong predictor of senators' votes on the Coburn Amendment. Comparing the two models in terms of the differences in the log-likelihood, we see that including many more covariates does indeed improve the model fit (higher likelihood); the likelihood ratio is 24 with a χ^2 value of 0.01. But the AIC improvement for the full model relative to the simpler one is modest, reflecting the fact that the full model includes eleven more parameters. Using the BIC, with its heavier penalty for complexity, we see a preference for the simpler model.

An alternative form of presentation is the coefficient plot. Figure 3.4 displays such a plot for the second model and omitting the intercept. In this display it is easy to see which coefficients are near special values, such as zero. The points represent coefficient point estimates and the lines are the ±2 standard error (approximate 95%) interval around the estimated coefficient. We find this a more satisfying way of displaying regression output than the standard BUTON with asterisks, but complications can ensue when variables are on wildly different scales.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

$\mathcal{R}$ Code Example 3.3 *Post Estimation Analysis*

```r
#scenarios: going from min to max of female wages
#comparing women with one young child to one with none.
#all other covariates at sample median/mode
library(MASS)
inv.logit<-function(x){
  1/(1+exp(-x))
}

#the wage values for the scenarios
lwg.range <- seq(from=min(lfp$wage), to=max(lfp$wage),by=.1)

#women w/o young kids
x.lo <- c(1, #intercept
          0, #young.kids
          median(lfp$school.kids),
          median(lfp$age),
          0, #college
          median(lfp$wage))
X.lo <- matrix(x.lo, nrow=length(x.lo), ncol=length(lwg.range))
X.lo[6,] <- lwg.range #replacing with different wage values

#women with young kid(s)
x.hi <- c(1, #intercept
          1, #young.kids
          median(lfp$school.kids),
          median(lfp$age),
          0, #college
          median(lfp$wage))
X.hi <- matrix(x.hi, nrow=length(x.hi), ncol=length(lwg.range))
X.hi[6,] <- lwg.range #replacing with different wage values

B.tilde <- mvrnorm(1000, coef(fit.glm), vcov(fit.glm)) #1000 draws of
     coefficient vectors
s.lo <- inv.logit(B.tilde %*% X.lo) #matrix of predicted probabilities
s.hi <- inv.logit(B.tilde %*% X.hi) #matrix of predicted probabilities
s.lo<-apply(s.lo, 2, quantile, c(0.025, 0.5, .975)) #95\% CI and median
s.hi<-apply(s.hi, 2, quantile, c(0.025, 0.5, .975)) #95\% CI and median

#plotting the results
plot(lwg.range, s.lo[2,], ylim=c(0,.9), xlab = "wage index",
    ylab = "Predicted Probability of LFP",
    main = "Wages, Children, and LFP", bty="n",
    col="white")
polygon(x=c(lwg.range, rev(lwg.range)), #confidence region
    y=c(s.lo[1,], rev(s.lo[3,])),
    col=grey(0.8), border=NA)
polygon(x=c(lwg.range, rev(lwg.range)), #confidence region
    y=c(s.hi[1,], rev(s.hi[3,])),
    col=grey(0.8), border=NA)
lines(lwg.range, s.hi[2,], lty=3, lwd=2)
lines(lwg.range, s.lo[2,], lwd=2)
legend(-2, 0.9, legend = c("No Young Children",
    "One Young Child "),lty = c(1,3),lwd=3)
```

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.8 *Replication of Table 1 in Uscinski and Klofstand (2010).*

	Dependent Variable: Vote "Nay" on Coburn Amendment	
	(Simple Model)	**(Full Model)**
Democrat	3.120	3.300
	(0.575)	(0.853)
Gender (Female)		−0.408
		(0.977)
Political Science Major in College		1.160
		(0.907)
Number of Top 20 Political Science Programs		2.080
		(0.996)
Number of Top 50 Political Science Programs		1.020
		(0.770)
Total Number of Political Science Programs		−0.183
		(0.433)
Percentage with Advanced Degrees		2.350
		(1.250)
Number of Amendment Petitioners		−0.011
		(0.016)
Number of NSF Grants 2008		0.157
		(0.345)
Years to Next Election		0.486
		(0.229)
Member of Labor HHS Subcommittee		1.390
		(0.968)
Seniority		−0.0002
		(0.041)
Constant	−0.802	−3.780
	(0.334)	(1.600)
n	98	98
$\log \mathcal{L}$	−42.800	−30.600
AIC	89.700	87.300
BIC	94.800	121.000

3.5.1 Model Evaluation

Logistic regression assigns observations probabilistically to one of two classes. There are a variety of diagnostics which help to assess the performance of such classifiers beyond the standard log-likelihood based measures. We take the view that, in general, single number summaries are, individually, inadequate. But it is useful to understand them.

In weather prediction, where forecasts are generally based on probabilities (30% chance of rain tomorrow, for example), the *Brier score* is often employed.

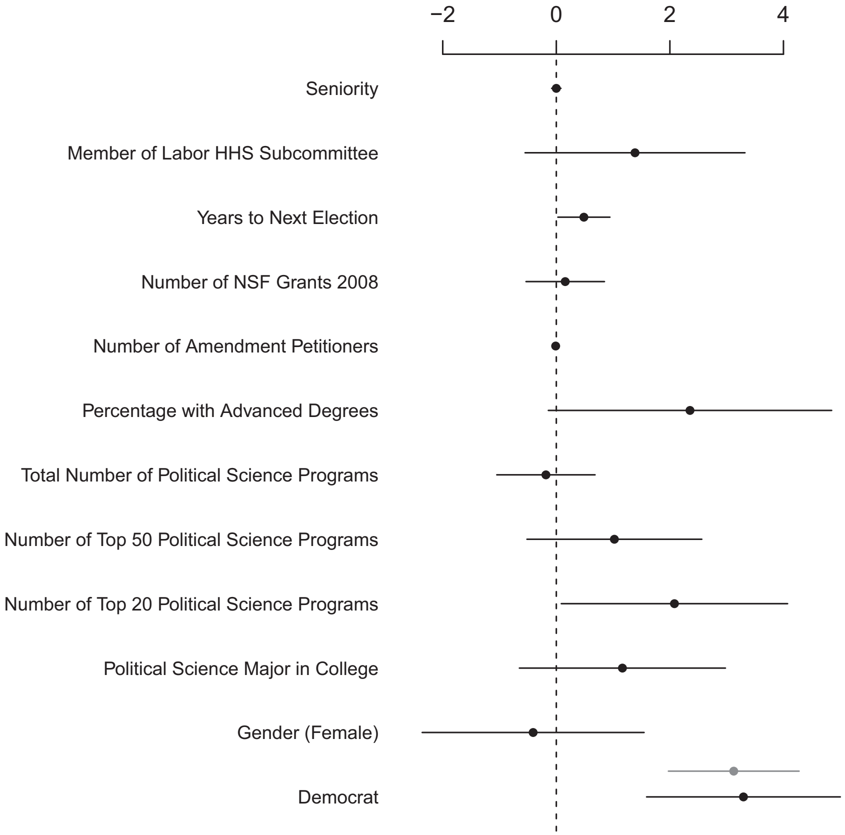

FIGURE 3.4 A coefficient plot of the logit regression of US Senate votes on the Coburn Amendment to eliminate NSF funding for political science. Horizontal bars are 95% confidence intervals. The lighter gray point and error bar represents estimates from the simple model.

Definition 3.1 (Brier Score). The Brier score (Brier, 1950) for a binary classifier such as logistic regression is defined as

$$B_b \equiv \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_i - y_i\right)^2,$$

where the predicted probabilities are $\hat{\theta}_i$ and the observed binary outcomes are given as y_i.

Lower values reflect better predictions. The Brier score for the logit model containing only partisanship is 0.14 against 0.09 for the more complicated model.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

Another approach is to generate a 2 × 2 table of predicted values against actual values, which will be distributed χ^2 with 1 degree of freedom. To do this, however, we must specify a specific threshold, t, that will map predicted probabilities into failures (0) and successes (1). Where $\hat{\theta}_i > t$ the model predicts 1s, otherwise the outcome is coded as a predicted 0. Often, and perhaps by default, analysts use $t = 0.5$ as a cutoff. Applying that same value to the predicted probabilities from the full model in the Coburn example, we recover the results in Table 3.9. There are 15 incorrect predictions; the χ^2 is 40, which is highly unlikely for a 2 × 2 table with 98 observations. But this also begs the question, "What is special about 0.5?," especially when the raw probability of voting "nay" is 0.65? Why focus on just one threshold?

ROC Curves

The *Receiver Operating Characteristic (ROC) Curve* (don't ask about the name) builds on the idea of comparing correct predictions against false positives. But rather than choosing one particular threshold, ROC curves display this comparison for all thresholds between 0 and 1. ROC curves are based on the idea that the relative costs of mis-predicting a failure (false negative) versus mis-predicting a success (false positive) can vary depending on the problem at had. More formally, let the cost of a false positive relative to the cost a false negative be denoted C. The optimal prediction for an event ($\hat{y} = 1$) that minimizes the total expected cost occurs when $t > 1/(1 + C)$; otherwise, $\hat{y} = 0$. Hence, if the false positives and false negatives are equally costly, then $C = 1$. If, however, the cost of mis-predicting an event is twice as costly as mis-predicting the absence of an event, then $C = 2$ and the cutoff would be at $t = 1/3$. The appropriate value for C in any particular application is, of course, a policy problem. The threshold should be established in terms of the human and physical costs of mis-predicting say, the absence of war, versus mis-predicting a war. The ROC curve is a way of summarizing the ratio of the rate of false positives to the rate of false negatives over the entire range of t.

ROC curves plot the true positive rate (percent of actual successes correctly predicted, for some fixed threshold) against the false positive rate (percent of

TABLE 3.9 *Predicted versus observed votes for the Coburn Amendment, with $t = 0.5$ for mapping probabilities into event space. Calculations are based on the second model in Table 3.8.*

	Observed 0	**Observed 1**
Predicted 0	27	8
Predicted 1	7	56

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

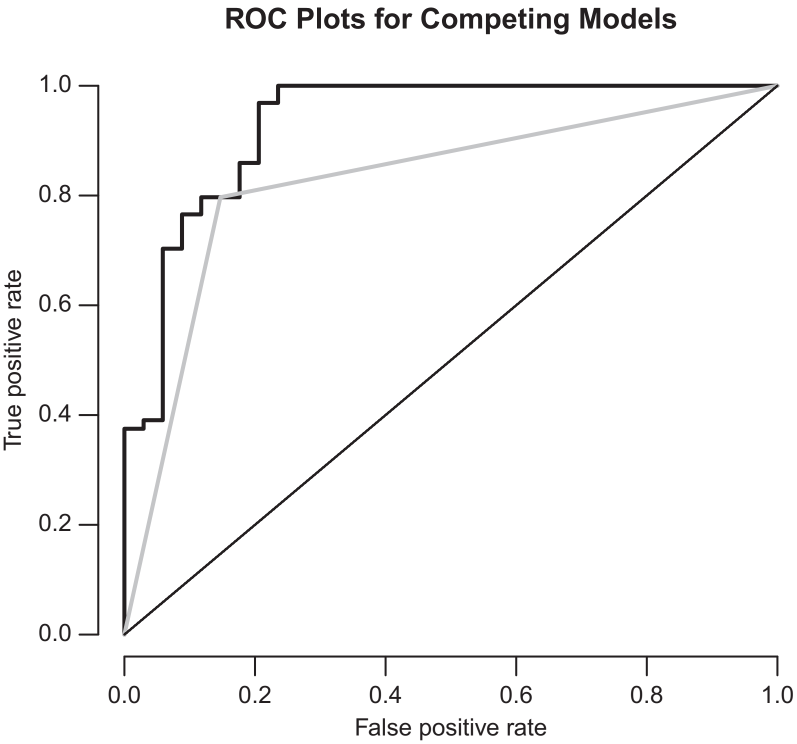

FIGURE 3.5 Receiver Operator Characteristics (ROC) curve for the logit models of the vote on the Coburn Amendment. The gray line is generated by the partisanship-only model (AUC= 0.83). The black line is generated by the full model (AUC= 0.93).

false positives out of actual failures).[1] The threshold values, t, that generates each particular point in the curve are not visible in the plot. Better models will have low false positive rates when they also have high true positive rates, whereas worse-performing models will have ROC curves that are close to the diagonal.

Figure 3.5 illustrates the ROC curves generated by both models in Table 3.8. The simpler model (the gray line) has one step, corresponding to the fact that the underlying model has only one (binary, categorical) predictor. ROC curves from competing models can be used heuristically to compare the fit of various model specifications, though this is a bit of an art. In this example the more complicated model, as reported in Uscinski and Klofstand's original paper, is better at predicting in-sample for all values of t.

[1] The false positive rate is also 1 – specificity, where specificity is the percent of failures correctly predicted as such.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

Some prefer the "area under the ROC curve" (AUC) as a single-number summary of model performance as derived from the ROC; we report these values in the caption. These single number summaries are unfortunate since they lend a false sense of precision and certainty in comparing model performance while discarding the fact that model performance may differ across t. Two models could conceivably provide very different trade-offs between false positives and false negatives relevant to decision makers. This would be masked by simple comparison of AUCs.

Separation Plots

Many models in the social sciences will not have any predicted probabilities that are greater than 0.5. This is not necessarily a symptom of a poorly fitting model. Rather, it is driven, in part, by the underlying frequency of events. When the observed number of events is small relative to the number of trials, as is the case with international conflict, we should expect small absolute predicted probabilities. What really concerns us is the model's ability to distinguish more likely events from less likely ones, even if they all have small absolute probabilities.

To visually compare different models' abilities to usefully discriminate between cases, we can sort the observations by their predicted probabilities and then compare this sorting to actual, observed events. This is exactly the strategy followed by Greenhill et al. (2011) in developing their *separation plot*. In these plots, dark vertical bars are observed events, in this case nay votes. Light bars are nonevents. If the model perfectly discriminated between events and nonevents, then successes would cluster to the right of the plot and failures to the left; the plot would appear as two starkly defined color blocks. A model that performs poorly would appear as a set of randomly distributed vertical lines.

Figure 3.6 displays separation plots for each of the models in Table 3.8. The predicted probabilities are ranked from low to high (shown as a black line), a red bar indicates a vote against the amendment, and a cream colored bar is a vote in favor. Consistent with the results from the ROC plot, we see that the full model is better at discriminating nay votes from the other senators.

Model Interpretation

There is some evidence that the more complicated full model of Uscinski and Klofstand is better at predicting Senate votes on the Coburn Amendment. How can we interpret these findings? One way is to use the exponentiated logit coefficients, as presented in Table 3.10. These values represent the odds ratios for different values of the covariate. This makes some sense in the context of partisanship; the odds of a Democrat voting nay are 27 times greater than a Republican. But it is harder to interpret what these values say about continuous predictors. Odds ratios are a tough sell.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

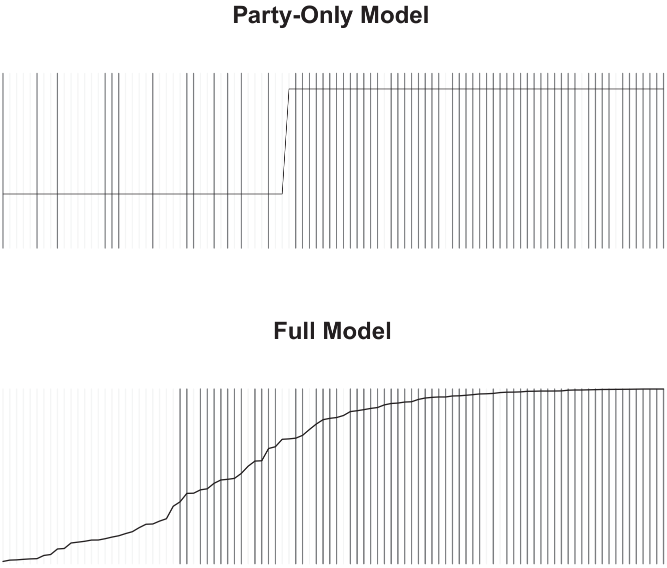

FIGURE 3.6 Separation plots for the partisan-only and full models of US Senate voting on the Coburn Amendment.

Using our strategy of constructing scenarios of interest and then simulating from the sampling distribution we can make more nuanced and easily expressed interpretations of the model. These interpretations have the added benefit of incorporating our uncertainty while reflecting the nonlinearity inherent in both the model and in categorical data. Here we are interested in how the predicted probability of nay vote changes as senators approach reelection for Democrats and non-Democrats.[2] To generate this scenario, we hold the other covariates at their mean or modal values and then vary the time-until-election from one to five years. We display these results in Figure 3.7.

US Congressional elections are only held in even-numbered years, so we only observe three values for the number of years-to-election: one, three, and five. This is reflected in the plot. The figure indicates three implications of the model. First, the gap in voting behavior is driven, unsurprisingly, by partisanship. Democrats are almost certain to vote “nay,” regardless of their proximity to reelection. Their predicted probability is nearly at the top of the plot. Second, the gap between Democrats and non-Democrats narrows modestly as reelection is further in the future. Third, the uncertainty around the predicted probability for non-Democrats is considerable. The predicted probability at five years until reelection is about 0.9, well within the confidence bounds

[2] There were two independent senators at the time, Joe Lieberman of Connecticut and Bernie Sanders of Vermont, both of whom voted nay.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.10 *Odds Ratios for the replication of table 1 in Uscinski and Klofstand (2010).*

	OR	2.5 %	97.5 %
Democrat	27.00	6.08	190.48
Gender (Female)	0.66	0.10	4.89
Political Science Major in College	3.20	0.57	21.44
Number of Top 20 Political Science Programs	8.01	1.42	89.54
Number of Top 50 Political Science Programs	2.78	0.65	14.47
Total Number of Political Science Programs	0.83	0.34	1.92
Percentage with Advanced Degrees	10.53	1.21	181.70
Number of Amendment Petitioners	0.99	0.96	1.02
Number of NSF Grants, 2008	1.17	0.57	2.30
Years before Next Election	1.63	1.06	2.66
Member of Labor HHS Subcommittee	4.00	0.65	31.22
Seniority	1.00	0.92	1.08

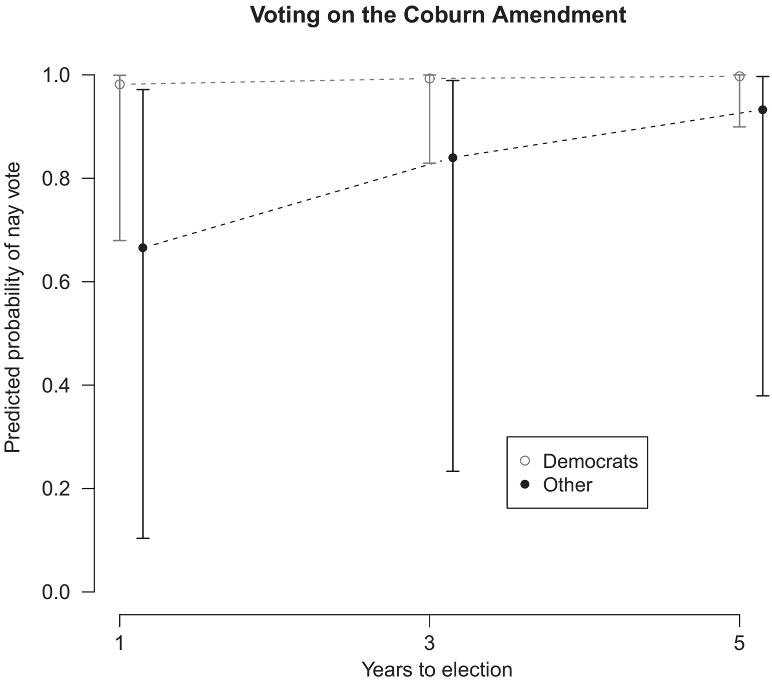

FIGURE 3.7 The predicted probability of voting “nay” on the Coburn Amendment as reelection approaches for Democrat and non-Democrat US senators. All other covariates are held at their central tendencies. The vertical bars are 95% confidence bands.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

for both a Democrat in that year as well as for the "typical" non-Democrat who is facing an election within the year. In other words, the "effect" of proximity to an election appears weak and only visible among non-Democrats. This insight would not be discernible simply by examining BUTON or odds ratios.

3.6 SOME EXTENSIONS

3.6.1 Latent Variable Formulation for Binary Data

Imagine that the phenomenon we seek to observe is actually continuous but, unfortunately, we are only able to observe it imperfectly. We observe the underlying *latent* variable when it passes some threshold. And even then we do not observe actual value but only the fact that it crossed the observability point. While we do not need such a specification for deriving the logit model, it is conceptually useful and underpins several more complicated models, so it bears discussion here.

Consider a discrete, binary outcome Y_i which is based on an unobserved latent variable Z_i as:

$$\begin{aligned}
Y_i &= \begin{cases} 1 & \iff \quad z_i > \tau \\ 0 & \iff \quad z_i \leq \tau \end{cases} \\
Z_i &= \mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i \\
\varepsilon_i &\sim f,
\end{aligned}$$

where τ is a threshold above which we observe an event, and f is some probability distribution. Since this threshold is unobserved and, in some sense, arbitrary, we set $\tau = 0$.[3]

Making different assumptions about the distribution of the error term, f, yields different models. If we assume the errors are distributed via a logistic probability distribution with mean zero and scale parameter fixed at 1, we arrive back at the logit model we derived above. If we instead assume that $\varepsilon_i \sim \mathcal{N}(0, 1)$ and denoting $\Phi(\cdot)$ as the corresponding cumulative distribution function, we arrive at the *probit* model:

$$\begin{aligned}
\Pr(Y_i = 1) &= \Pr(z_i > 0) \\
&= \Pr\left(\mathbf{x}_i^\top \boldsymbol{\beta} + \varepsilon_i > 0\right) \\
&= \Pr\left(\varepsilon_i > -\mathbf{x}_i^\top \boldsymbol{\beta}\right) \\
&= 1 - \Phi\left(-\mathbf{x}_i^\top \boldsymbol{\beta}\right) \\
&= \Phi\left(\mathbf{x}_i^\top \boldsymbol{\beta}\right),
\end{aligned}$$

[3] We could set τ to any other value and then simply subtract that constant from the linear predictor to yield an equivalent model. In other words, τ is not identified.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

where the last step follows from the symmetry of the normal distribution, i.e., $\Phi(x) = 1 - \Phi(-x)$. Predicted probabilities and statistical inference under a probit specification are nearly identical to those from a logit. The coefficient values will differ; it turns out that logit coefficients divided by 1.6 should give a close approximation to the estimated coefficients from a probit regression. Our approach for generating meaningful interpretations of model estimates and implications fits the probit case just as easily.

In case you were wondering ... 3.1 Logistic distribution

Let Y be a random variable with support along the entire real line. Suppose that the cumulative distribution function (CDF) for Y can be written as

$$\Pr\left(Y \leq y\right) = \Lambda(y;\mu,\sigma) = \frac{1}{1+\exp\left(-\frac{y-\mu}{\sigma}\right)}. \tag{3.1}$$

We say that Y follows a *logistic distribution* with parameter vector $\boldsymbol{\theta} = (\mu, \sigma)$. We write $Y \sim f_L(y;\mu,\sigma)$. $\text{E}[Y] = \mu$ and $\text{var}(Y) = \sigma^2\pi^2/3$.

There is rarely, if ever, a statistical reason for preferring a logit model to a probit or vice versa. Some disciplines prefer the probit, others use the logit more frequently. Sometimes there are aesthetic reasons to prefer one to the other. For example, the *strategic probit* model (Signorino, 1999) emerges from an extensive form game where the stochastic components are normal distributions. Occasionally there are computational reasons to prefer one specification over another but, in general, the choice between logit and probit is inconsequential for the purposes of inference and prediction.

3.6.2 Heteroskedastic Probit

The standard probit model assumes that the error distribution has fixed, unit variance. But it may be the case that the error variance will depend on variables that are included in the model. For example, in survey data different subgroups might differ in the levels of information they have about a topic, so the variance around their group-specific average (latent) response may also differ. Davidson and MacKinnon (1984) developed an extension to the probit model that can account for this variance relationship. This model goes by the name *heteroskedastic probit*.

The heteroskedastic probit model relies on the fact that we can transform a $\mathcal{N}(0,\sigma^2)$ variable into a $\mathcal{N}(0,1)$ variable by dividing the standard deviation, σ. So we can by rewrite the standard probit model as

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

$$\begin{aligned}
\Pr(Y_i = 1) &= \Pr\left(\frac{\varepsilon_i}{\sigma} > -\frac{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\right) \\
&= \Phi\left(\frac{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\sigma}\right).
\end{aligned}$$

To complete the model we must specify a relationship between covariates and σ_i. Since the standard deviation is nonnegative the exponential function is useful here:

$$\sigma_i = \exp(\mathbf{v}_i^{\mathsf{T}}\boldsymbol{\gamma}),$$

where $\mathbf{v}$ is the vector of covariates thought to be linked to the variance, and γ is a vector of the to-be-estimated parameters. Given $\mathbf{X}$ and $\mathbf{V}$ we can now state the log-likelihood:

$$\begin{aligned}
&\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathrm{X}, \mathrm{V}, \mathrm{y}) \\
&= \sum_{i=1}^{n} \left[y_i \log \Phi\left(\frac{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\exp(\mathbf{v}_i^{\mathsf{T}}\boldsymbol{\gamma})}\right) + (1 - y_i) \log \left[1 - \Phi\left(\frac{\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}}{\exp(\mathbf{v}_i^{\mathsf{T}}\boldsymbol{\gamma})}\right)\right]\right] \\
&= \sum_{i=1}^{n} \left[y_i \log \Phi\left(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} \exp(-\mathbf{v}_i^{\mathsf{T}}\boldsymbol{\gamma})\right) + (1 - y_i) \log \Phi\left(-\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta} \exp(-\mathbf{v}_i^{\mathsf{T}}\boldsymbol{\gamma})\right)\right].
\end{aligned}$$

3.6.3 Complimentary Logistic Model, a.k.a. *cloglog*

Another complication arises when we consider the fact that both the logistic and normal distributions are symmetric. This symmetry carries over to the logit and probit transformations: $\text{logit}(\theta) = -\text{logit}(1 - \theta)$ and $\Phi^{-1}(\theta) = -\Phi^{-1}(1 - \theta)$. This implies that as a predictor, x, becomes large $\theta(x)$ approaches 1 at the same rate that it approaches 0 as x becomes small.

In some circumstances we might worry that the process we are examining is not symmetric in this way. This can occur, for example, when the probability of an event is extremely small or very large, even with good variation in the predictor variable. A different choice of function mapping $\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ into $[0, 1]$ can allow for asymmetry.

The complimentary logistic model, often called the *complementary log-log* or *cloglog*, is one possibility. The model is still based on Bernoulli trials but now $\log(-\log(1 - \theta_i)) = \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}$ or, equivalently, $\theta_i = 1 - \exp(-\exp(\mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}))$. As displayed in Figure 3.8, we can see that the predicted probabilities approach 1 much faster than 0 under the cloglog specification.

An immediate difference between the cloglog model and logit/probit is that the model is not symmetric with respect to the coding of the outcome variable. In the Coburn Amendment example, swapping the coding of the dependent

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

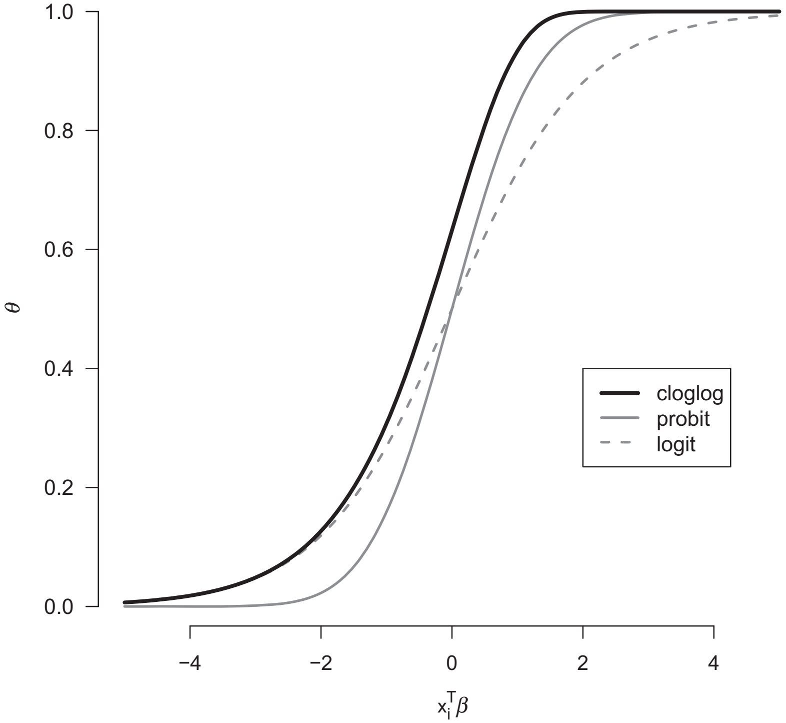

FIGURE 3.8 The (inverse) logit, probit, and complementary log-log functions, mapping the linear term into θ. Logit and probit are symmetric around $\theta = 0.5$, whereas the cloglog approaches 1 faster and 0 more slowly.

variable such that "nays" are 0 would have no effect on the estimates reported in Table 3.8 beyond flipping their signs. Nothing else, including model fit, would change. In a cloglog model, however, this symmetry does not hold.

3.6.4 An Example: Civil Conflict

In new democracies, prior autocratic leaders with military ties often retain a strong political presence in the new regime. Such legacies may have consequences for subsequent political stability. Cook and Savun (2016) examined how the structure of prior authoritarian government relates to a binary coding of civil conflict in new democracies, 1946–2009. Their primary analysis relied on logistic regression.

The Cook and Savun data contain a heterogeneous group of countries and a comparatively small number of conflicts (69) relative to the number of country-years, presenting an opportunity to apply the modeling extensions just described. Table 3.11 is the BUTON reflecting this exercise. In the first column

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

TABLE 3.11 *Replication and extension of table II in Cook and Savun (2016).*

	Dependent Variable:				
	Civil Conflict Onset				
	logit (1)	**probit** (2)	**het. probit** (3)	**cloglog** (4)	**log-log** (5)
Military	0.96	0.45	0.17	0.87	−0.32
	(0.28)	(0.14)	(0.07)	(0.27)	(0.11)
Personal	0.15	0.09	0.07	0.14	−0.09
	(0.65)	(0.32)	(0.10)	(0.62)	(0.24)
Party	−0.45	−0.22	−0.05	−0.40	0.14
	(0.67)	(0.30)	(0.10)	(0.65)	(0.21)
GDP (lagged)	−0.30	−0.15	−0.04	−0.27	0.11
	(0.17)	(0.08)	(0.03)	(0.17)	(0.05)
Population (lagged)	0.53	0.24	0.13	0.51	−0.16
	(0.12)	(0.06)	(0.03)	(0.10)	(0.05)
Peace years	−0.45	−0.23	−0.06	−0.42	0.17
	(0.18)	(0.09)	(0.04)	(0.16)	(0.07)
n	2575	2575	2575	2575	2575
$\log \mathcal{L}$	−239	−240	−236	−238	−242
AIC	498	500	495	496	503
BIC	557	558	559	555	562

Note: Following Cook and Savun, standard errors are clustered by country and cubic splines estimated but omitted from the table. Lagged population is the only variance-term covariate for the heteroskedastic probit model.

we repeat their logit analysis, while the second column fits the same model as a probit. In the third column we fit a heteroskedastic probit using population as the only covariate in the model for σ. In the last two columns we fit two cloglog models. The first of these continues the coding, assigning conflict onset a 1 and nonconflict a 0, whereas in the "log-log" model we reverse the coding of the dependent variable.

Unsurprisingly, different distributional assumptions produce different numerical estimates for the regression parameters. But the raw values reported in the table turned out to yield very similar descriptions of the DGP, as indicated by the log-likelihood, AIC, and BIC values. This is further confirmed in Figure 3.9, which shows that model fit is virtually identical across these alternatives. The ratio of coefficient estimates to their standard errors are also similar across models. Note, however, that the parameter estimates for the cloglog and log-log models are not opposites, reflecting the asymmetry in the assumed distribution. In this example several model variations produce effectively the same answer.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

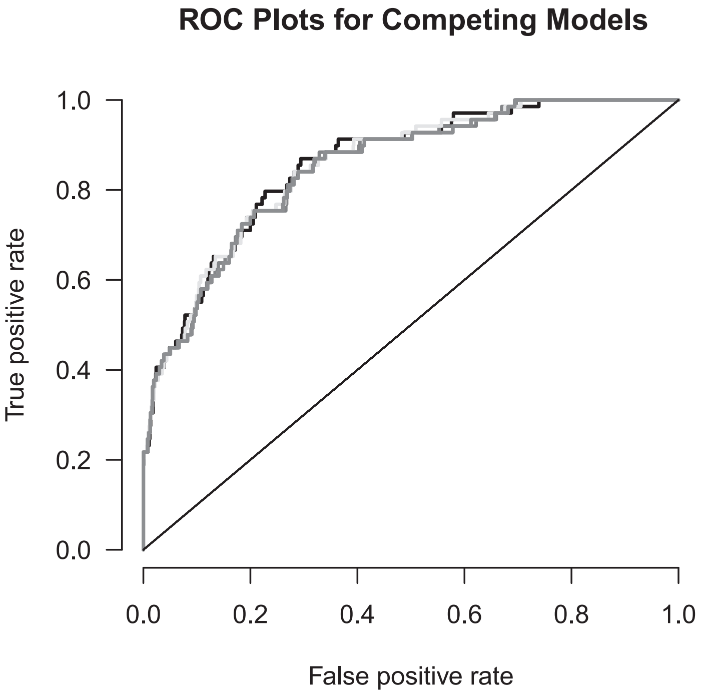

FIGURE 3.9 ROC curves for the five models reported in Table 3.11. The models all fit the data similarly.

3.7 SUMMARY OF GENERAL APPROACH TO BINARY VARIABLES

When outcomes are binary, analysis tools that assume a normally distributed response are often inappropriate. If we imagine that our outcome can be modeled as a Bernoulli trial, combined with a linear predictor term that can be mapped into the probability space, then we can derive a variety of models for binary data. Given a link function and a probability distribution, these models are easily estimated using maximum likelihood. Logit and probit are the most commonly used and can be considered interchangeable.

Interpretation of nonlinear models is more involved than for OLS since marginal effects (and the uncertainty around them) depend on the values of the covariates. Standard tables of coefficients and standard errors are not terribly informative about the model's implications. We advocate for a general procedure in which the analyst constructs substantively meaningful and defensible scenarios (i.e., vectors of covariate values) and then uses a large number of samples from the limiting distribution to generate many values of particular quantities of interest. In this way, the analyst can both

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

make substantive interpretations at meaningful values while communicating her uncertainty around those values.

3.8 FURTHER READING

Applications

Logit and probit models are widely employed across the social sciences in both observational and experimental work. Among many recent examples, Einstein and Glick (2017) look at how a constituent's race affects whether and how public housing bureaucrats respond to information requests. Baturo (2017) examines the predictors of political leaders' activities after leaving office.

Past Work

Joseph Berkson was a statistician with the Mayo Clinic and most well-known for his pioneering studies of the link between tobacco smoking and lung cancer in the 1960s. Berkson's work on the logistic regression – as an explicit correction to using a normal distribution for studying probabilities – began in 1944 (Berkson, 1944). Subsequently, Berkson produced several important works relating to this topic (Berkson, 1946, 1953, 1955). However, work on using sigmoid curves (akin to logistic curves) dates back to the work of Bliss (1935), who worked with Fisher at the Galton Laboratory, University College, London; R. A. Fisher followed Karl Pearson as the Galton Professor at the UCL in 1934.

Glasgow and Alvarez (2008) provide a recent summary of likelihood-based models for discrete choice, along with their extensions. Alvarez and Brehm (1995, 2002) is an early derivation and application of the heteroskedastic probit in the study of American public opinion. Nagler (1994) proposes the *scobit* model that allows for a logistic distribution that is scaled so as to not require an assumption of symmetry around 0. A challenge with the scobit model is that there is often insufficient information in the covariates to cleanly estimate both the regression weights and the ancillary parameter governing symmetry of the distribution.

Software Notes

We used the `arm` library's `coefplot` command (Gelman and Su, 2016), but the `coefplot` (Lander, 2016) extends this functionality in a number of ways. The `verification` (NCAR – Research Applications Laboratory, 2015) and `scoring` (Merkle and Steyvers, 2013) libraries can be used to calculate the Brier score. The `separationplot` library (Greenhill et al., 2015) calculates

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press

and displays separation plots. glmx (Zeileis et al., 2015) provides a way to estimate probits (and other models) allowing for heteroskedasticity.

The $\mathcal{R}$ package Zelig package (Imai et al., 2008, 2009) and its progeny have implemented a general syntax for estimating several classes of models, including logit, probit, and many other types of models. It facilitates for nonprogrammers the calculation and display of quantities of interest and associated uncertainty under different scenarios.

https://doi.org/10.1017/9781316888544.006 Published online by Cambridge University Press