ECOM40006/ECOM90013 Econometrics 3
Department of Economics
University of Melbourne

Binary Response Models

Semester 1, 2025

Version: May 19, 2025

Contents

1 The Latent and Observable Models

We postulate a latent model

$$y_i^* = x_i'\beta + u_i \qquad i = 1, \ldots, n, \tag{1}$$

where x_i is a vector of observations on k explanatory variables and β is a k-vector of coefficients. We shall assume that $u_i|x_i \sim f(u_i)$. We shall further assume that $f(u_i)$ is symmetric about zero, which implies that $\mathrm{E}[u_i|x_i] = 0$. For later use we define the conditional distribution function of u_i given x_i as

$$F(a) = \int_{-\infty}^{a} f(u_i)\, du_i.$$

In particular, observe that

$$\frac{\mathrm{d}F(a)}{\mathrm{d}a} = f(a),$$

where $f(a)$ denotes the density of u_i given x_i evaluated at the point a. We shall on occasion wish to stack the variables and to this end we define the $n \times k$ matrix $X = [x_1, \ldots, x_n]'$ and the n-vectors $y^* = [y_1^*, \ldots, y_n^*]'$ and $u = [u_1, \ldots, u_n]'$. We shall assume that X can be treated as being predetermined.

The distinguishing feature of (1) is that y_i^* is unobservable. Instead we have the observation rule

$$y_i = \begin{cases} 1, & \text{if } y_i^* > 0 \Rightarrow x_i'\beta + u_i > 0 \Rightarrow u_i > -x_i'\beta, \\ 0, & \text{otherwise.} \end{cases}$$

Note that, in essence, we only observe information about the sign of y_i^* and so it is impossible to estimate its variance, σ^2 say, because such an estimate would require information about the magnitude of y_i^*. In recognition of this fact it is standard practice to normalize the variance of u_i given x_i to unity. Normalization of (1) in this way means that the only estimable parameter is β/σ. This has implications for the interpretation of final results but is such common practice that it typically passes without further comment. Because there is no prospect of unravelling our estimates of β/σ to obtain separate estimates of β and σ^2 we shall henceforth denote the ratio β/σ generically by β.

It should be noted that the way the model has been set up predisposes the reader to the mindset that what we have here is weakness of data leading to a paucity of information. This tends to be the way economists view these things but it should be noted that there are many variables that are inherently binary in nature ; e.g. alive/dead, male/female, loving econometrics/not so much, that a latent model seems artificial. Indeed, statisticians probably wouldn't necessarily think of approaching the problem in this way, taking the observable model as their starting point. There is no right or wrong here, merely an admonition to think about your model and why you are doing things the way that you are.

2 The Linear Probability Model

One possible approach to estimating the βs is to fit a model involving the observable data using OLS, on the grounds that there is nothing in the assumptions that specifically precludes the dependent variable from only taking the values 0 or 1. What is to stop one from estimating β according to $\hat{\beta}_n = (X'X)^{-1}X'y$, where $y = [y_1, \ldots, y_n]'$? The short answer is nothing. However, there are consequences for the interpretation and application of the results so obtained, of which you should be aware. First, the reason that linear models are called *regression models* is because the quantity being estimated is actually a

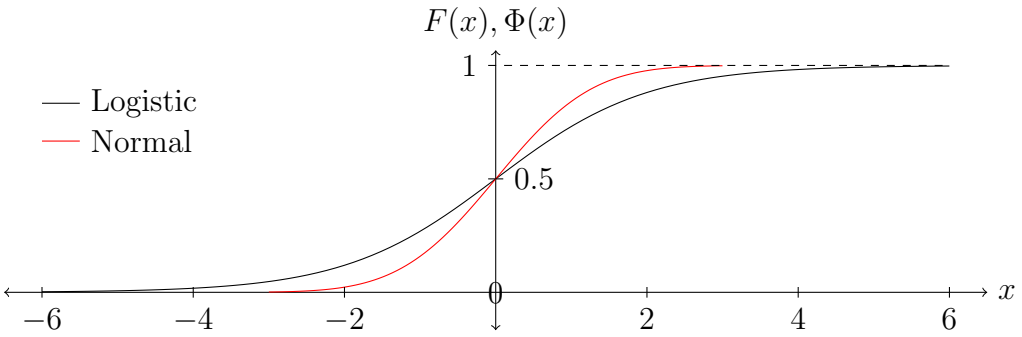

Figure 1: Two Distribution Functions: Logistic and Standard Normal

conditional expectation of y given X. In this circumstance the βs that we are interested in belong to (1) rather than the observable model

$$y = X\beta + \upsilon,$$

where the unobserved disturbance term υ will simply be something different to u. Our fitted model $X\hat{\beta}_n$ is an estimate of a regression function which, by definition is the conditional expectation of y given X. If we look at just a single observation then we observe that, because it can only take values zero or unity, y_i is a Bernoulli random variable which implies that

$$\begin{aligned} \mathrm{E}\left[y_i \middle| x_i\right] &= \Pr\left(y_i = 1 \mid x_i\right) = \Pr\left(y_i^* > 0 \mid x_i\right) \\ &= \Pr\left(u_i > -x_i'\beta \mid x_i\right) = 1 - F\left(-x_i'\beta\right) = F\left(x_i'\beta\right), \end{aligned}$$

where the final equality follows by appealing to the symmetry about zero of $f\left(u_i\right)$. Before moving on, it is worth noting that the argument of the probability function is $x_i'\beta$, the same as for a linear regression model, where we note that the linearity here refers to the that this function is a weighted sum of the elements of β. That is, *linear* here refers to linear in the parameters. Even though the regression function that is our conditional expectation is a non-linear function of its argument, its argument is linear in the elements of β. For this reason, this model, and all such models, are jointly referred to as *linear index models*.

Now, recall that the distribution function of a random variable, evaluated at some value tells us the probability of the random variable being less than or equal to that value. There are two things to notice about this expression. First, our fitted model, $X\hat{\beta}_n$ is modelling a probability function. Hence, the interpretation of the estimated coefficients is different to that often ascribed to those of a linear regression model. Here, β does not tell us the marginal impact of a change in x_i on y_i, but rather it tells us something about the impact of a change in x_i on the probability that $y_i = 1$. Because probability functions are likely to be extremely non-linear functions of their arguments, great care must be taken when making such interpretations. Consider Figure 1, which depicts two distribution functions, one being that of the logistic distribution

$$F(x) = (1 + e^{-x})^{-1}$$

and the other being that of the standard Normal distribution

$$\Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp\{-z^2/2\}\, \mathrm{d}z.$$

There are a number of observations that can be made here. First, observe that there is no closed form solution for the standard Normal distribution function, or any other Normal distribution function for that matter, whereas the logistic distributions has a relatively simple closed form expression.[1] Second, the standard Normal distribution has much thinner tails than does the logistic, by which is meant that the probability of observing relatively large positive or negative draws is much higher for a logistic distribution than it is for a standard Normal. Finally, and of particular relevance to this discussion, is that the tangents to these functions have very different slopes depending on the value at which the functions are evaluated. For example, the tangents $f(x) = F'(x)$ and $\phi(x) = \Phi'(x)$ are relatively steep at $x = 0$. By way of contrast, at $x = \pm 5$, the tangents to both functions have slopes very close to zero. Consequently, the impact of a change in x on the probability of $y = 1$ will vary depending on where you choose to measure the effect. There is a commonly used strategy that people use to address this problem, although it is implemented in two distinct ways in the literature. The common strategy is to provide the average response to a change in x. One way that this is done is to set each of the explanatory variables to their average value and report the impact of a change in the j-th element of x' as

$$\left.\frac{\partial F(x'\beta)}{\partial \dot{x}_j}\right|_{x=\overline{x}} = \left.\frac{\partial F(x'\beta)}{\partial x'\beta}\frac{\partial x'\beta}{\partial \dot{x}_j}\right|_{x=\overline{x}} = f(\overline{x}'\beta)\beta_j,$$

where $\dot{x}_j$ denotes the j-th element of the vector x. This highlights the fact that the marginal response is not just β_j as it is in a linear model. One drawback with this approach is that there may be no single element in the sample that actually looks like $\overline{x}$. It is the usual problem of the typical family being comprised of 1.2 adults, 2.7 kids and 0.9 pets, even though there is no such family. The alternative, and preferred, approach is to report an average response of the form

$$\frac{\partial F(x'\beta)}{\partial \dot{x}_j} = \frac{\beta_j}{n}\sum_{i=1}^{n} f(x_i'\beta).$$

The advantage of this approach is that each observation in the sample contributes to the outcome and, moreover, each contribution to the outcome is an observable value for x_i. This approach will be assumed when thinking about subsequent models encountered.

So, interpretation of the coefficients can be problematic. Clearly a unit change in x_j does not result in a β_j unit change in the dependent variable. Nevertheless, the sign of β_j is indicative of the direction in which the probability will move and we do have techniques in place that allow us to think about what the average response to a change in x_j might be. So what other problems might we encounter? One fairly obvious one stems from the Bernoulli nature of y_i. Observe that[2]

$$\begin{aligned} \mathrm{V}\left[y_i \middle| x_i\right] &= \mathrm{E}\left[y_i^2 \middle| x_i\right] - \left\{\mathrm{E}\left[y_i \middle| x_i\right]\right\}^2 = F\left(x_i'\beta\right) - \left\{F\left(x_i'\beta\right)\right\}^2 \\ &= F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]. \end{aligned} \tag{2}$$

The thing to notice here is not that the variance appears to be a moderately icky function but rather it is a moderately icky function indexed by i, the observation number. That is,

[1] A *closed form* expression is one that can be expressed as other than an unresolved integral or infinite series, although there are some special functions defined in term of infinite series that would typically be thought of as closed form expressions.

[2] Because y_i can only take the values 0 or 1, it follows that $y_i^2 = y_i$. Hence, $\mathrm{E}\left[y_i^2 \middle| x_i\right] = \mathrm{E}\left[y_i \middle| x_i\right] = F\left(x_i'\beta\right)$.

the linear probability model has heteroskedastic disturbances. The only real implication of this is that the standard OLS standard errors and t-statistics are incorrect. Of course, we have discussed how one might think about constructing a feasible generalized least squares estimator by constructing a model of the disturbance variance. You might think that we are well placed to do this given that we know (2). Clearly one problem is that the u_i are unobservable and it remains for us to establish that the sample counterparts of the υ_i, $i = 1, \ldots, n$, are able to be of much help here. In particular, the biggest issue is the fact that we observe no measure of the magnitude of the y_i^* in the y_i. An obvious alternative, that may work for us is to simply 'White-wash' the OLS standard errors.[3] At least this would yield valid standard errors for the construction of confidence intervals and test statistics, assuming that the 'estimated' υ_i are up to the task.

An immediate problem with falling back on OLS in this context is that the non-linearity of the model implies that OLS will almost certainly yield a biased estimator and, as it transpires, one that is inconsistent too (see, Horrace and Oaxaca, 2006). If you cannot estimate the coefficients adequately then it is not clear how the residuals will behave.

There remains one final problem that many view as the straw that breaks the camel's back for the linear probability model and that is the certainty that any linear model with non-zero β will predict values for the dependent variable that are outside the interval $[0, 1]$. That is, if you extrapolate any straight line, with non-zero slope far enough in either direction, then the line will eventually move outside the interval $[0, 1]$. This is clearly an issue for forecast probabilities. Indeed, if one moves far enough to the left, so that the forecast value of the line is negative, an implication is that the corresponding variance will be negative too. All up, a bad day at the office for the linear probability model.

The combination of all of these shortcomings drove people to consider non-linear regression model in these cases. The essential feature that such models had to have was that they need to behave like probabilities and observe the rules that probabilities observe. An obvious class of functions for this are probability functions and, in particular, distribution functions, because that is what we are trying to model. Once working with probability functions, maximum likelihood immediately suggests itself as the appropriate technique, although it will be clear how GMM might also be applied. It is to this that we turn next.

3 Maximum Likelihood in Binary Response Models

3.1 The Log-Likelihood Function

For a simple random sample of size n, y_i will take the value unity n_1 (say) times and the value zero n_2 times, where $n = n_1 + n_2$. If we order our sample so that the first n_1 observations are those for which y_i takes the value unity, so that $y_i = 1$ $(i = 1, \ldots, n_1)$ and $y_i = 0$ $(i = n_1 + 1, \ldots, n)$, the joint density of our sample is

$$g(y|X) = \underbrace{F\left(x_1'\beta\right) \times \ldots \times F\left(x_{n_1}'\beta\right)}_{n_1 \ terms}$$

[3] To 'White-wash' the OLS standard errors is to use White's heteroskedastic-robust standard errors, as discussed at Equation (7) of the GLS handout.

$$\times \underbrace{\left[1 - F\left(x'_{n_1+1}\beta\right)\right] \times \ldots \times \left[1 - F\left(x'_n\beta\right)\right]}_{n_2 \text{ terms}}$$

$$= \left\{\prod_{i=1}^{n_1} F\left(x'_i\beta\right)\right\}\left\{\prod_{i=n_1+1}^{n} \left[1 - F\left(x'_i\beta\right)\right]\right\} = \prod_{i=1}^{n} \left[F\left(x'_i\beta\right)\right]^{y_i}\left[1 - F\left(x'_i\beta\right)\right]^{(1-y_i)},$$

where $y = (y_1, \ldots, y_n)'$ and the $n \times k$ matrix $X = [x_1, \ldots, x_n]'$.[4] We then have the log-likelihood function

$$\ln \mathcal{L}(\beta; y, X) = \sum_{i=1}^{n} \left\{y_i \ln\left[F\left(x'_i\beta\right)\right] + (1 - y_i) \ln\left[1 - F\left(x'_i\beta\right)\right]\right\}.$$

Importantly, the log-likelihood function has this general form regardless of choice of $F(\cdot)$.

3.2 The Score Function

We shall adopt the notation $S(\beta)$ to denote the score function; hence, denoting by $\partial/\partial\beta$ the $k \times 1$ vector of partial differential operators $[\partial/\partial\beta_1, \ldots, \partial/\partial\beta_k]'$,

$$\begin{aligned}
S(\beta) &= \frac{\partial \ln \mathcal{L}(\beta; y, X)}{\partial \beta} \\
&= \frac{\partial}{\partial \beta} \sum_{i=1}^{n} \left\{y_i \ln\left[F\left(x'_i\beta\right)\right] + (1 - y_i) \ln\left[1 - F\left(x'_i\beta\right)\right]\right\} \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta} \left\{y_i \ln\left[F\left(x'_i\beta\right)\right] + (1 - y_i) \ln\left[1 - F\left(x'_i\beta\right)\right]\right\} \\
&= \sum_{i=1}^{n} \left\{y_i \frac{\partial \ln\left[F\left(x'_i\beta\right)\right]}{\partial \beta} + (1 - y_i) \frac{\partial \ln\left[1 - F\left(x'_i\beta\right)\right]}{\partial \beta}\right\} \\
&= \sum_{i=1}^{n} \left\{y_i \times \frac{\partial \ln\left[F\left(x'_i\beta\right)\right]}{\partial F\left(x'_i\beta\right)} \times \frac{\partial F\left(x'_i\beta\right)}{\partial x'_i\beta} \times \frac{\partial x'_i\beta}{\partial \beta}\right. \\
&\qquad \left. + (1 - y_i) \times \frac{\partial \ln\left[1 - F\left(x'_i\beta\right)\right]}{\partial\left[1 - F\left(x'_i\beta\right)\right]} \times \frac{\partial\left[1 - F\left(x'_i\beta\right)\right]}{\partial x'_i\beta} \times \frac{\partial x'_i\beta}{\partial \beta}\right\} \\
&= \sum_{i=1}^{n} \left\{y_i \times \frac{1}{F\left(x'_i\beta\right)} \times f\left(x'_i\beta\right) \times x_i\right. \\
&\qquad \left. + (1 - y_i) \times \frac{1}{\left[1 - F\left(x'_i\beta\right)\right]} \times \left[-f\left(x'_i\beta\right)\right] \times x_i\right\} \\
&= \sum_{i=1}^{n} x_i \left\{\frac{y_i}{F\left(x'_i\beta\right)} - \frac{(1 - y_i)}{\left[1 - F\left(x'_i\beta\right)\right]}\right\} f\left(x'_i\beta\right) \\
&= \sum_{i=1}^{n} x_i \frac{\left[y_i - F\left(x'_i\beta\right)\right] f\left(x'_i\beta\right)}{F\left(x'_i\beta\right)\left[1 - F\left(x'_i\beta\right)\right]}
\end{aligned}$$

[4] In the final expression remember that y_i only takes the values 0 or 1 hence so too does $1 - y_i$. However, the latter is equal to zero when the former is equal to 1 and vice versa. So,

$$\left[F\left(x'_i\beta\right)\right]^{y_i}\left[1 - F\left(x'_i\beta\right)\right]^{(1-y_i)} = \begin{cases} F\left(x'_i\beta\right), & \text{if } y_i = 1, \\ 1 - F\left(x'_i\beta\right), & \text{if } y_i = 0. \end{cases}$$

$$= \sum_{i=1}^{n} x_i \nu_i(\beta), \tag{3}$$

where

$$\nu_i(\beta) = \frac{[y_i - F(x_i'\beta)]\, f(x_i'\beta)}{F(x_i'\beta)\,[1 - F(x_i'\beta)]}. \tag{4}$$

When evaluated at the maximum likelihood estimator $\hat{\beta}$, we shall call $\nu_i(\hat{\beta})$ a *generalized residual*. This name comes by analogy with the linear regression model where residuals, $e_i = y_i - x_i'\hat{\beta}$ say, satisfy the orthogonality condition

$$\sum_{i=1}^{n} x_i e_i = 0 \quad \text{or} \quad X'e = 0,$$

where $e = [e_1, \ldots, e_n]'$. Here the score function is $S(\beta) = X'\nu(\beta)$, where $\nu(\beta) = [\nu_1(\beta), \ldots, \nu_n(\beta)]'$, and we see that the maximum likelihood estimator is that value $\hat{\beta}$ which satisfies

$$S(\hat{\beta}) = \sum_{i=1}^{n} x_i \nu_i(\hat{\beta}) = 0 \quad \text{or} \quad S(\hat{\beta}) = X'\nu(\hat{\beta}) = 0.$$

Next observe that, conditional on x_i, both of $f(x_i'\beta)$ and $F(x_i'\beta)$ are non-random. Hence

$$\begin{aligned} \mathrm{E}\left[\nu_i(\beta) \middle| x_i\right] &= \mathrm{E}\left[\frac{[y_i - F(x_i'\beta)]\, f(x_i'\beta)}{F(x_i'\beta)\,[1 - F(x_i'\beta)]} \middle| x_i \right] \\ &= \frac{[\mathrm{E}[y_i | x_i] - F(x_i'\beta)]\, f(x_i'\beta)}{F(x_i'\beta)\,[1 - F(x_i'\beta)]} \\ &= 0, \end{aligned} \tag{5}$$

where the final equality follows from the fact that $\mathrm{E}[y_i | x_i] = F(x_i'\beta)$. An implication of this result is that

$$\mathrm{E}[S(\beta) \mid X] = \mathrm{E}\left[\sum_{i=1}^{n} x_i \nu_i(\beta) \mid X \right] = \sum_{i=1}^{n} x_i \,\mathrm{E}[\nu_i(\beta)] = 0.$$

That is, the score has zero expectation, which is a general property of maximum likelihood estimators.

3.3 The Information Matrix

We have just shown that the score has zero expectation with respect to the true density of y_i given x_i. It follows that the conditional variance of $S(\beta)$ is

$$\begin{aligned} \mathrm{V}\left[S(\beta) \middle| X\right] &= \mathrm{E}\left[S(\beta) S(\beta)' \middle| X\right] \\ &= \mathrm{E}\left[\left\{ \sum_{i=1}^{n} x_i \nu_i(\beta) \right\} \left\{ \sum_{j=1}^{n} x_j' \nu_j(\beta) \right\} \middle| X \right] \\ &= \mathrm{E}\left[\sum_{i=1}^{n} x_i x_i' \nu_i(\beta)^2 + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} x_i x_j' \nu_i(\beta) \nu_j(\beta) \middle| X \right] \end{aligned}$$

$$= \sum_{i=1}^{n} x_i x_i' \mathrm{E}\left[\nu_i(\beta)^2 \middle| x_i\right] + \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} x_i x_j' \mathrm{E}\left[\nu_i(\beta) \middle| x_i\right] \mathrm{E}\left[\nu_j(\beta) \middle| x_j\right],$$

where the previous equality follows from the independence of y_i and y_j for all $i \neq j$ which, in turn, is a consequence of our assumption that the data comes from a simple random sample. From (5) we see that the cross-product terms are zero and so

$$\mathrm{V}\left[S(\beta) \middle| X\right] = \sum_{i=1}^{n} x_i x_i' \mathrm{E}\left[\nu_i(\beta)^2 \middle| x_i\right],$$

meaning that our problem reduces to the calculation of $\mathrm{E}\left[\nu_i(\beta)^2 \middle| x_i\right]$. As an aside, given that $\mathrm{E}\left[\nu_i(\beta) \middle| x_i\right] = 0$ for all i, it follows that $\mathrm{E}\left[\nu_i(\beta)^2 \middle| x_i\right] = \mathrm{V}\left[\nu_i(\beta) \middle| x_i\right]$. Noting that

$$\nu_i(\beta)^2 = \frac{\left[y_i - F\left(x_i'\beta\right)\right]^2 f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)^2 \left[1 - F\left(x_i'\beta\right)\right]^2},$$

we see that, using (2),

$$\begin{aligned}
\mathrm{E}\left[\nu_i(\beta)^2 \middle| x_i\right] &= \frac{\mathrm{E}\left[\left[y_i - F\left(x_i'\beta\right)\right]^2 \middle| x_i\right] f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)^2 \left[1 - F\left(x_i'\beta\right)\right]^2} \\
&= \frac{\mathrm{V}\left[y_i \middle| x_i\right] f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)^2 \left[1 - F\left(x_i'\beta\right)\right]^2} \\
&= \frac{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right] f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)^2 \left[1 - F\left(x_i'\beta\right)\right]^2} \\
&= \frac{f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]}.
\end{aligned}$$

Consequently, we obtain

$$\mathrm{V}\left[S(\beta) \middle| X\right] = \sum_{i=1}^{n} \frac{f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]} x_i x_i' = \sum_{i=1}^{n} \delta_i x_i x_i', \tag{6}$$

where

$$\delta_i = \frac{f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]}.$$

Note that writing $\Delta = \mathrm{diag}\{\delta_1, \ldots, \delta_n\}$ we can write $\mathrm{V}\left[S(\beta) \middle| X\right] = X'\Delta X$, which is sometimes a more convenient notation. Finally observe that the variance of $S(\beta)$ is simply the information matrix $\mathcal{I}$, so that $\mathcal{I} = X'\Delta X$.

3.4 The Hessian

A hessian matrix is a matrix of second derivatives. Let us define the column vector of differential operators $\partial/\partial\beta = [\partial/\partial\beta_1, \ldots, \partial/\partial\beta_k]'$ and observe that if $h(\beta)$ is a $p \times 1$ vector then

$$\frac{\partial h(\beta)'}{\partial \beta} = \left[\frac{\partial h(\beta)}{\partial \beta_1}, \ldots, \frac{\partial h(\beta)}{\partial \beta_k}\right]'$$

is a $k \times p$ matrix, i.e. differentiating a row vector with respect to a scalar yields a vector of the same dimension as the one being differentiated. Here we are seeking

$$H(\beta) = \frac{\partial^2 \ln \mathcal{L}(\beta; y, X)}{\partial \beta \, \partial \beta'} = \frac{\partial S(\beta)'}{\partial \beta} = \sum_{i=1}^{n} x_i' \frac{\partial \nu_i(\beta)}{\partial \beta}.$$

Let us write

$$\nu_i(\beta) = \frac{[y_i - F(x_i'\beta)] f(x_i'\beta)}{F(x_i'\beta)[1 - F(x_i'\beta)]} = \frac{\gamma_i(\beta)}{\theta_i(\beta)},$$

whence it follows that the elements of $\partial \nu_i(\beta)/\partial \beta'$ are

$$\frac{\partial \nu_i(\beta)}{\partial \beta_j} = \frac{\Gamma_{ij}(\beta)\theta_i(\beta) - \gamma_i(\beta)\Theta_{ij}(\beta)}{\theta_i(\beta)^2}, \qquad j = 1, \ldots, k,$$

where

$$\Gamma_{ij}(\beta) = \frac{\partial \gamma_i(\beta)}{\partial \beta_j} \qquad \text{and} \qquad \Theta_{ij}(\beta) = \frac{\partial \theta_i(\beta)}{\partial \beta_j}.$$

Now,

$$\gamma_i(\beta) = [y_i - F(x_i'\beta)] f(x_i'\beta)$$

and so

$$\begin{aligned}\Gamma_{ij}(\beta) &= -\frac{\partial F(x_i'\beta)}{\partial x_i'\beta} \frac{\partial x_i'\beta}{\partial \beta_j} f(x_i'\beta) + [y_i - F(x_i'\beta)] \frac{\partial f(x_i'\beta)}{\partial x_i'\beta} \frac{\partial x_i'\beta}{\partial \beta_j} \\ &= \left\{ -f(x_i'\beta)^2 + [y_i - F(x_i'\beta)] \frac{\partial f(x_i'\beta)}{\partial x_i'\beta} \right\} x_{ij}.\end{aligned}$$

Similarly,

$$\theta_i(\beta) = F(x_i'\beta)[1 - F(x_i'\beta)]$$

and so

$$\Theta_{ij}(\beta) = [1 - 2F(x_i'\beta)] f(x_i'\beta) x_{ij}.$$

Gathering these results yields

$$\begin{aligned}\frac{\partial \nu_i(\beta)}{\partial \beta_j} = &\left[\left\{ -f(x_i'\beta)^2 + [y_i - F(x_i'\beta)] \frac{\partial f(x_i'\beta)}{\partial x_i'\beta} \right\} F(x_i'\beta)[1 - F(x_i'\beta)] x_{ij} \right. \\ &\left. - [y_i - F(x_i'\beta)][1 - 2F(x_i'\beta)] f(x_i'\beta)^2 x_{ij} \right] \\ &\div \left\{ F(x_i'\beta)[1 - F(x_i'\beta)] \right\}^2.\end{aligned}$$

If we gather these results for all $j = 1, \ldots, k$ we obtain

$$\frac{\partial \nu_i(\beta)}{\partial \beta} = \alpha_i x_i,$$

where

$$\begin{aligned}\alpha_i = &\left[\left\{ -f(x_i'\beta)^2 + [y_i - F(x_i'\beta)] \frac{\partial f(x_i'\beta)}{\partial x_i'\beta} \right\} F(x_i'\beta)[1 - F(x_i'\beta)] \right. \\ &\left. - [y_i - F(x_i'\beta)][1 - 2F(x_i'\beta)] f(x_i'\beta)^2 \right] \\ &\div \left\{ F(x_i'\beta)[1 - F(x_i'\beta)] \right\}^2.\end{aligned}$$

Finally, we see that

$$H(\beta) = \sum_{i=1}^{n} \alpha_i x_i x_i'.$$

Clearly our expression for the hessian is fairly messy although, on the bright side, it is rarely what we are interested in. Consider next

$$\mathrm{E}\left[H(\beta)\middle| X\right] = \sum_{i=1}^{n} \mathrm{E}\left[\alpha_i\middle| x_i\right] x_i x_i'.$$

Conditional on x_i the only random terms are the y_i but these only occur in the form $y_i - F\left(x_i'\beta\right)$, which we have already established has zero expectation. Hence we have the reduction

$$\mathrm{E}\left[\alpha_i\middle| x_i\right] = \frac{-f\left(x_i'\beta\right)^2}{F\left(x_i'\beta\right)\left[1 - F\left(x_i'\beta\right)\right]} = -\delta_i.$$

This yields our final result of $-\mathrm{E}\left[H(\beta)\middle| X\right] = \mathrm{E}\left[S(\beta)S(\beta)'\middle| X\right] = \mathcal{I}$, which is simply a demonstration of the information equality that holds for all maximum likelihood estimators.

4 Choices for $F(x_i'\beta)$: Probit and Logit models

The preceding analysis has been quite general in the sense that the conditional density of the disturbances in (1) is arbitrary (beyond our symmetry assumption). Here we will explore the two most common choices.

4.1 The Probit Model

If we suppose that $u_i|x_i \sim N(0,1)\ \forall\ i = 1, \ldots, n$, then we have

$$f\left(u_i\right) \equiv \phi\left(u_i\right) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u_i^2}{2}\right\}$$

and

$$F\left(x_i'\beta\right) \equiv \Phi\left(x_i'\beta\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x_i'\beta} \exp\left\{-\frac{u_i^2}{2}\right\} du_i.$$

In this model, which is sometimes called a *normit* model because of the normality assumption, the generalized residual is

$$\nu_i(\beta) = \frac{\left[y_i - \Phi\left(x_i'\beta\right)\right]\phi\left(x_i'\beta\right)}{\Phi\left(x_i'\beta\right)\left[1 - \Phi\left(x_i'\beta\right)\right]}. \tag{7}$$

Note that $\Phi\left(x_i'\beta\right)$ is a tail probability that can be obtained from tables of the standard normal distribution (or calculated numerically by computer) and

$$\phi\left(x_i'\beta\right) = (2\pi)^{-1/2} \exp\left\{-0.5\left(x_i'\beta\right)^2\right\}$$

is simply a number that can be calculated without any problems. The only issue that sometimes arises in the calculation of $\nu_i(\beta)$ is that some computer packages give $\Phi\left(x_i'\beta\right)$

to relatively few decimal places and so, for large values of $|x_i'\beta|$, one of $\Phi(x_i'\beta)$ and $1-\Phi(x_i'\beta)$ will be treated as unity while the other will take the value zero. This will generate errors of the *Division by zero* kind which can be a problem. Should you ever encounter such a situation there are approximations available that will let you calculate the $\nu_i(\beta)$ without such divisions.

The definition of the score in probit models follows immediately from (3) and (7). Similarly, the information matrix follows directly from (6) on writing

$$\delta_i = \frac{\phi(x_i'\beta)^2}{\Phi(x_i'\beta)\left[1-\Phi(x_i'\beta)\right]}.$$

4.2 The Logit Model

The logit model is named for the choice of the logistic distribution function for $F(x_i'\beta)$. In particular,

$$F(x_i'\beta) \equiv \Lambda(x_i'\beta) = \frac{\exp\{x_i'\beta\}}{1+\exp\{x_i'\beta\}}.$$

The corresponding density function of u_i given x_i is

$$f(u_i) \equiv \lambda(u_i) = \frac{\exp\{u_i\}}{\left[1+\exp\{u_i\}\right]^2},$$

is called the sech-squared distribution.[5] Given these choices we see that the generalized residual becomes

$$\nu_i(\beta) = \frac{\left[y_i - \Lambda(x_i'\beta)\right]\lambda(x_i'\beta)}{\Lambda(x_i'\beta)\left[1-\Lambda(x_i'\beta)\right]}. \tag{8}$$

Obviously the score follows immediately on substitution of (8) into (3). To obtain the information matrix observe that

$$\begin{aligned}
\delta_i &= \frac{\lambda(x_i'\beta)^2}{\Lambda(x_i'\beta)\left[1-\Lambda(x_i'\beta)\right]} \\
&= \frac{\exp\{2x_i'\beta\}/\left[1+\exp\{x_i'\beta\}\right]^4}{\exp\{x_i'\beta\}/\left[1+\exp\{x_i'\beta\}\right]^2} \\
&= \frac{\exp\{x_i'\beta\}}{\left[1+\exp\{x_i'\beta\}\right]^2} = \lambda(x_i'\beta).
\end{aligned} \tag{9}$$

One attractive feature of this model is that it yields closed form results, there are no unresolved integrals as seen in the probit model with $\Phi(x_i'\beta)$. Another attractive feature is that exponential functions are readily available in most software.

4.3 Comparing Probit and Logit Estimates

Consider two random variables $\eta_1 \sim N(0,1)$ and η_2 has a logistic distribution. It can be shown that $V[\eta_1] = 1$ and $V[\eta_2] = \pi^2/3$. Consequently, our two sets of estimates are not comparable because they are estimates from models with different variances. To make the estimates comparable we must make the distribution variances the same. This means

[5] This name comes from an alternative representation of the density function which we will not pursue further here.

that we must either divide the logit estimates by $\pi/\sqrt{3}$ to make them comparable to the probit estimates or multiply the probit estimates by this same amount to make them comparable with the logit estimates. It has been argued that a better approximation would involve either multiplying the probit estimates by 1.6 or, equivalently, multiplying the logit estimates by $1/1.6 \approx 0.625$. Unfortunately this approximation works best at $x_i'\beta = 0$, which is where $F(x_i'\beta) = 0.5$, i.e. the centre of the distribution. The approximation works less well as $x_i'\beta$ becomes large in absolute value.

As an aside, it is worth noting that approximations have also been determined which relate the coefficient estimates from (say) a logit model to those of a linear probability model. If we denote by $\hat{\beta}_{LP}$ estimates obtained from the linear probability model and by $\hat{\beta}_L$ estimates obtained from the corresponding logit model, it can be shown that

$$\hat{\beta}_{LP} \approx \begin{cases} 0.25\hat{\beta}_L & \text{except for the constant term} \\ 0.25\hat{\beta}_L + 0.5 & \text{for the constant term.} \end{cases}$$

Similar kinds of relationships can be determined between the coefficients of the linear probability model and those of the probit model. Again, however, such approximate relationships are likely to work best in the centre of the distribution but less well out in the tails.

Finally, the optimality properties of maximum likelihood estimates require correct distributional assumptions. Therefore it cannot be the case that both probit and logit estimators are consistent. If one is, the other will not be because it is based on the wrong distributional assumption.

4.4 Interpreting coefficients

The different models have different interpretations. In each case the fitted model is an estimate of the regression function, or conditional expectation of y_i given x_i, which is simply the probability that y_i takes the value unity for given x_i. For the linear probability model the regression function is $x_i'\beta$, it is $\Phi(x_i'\beta)$ and $\Lambda(x_i'\beta)$ for the probit and logit models, respectively. More generally, we have $\Pr(y_i = 1) = F(x_i'\beta)$. Letting x_{ij} denote the j-th element of x_i, $j = 1, \ldots, k$, and β_j the corresponding coefficient, we have

$$\frac{\partial x_i'\beta}{\partial x_{ij}} = \beta_j \qquad \text{(Linear Model)}$$

$$\frac{\partial \Phi(x_i'\beta)}{\partial x_{ij}} = \phi(x_i'\beta)\beta_j \qquad \text{(Probit Model)}$$

$$\frac{\partial \Lambda(x_i'\beta)}{\partial x_{ij}} = \frac{\exp\{x_i'\beta\}}{[1+\exp\{x_i'\beta\}]^2}\beta_j = \lambda(x_i'\beta)\ \beta_j \qquad \text{(Logit Model)}$$

and, in general,

$$\frac{\partial F(x_i'\beta)}{\partial x_{ij}} = \frac{\partial F(x_i'\beta)}{\partial x_i'\beta}\frac{\partial x_i'\beta}{\partial x_{ij}} = f(x_i'\beta)\beta_j.$$

The linear probability model predicts a constant response to *ceteris paribus* changes in x_{ij} (say), whereas the other two models predict non-linear responses. Although the non-linearity of these responses makes interpretation difficult we can say the following in response to marginal changes in a single explanatory variables x_{ij}:

1. The sign of the marginal effect will be the same as that of the coefficient β_j whenever $f(x_i'\beta) > 0$ (i.e., pretty much always).

2. The magnitude of the effect depends upon:

   (a) the magnitude of β_j;

   (b) the magnitude of $x_i'\beta$, i.e., all the x's;

   (c) choice of $f(\cdot)$.

3. The largest marginal effect will occur at that x_i where $\Pr(y_i = 1) = 0.5$ and will be symmetric about that point (because we typically assume that $f(\cdot)$ is symmetric about zero, e.g., the disturbances are symmetrically distributed about zero, which then implies that they have zero expectation).

Note that, for any choice of x_i,

$$\frac{\frac{\partial F(x_i'\beta)}{\partial x_{ik}}}{\frac{\partial F(x_i'\beta)}{\partial x_{ij}}} = \frac{f(x_i'\beta)\,\beta_k}{f(x_i'\beta)\,\beta_j} = \frac{\beta_k}{\beta_j}$$

so that the ratio of coefficients gives the ratio of marginal effects.

4.5 An Alternative Derivation of the Logit Model

Consider the odds-ratio

$$R_i = \frac{\Pr(y_i = 1)}{\Pr(y_i = 0)} = \frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)}.$$

Now, $0 \leq \Pr(y_i = 1) \leq 1$. At the extremes of this range we see that if $\Pr(y_i = 1) = 0$ then $R_i = 0$, and if $\Pr(y_i = 1) = 1$ then R_i diverges to ∞. Clearly, $-\infty < \ln R_i < \infty$. The quantity $\ln R_i$, the log odds-ratio, is known as the logit. The question then is, how can you model the logit as a function of some explanatory variables?

Suppose that we express the odds ratio as a function of the explanatory variables: $R_i = R(x_i)$. A first step to modelling the logit is to specify some functional form. The simplest (non-constant) form that we might think of is simply a linear model, e.g., $\ln R(x_i) = x_i'\beta$. This choice leads us to the logit model. In particular, if $\ln R_i = x_i'\beta$ then

$$\begin{aligned}
& R(x_i) = \exp\{x_i'\beta\} \\
\Rightarrow \quad & \frac{\Pr(y_i = 1)}{1 - \Pr(y_i = 1)} = \exp\{x_i'\beta\} \\
\Rightarrow \quad & \Pr(y_i = 1) = \exp\{x_i'\beta\}\,[1 - \Pr(y_i = 1)] \\
\Rightarrow \quad & \Pr(y_i = 1)\,[1 + \exp\{x_i'\beta\}] = \exp\{x_i'\beta\} \\
\Rightarrow \quad & \Pr(y_i = 1) = \frac{\exp\{x_i'\beta\}}{1 + \exp\{x_i'\beta\}} = \frac{1}{1 + \exp\{-x_i'\beta\}}
\end{aligned}$$

One reason people find the odds-ratio of interest is because it provides a relatively straight-forward interpretation of parameters. Obviously $\partial \ln R(x_i)/\partial x_{ij} = \beta_j$. Similarly, if we define $\overline{x}_{ij}$ to be x_i with the element x_{ij} removed then we might define $R(x_i) \equiv$

$R(\overline{x}_{ij}, x_{ij})$. Now suppose that we perturb x_{ij} by δ, then the proportionate change in the odds ratio is

$$\frac{R(\overline{x}_{ij}, x_{ij}+\delta)}{R(\overline{x}_{ij}, x_{ij})} = \exp\{\beta_j \delta\}.$$

For $\delta = 1$ this reduces to $\exp\{\beta_j\}$. In the general case we see that the percentage change in the odds-ratio is $100\,(\exp\{\beta_j\delta\} - 1)$.

5 Hypothesis Testing in Probit and Logit Models

The models that we have explored are all models that can be estimated by maximum likelihood and, as such all of the trinity of likelihood-based tests are available to us. Of particular interest is the fact that, if one wishes to test single restrictions, like exclusion restrictions on single coefficients, then t-tests with critical values from a standard Normal distribution will be applicable. Otherwise, people are likely to reach for LR tests, because the maximized log-likelihood is one statistic that is likely to be printed out with standard estimation routines. A really useful reference for LM-like tests in the context of unit record data (which is basically survey information about individuals, be they people, households, businesses, or whatever) is Pagan and Vella (1989).

6 R

In the broader statistical world, probit and logit model are generally thought of a generalized linear models with probit and logit link functions, respectively. We will not explore this further than to say that the relevant estimation commands for fitting a probit model are of the form

```
mdl.glm=glm(dependent variable ~ list of explanators, family = binomial(link
= "probit"))
mdl.glm.stats=summary.glm(mdl.glm)
```

In order to fit a logit model, one simply need replace the word "probit" with "logit", as appropriate. Like any R command the are very many options and tuning parameters that can be used with this command, you can investigate them for yourselves as and when the need arises.

7 An Example

In this example I am going to fit a linear regression model, a probit model and a logit model to the same set of data. You can find the data in the Stata file `my_loanapp.dta`. When you import the data you will see that the data comes with definitions of the variable. The code that I used to generate the results is available in the file `BRM.R`. The various results can be found in the subsequent tables of code listings.

Looking first to Listing 1 we see that the `lm` command gives coefficient estimates and very little else. More information is available from the `summary.lm` command but not the maximized log-likelihood (which is a surprising omission). Thankfully, it can be recover from the output of the `lm` command using the `logLik` command. Strictly, it could have

been derived from the information already there on noting that the Residual standard error is given by the formula

$$s_e = \sqrt{\frac{SSE}{n-k}},$$

where n is the sample size, k is the total number of coefficients estimated in the regression function (including the intercept!) and SSE is the sum of squared residuals. Hence, we can recover SSE from this quantity. Note that $n-k$ is listed here as the degrees of freedom (There were originally 1989 observations in the data set, of which 18 are excluded on account of missing data in one or more variables, giving an effective sample size of $n = 1971$ in this example), and k is the numerator degrees of freedom from the F-statistic plus 1 (15 in this example). The extra 1 is needed because the F-statistic is that used to test the null hypothesis that all of the slope coefficients are zero against the alternative that at least one of them is not. There are 14 variables in this example plus a constant, making $k = 15$. In summary, the $n-k$ that we need to work with is 1956. Hence

$$SSE = 1956 \times (0.3057)^2 = 1956 \times (0.3057)^2 = 182.7454.$$

Note that, in order to minimize rounding error, I actually performed this calculation in R using the command:

```
SSE=1956*ols.summary$sigma*ols.summary$sigma
```

as `ols.summary$sigma` is the name under which s_e has been stored by R in this example.

Next, recalling that the log-likelihood here is of the form

$$\ln \mathcal{L}(\beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta),$$

with $\hat{\beta} = (X'X)^{-1}X'y$ and $\hat{\sigma}^2 = (y - X\hat{\beta})'(y - X\hat{\beta})/n = SSE/n$, we can obtain the maximized value of the log-likelihood on making the relevant substitutions. Before doing so, a word of warning. Whenever you use a package like R you need to check exactly what it returns as the value of the maximized log-likelihood. Some packages include the various constants and some don't.[6] So let's check what R does.

To begin, observe that

$$\begin{aligned}
\ln \mathcal{L}(\hat{\beta}, \hat{\sigma}^2) &= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}(y - X\hat{\beta})'(y - X\hat{\beta}) \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\frac{SSE}{n}\right) - \frac{n}{2SSE}SSE \\
&= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}[\ln(SSE) - \ln(n)] - \frac{n}{2} \\
&= -\frac{n}{2}[\ln(2\pi) - \ln(n) + 1] - \frac{n}{2}\ln(SSE) \\
&= -\frac{n}{2}[\ln\left(\frac{2\pi * SSE}{n} + 1\right).
\end{aligned}$$

Using the values from R we have

$$\ln \mathcal{L}(\hat{\beta}, \hat{\sigma}^2) = -\frac{1971}{2}\left[\ln\left(\frac{2\pi * 182.7454}{1971(\log\right)\right) + 1\right] = -453.0094,$$

[6] Given that the primary use of the maximized value of the log-likelihood is going to be the construction of LR tests, the constants will cancel out when the difference is taken between the restricted and unrestricted log-likelihoods.

as required. (Yeah!) So, R is giving the complete maximized log-likelihood and not some truncated version thereof.

As mentioned in the previous section, R treats the probit and logit models as members of the class of generalized linear models. Nevertheless, the `glm` command gives us maximum likelihood estimates. Looking at Listing 2 we see that `glm` command gives more information than does the `lm` command but less than does the `summary.glm` command and so we shall focus attention on the output of this latter command.[7] Now, most of the output should look familiar but there are three statistics that may be new to you and so merit further discussion.

The first of these statistics is called the *Null Deviance*. This is minus twice the maximized value for the log-likelihood in a model containing only the intercept as an explanatory variable.[8] This model corresponds to the restricted model used to construct the F-statistic in Listing 1. You will see that in this example the corresponding degrees of freedom is $n - k = 1970$. That is, the $n = 1971$ effective observations, after observations containing missing values have been deleted, minus the $k = 1$ coefficient to be estimated. (Recall that we can't estimate a variance in a binary response model and so it is essentially normalized to unity and treated as known.) In Listing 3, I have fitted a model with just an intercept and we see that the corresponding log-likelihood is -740.3466 and minus twice this amount does yield (roughly) the Null Deviance of 1481. Somewhat surprisingly, it differs from the Null Deviance in Listing 2, albeit by a small amount. I am not sure why this is.

The second statistic to discuss is the *Residual Deviance*. This is minus twice the maximized log-likelihood for the unrestricted model. Looking at the output in Listing 2 we see that

$$\text{log-likelihood} = -618.233$$

and that

$$-2 * \text{log-likelihood} = -2 \times -618.233 = 1236.466 \approx 1236.5 = \text{Residual Deviance},$$

as expected. Note that in the unrestricted model there are $n = 1971$ effective observations (nett of observations with missing values) and there are $k = 15$ coefficients estimated in the unrestricted model (14 slope coefficients and the intercept) and so the degrees of freedom here is $n - k = 1971 - 15 = 1956$. Observe also that

$$\begin{aligned} LR &= -2(\text{Restricted log-likelihood} - \text{Unrestricted log-likelihood}) \\ &= -2((-740.3466) - (-618.233)) = 244.2272 \end{aligned}$$

is the calculated LR test statistic of the null hypothesis that all of the slope coefficients are equal to zero against the two-sided alternative that at least one of the slope coefficients is not equal to zero. In Listing refprobit2 I provide estimates of the restricted model here; that is, the linear index is comprised solely of an intercept. Note that, letting $\hat{\beta}_1$ denote our estimate of the intercept in the restricted model, we should find that $\Phi(\hat{\beta}_1) = \bar{y}$.[9]

[7] Instead of the `summary.lm` and `summary.glm` commands we could have just used the `summary` command in each case. However, if you seek help on the `summary` command, via either `help(summary)` or `?summary` the R is not very helpful. It is somewhat more responsive if you ask for help on either of the commands `summary.lm` or `summary.glm`, respectively.

[8] Again, you would want to check that R is not using a truncated version of the log-likelihood obtained by omitting some constants.

[9] I note in passing that the sample average of the variable `approve` is approximately 0.877 and that $\Phi(1.162) \approx 0.877$, as predicted.

(Why?) In any event, it should also be the case that

$$LR = \text{Null Deviance} - \text{Residual Deviance} = 1476 - 1236.5 = 239.5$$

which is similar to the value of LR based on the likelihood functions. If, instead, we take the Null Deviance from Listing 3 then we find that

$$LR = 1481 - 1236.5 = 244.5,$$

which is near enough, making due allowance for rounding error.

The remaining statistic is the `AIC`, which stands for Akaike's Information Criterion. The AIC arose from the recognition that choosing a model to maximize the log-likelihood tends to favour overly large models, which is sometimes called over-fitting. A direct analogy is the coefficient of determination (R^2) in linear models. If maximizing R^2 is your objective then you will always favour larger models over more parsimonious models. In the linear model, the adjusted coefficient of determination ($\bar{R}^2$), is an attempt to add a penalty to R^2 for including variables, with relatively little explanatory power, simply because their inclusion in the model increases R^2. So too with the AIC. It penalizes the log-likelihood function for including too many variables if they are providing little explanatory power. The definition of AIC is

$$AIC = 2p - 2\ln\mathcal{L}(\hat{\theta}) = 2p - \text{Residual Deviance},$$

where p is the total number of parameters estimated in the model. In our example here $p = 15$. Therefore, we see that

$$AIC = 2 * 15 - 2 * (-618.233) = 30 + 1236.466 \approx 1266.5$$

as observed in the output.

Finally, Listing 4 provides results for the same model except now using a logit model rather than a probit model. You will see that the coefficient estimates have moved around but that the statistical significance of various coefficients remains essentially the same. Similarly the maximized log-likelihood are fairly similar. There is not much difference between a probit an a logit, although the latter distribution has slightly fatter tails and tends to work a little better in the real world. Also the availability of closed form solutions is a practical advantage, albeit not a huge one.

Bibliography

Horrace, W. C. and R. L. Oaxaca (2006). Results on the bias and inconsistency of Ordinary Least Squares for the linear probability model. *Economics Letters* 90(3), 321 – 327, ISSN 0165-1765, doi:https://doi.org/10.1016/j.econlet.2005.08.024. 5

Pagan, A. R. and F. Vella (1989). Diagnostic tests for models based on unit record data: A survey. *Journal of Applied Econometrics* 4, S29–S59. 14

Listing 1: ML Estimation of Linear Regression Model

```
> ols=lm(approve~hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+
> pubrec+mortlat1+mortlat2+vr+white, data=my_loanapp)
>
> ols

Call:
lm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign
\\
+ pubrec + mortlat1 + mortlat2 +
    vr + white, data = my_loanapp)

Coefficients:
(Intercept)         hrat        obrat      loanprc         unem         male      married
   1.044592     0.002354    -0.006025    -0.151291    -0.006550    -0.003009     0.041296
        dep          sch       cosign       pubrec     mortlat1     mortlat2           vr
  -0.005536     0.006676     0.005063    -0.285983    -0.067845    -0.132650    -0.031439
      white
   0.141490

> ols.summary=summary.lm(ols)
> ols.summary

Call:
lm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign + pubrec + mortlat1 + mortlat2 +
    vr + white, data = my_loanapp)

Residuals:
     Min       1Q   Median       3Q      Max
-1.04833  0.01876  0.07418  0.12677  0.65727

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.044592   0.050967  20.495  < 2e-16 ***
hrat         0.002354   0.001276   1.845   0.0652 .
obrat       -0.006025   0.001111  -5.421 6.66e-08 ***
loanprc     -0.151291   0.037957  -3.986 6.97e-05 ***
unem        -0.006550   0.003234  -2.025   0.0430 *
male        -0.003009   0.019087  -0.158   0.8748
married      0.041296   0.016488   2.505   0.0123 *
dep         -0.005536   0.006778  -0.817   0.4141
sch          0.006676   0.016832   0.397   0.6917
cosign       0.005063   0.041622   0.122   0.9032
pubrec      -0.285983   0.027823 -10.279  < 2e-16 ***
mortlat1    -0.067845   0.050582  -1.341   0.1800
mortlat2    -0.132650   0.067722  -1.959   0.0503 .
vr          -0.031439   0.014198  -2.214   0.0269 *
white        0.141490   0.019879   7.117 1.54e-12 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 0.3057 on 1956 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.1452,    Adjusted R-squared:  0.1391
F-statistic: 23.74 on 14 and 1956 DF,  p-value: < 2.2e-16

> ols.logLik=logLik(ols)
>
> ols.logLik
'log Lik.' -453.0094 (df=16)
```

Listing 2: ML Estimation of Probit Model

```
> probit=glm(approve~hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+
> pubrec+mortlat1+mortlat2+vr+white, family=binomial(link="probit"), data=my_loanapp)
>
> probit

Call:  glm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign + pubrec + mortlat1 + mortlat2 +
    vr + white, family = binomial(link = "probit"), data = my_loanapp)

Coefficients:
(Intercept)         hrat        obrat      loanprc         unem         male      married
    2.46739      0.01047     -0.02988     -1.00572     -0.03117     -0.03775      0.22907
        dep          sch       cosign       pubrec     mortlat1     mortlat2           vr
   -0.03725      0.04158      0.04407     -0.96528     -0.25622     -0.58443     -0.18748
      white
    0.57001

Degrees of Freedom: 1970 Total (i.e. Null);  1956 Residual
  (18 observations deleted due to missingness)
Null Deviance:      1476
Residual Deviance: 1236         AIC: 1266
> probit.summary=summary.glm(probit)
> probit.summary

Call:
glm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign + pubrec + mortlat1 + mortlat2 +
    vr + white, family = binomial(link = "probit"), data = my_loanapp)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.467391   0.306604   8.047 8.45e-16 ***
hrat         0.010469   0.006898   1.518   0.1291
obrat       -0.029881   0.006035  -4.951 7.37e-07 ***
loanprc     -1.005723   0.239389  -4.201 2.65e-05 ***
unem        -0.031173   0.017472  -1.784   0.0744 .
male        -0.037751   0.108366  -0.348   0.7276
married      0.229066   0.093089   2.461   0.0139 *
dep         -0.037247   0.038513  -0.967   0.3335
sch          0.041582   0.093806   0.443   0.6576
cosign       0.044068   0.234173   0.188   0.8507
pubrec      -0.965282   0.121759  -7.928 2.23e-15 ***
mortlat1    -0.256219   0.255706  -1.002   0.3163
mortlat2    -0.584431   0.317975  -1.838   0.0661 .
vr          -0.187485   0.080409  -2.332   0.0197 *
white        0.570011   0.095110   5.993 2.06e-09 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1476.0  on 1970  degrees of freedom
Residual deviance: 1236.5  on 1956  degrees of freedom
  (18 observations deleted due to missingness)
AIC: 1266.5

Number of Fisher Scoring iterations: 5

> probit.logLik=logLik(probit)
> probit.logLik
'log Lik.' -618.233 (df=15)
```

Listing 3: ML Estimation of the Restricted Probit Model

```
> probit2=glm(approve~1, family=binomial(link="probit"), data=my_loanapp)
>
> probit2

Call:  glm(formula = approve ~ 1, family = binomial(link = "probit"),
    data = my_loanapp)

Coefficients:
(Intercept)
      1.162

Degrees of Freedom: 1988 Total (i.e. Null);  1988 Residual
Null Deviance:      1481
Residual Deviance: 1481         AIC: 1483
> probit2.summary=summary(probit2)
>
> probit2.summary

Call:
glm(formula = approve ~ 1, family = binomial(link = "probit"),
    data = my_loanapp)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.16172    0.03621   32.09   <2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1480.7  on 1988  degrees of freedom
Residual deviance: 1480.7  on 1988  degrees of freedom
AIC: 1482.7

Number of Fisher Scoring iterations: 4

> probit2.logLik=logLik(probit2)
> probit2.logLik
'log Lik.' -740.3466 (df=1)
```

Listing 4: ML Estimation of Logit Model

```
> logit=glm(approve~hrat+obrat+loanprc+unem+male+married+dep+sch+cosign+
> pubrec+mortlat1+mortlat2+vr+white, family=binomial(link="logit"), data=my_loanapp)
> logit

Call:  glm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign + pubrec + mortlat1 + mortlat2 +
    vr + white, family = binomial(link = "logit"), data = my_loanapp)

Coefficients:
(Intercept)         hrat        obrat      loanprc         unem         male      married
    4.52976      0.01852     -0.05715     -1.89846     -0.05610     -0.06646      0.42606
        dep          sch       cosign       pubrec     mortlat1     mortlat2           vr
   -0.06529      0.07918      0.09046     -1.66816     -0.41786     -1.07987     -0.33328
      white
    1.03621

Degrees of Freedom: 1970 Total (i.e. Null);  1956 Residual
  (18 observations deleted due to missingness)
Null Deviance:      1476
Residual Deviance: 1237         AIC: 1267
> logit.summary=summary.glm(logit)
> logit.summary

Call:
glm(formula = approve ~ hrat + obrat + loanprc + unem + male +
    married + dep + sch + cosign + pubrec + mortlat1 + mortlat2 +
    vr + white, family = binomial(link = "logit"), data = my_loanapp)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.52976    0.57986   7.812 5.63e-15 ***
hrat         0.01852    0.01261   1.468   0.1420
obrat       -0.05715    0.01103  -5.181 2.21e-07 ***
loanprc     -1.89846    0.45981  -4.129 3.65e-05 ***
unem        -0.05610    0.03235  -1.734   0.0829 .
male        -0.06646    0.20289  -0.328   0.7432
married      0.42606    0.17465   2.439   0.0147 *
dep         -0.06529    0.07202  -0.906   0.3647
sch          0.07918    0.17546   0.451   0.6518
cosign       0.09046    0.43311   0.209   0.8345
pubrec      -1.66816    0.20667  -8.072 6.94e-16 ***
mortlat1    -0.41786    0.46528  -0.898   0.3691
mortlat2    -1.07987    0.54908  -1.967   0.0492 *
vr          -0.33328    0.15167  -2.197   0.0280 *
white        1.03621    0.16878   6.139 8.29e-10 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1476.0  on 1970  degrees of freedom
Residual deviance: 1237.3  on 1956  degrees of freedom
  (18 observations deleted due to missingness)
AIC: 1267.3

Number of Fisher Scoring iterations: 5

> logit.logLik=logLik(logit)
> logit.logLik
'log Lik.' -618.6517 (df=15)
```