The Latent Variable Model in Binary Regressions

Richard Williams, University of Notre Dame, https://www3.nd.edu/~rwilliam/

Last revised February 20, 2022

1. As Long & Freese note, there are at least two ways of motivating the logistic (and other binary) regression models.

- The first approach (which is what we have tended to emphasize so far) is the *Nonlinear Probability Model*. The independent variables have linear effects on the Log Odds of the event occurring, and the Log Odds in turn have a nonlinear relationship with the probability of the event.
- A second approach is known as the *latent variable model*. The idea is that there is a latent, unobserved variable y*, e.g. the extent to which one favors the job the President is doing. Once people cross a threshold on y*, the observed binary variable y switches from 0 to 1, e.g. the respondent switches from saying disapprove to approve.
- The mathematics are the same either way, but some problems can be more easily conceptualized using one approach or the other.

2. The latent variable model in binary regressions can be written as

$$y^* = \alpha + \sum X\beta + \varepsilon_{y^*}$$

If y* >= 0, y = 1

If y* < 0, y = 0

In logistic regression, $\varepsilon_{y^*} \sim \text{Standard Logistic Distribution}$. A *standard logistic distribution* has a mean of 0 and a variance of $\pi^2/3$, or about 3.29. It is very similar to a N(0, $\pi^2/3$) distribution. A standard logistic distribution has nice mathematical properties (e.g. it makes it easy to compute the odds and predicted probabilities) but we could just as easily use a standardized logistic distribution with mean 0 and variance 1; or alternatively, we could set the variance of y* to 1 (which, as we will see, can be very useful). Such changes will affect the scaling of parameters but not the predicted probabilities.

3. If two variables A & B are independent, then V(A + B) = V(A) + V(B). Since the residual term is uncorrelated with the X variables in the equation, it follows that

$$V(y^*) = V(\alpha + \sum X\beta) + V(\varepsilon_{y^*}) = V\left(\alpha + \sum X\beta\right) + \pi^2/3 = V\left(\alpha + \sum X\beta\right) + 3.29$$

The last equalities follow from the fact that the variance of the residual is $\pi^2/3$, or about 3.29. In the sample, the estimated variance of y* is

$$V(y^*) = V(a + \sum Xb) + V(\varepsilon_{y^*}) = V(Z_i) + 3.29$$

4. The variance of y* and ε_{Y^*} (which is always 3.29) are reported by Long & Freese's `fitstat` command, which is part of the `spost13` set of routines.

```
. use https://www3.nd.edu/~rwilliam/statafiles/glm-logit.dta, clear
. quietly logit grade gpa tuce i.psi
. fitstat

                         |       logit
-------------------------+-------------
Log-likelihood           |
                   Model |     -12.890
          Intercept-only |     -20.592
-------------------------+-------------
Chi-square               |
        Deviance (df=28) |      25.779
               LR (df=3) |      15.404
                 p-value |       0.002
-------------------------+-------------
R2                       |
                McFadden |       0.374
     McFadden (adjusted) |       0.180
      McKelvey & Zavoina |       0.544
            Cox-Snell/ML |       0.382
  Cragg-Uhler/Nagelkerke |       0.528
                   Efron |       0.426
                Tjur's D |       0.429
                   Count |       0.813
        Count (adjusted) |       0.455
-------------------------+-------------
IC                       |
                     AIC |      33.779
        AIC divided by N |       1.056
               BIC (df=4) |      39.642
-------------------------+-------------
Variance of              |
                       e |       3.290
                  y-star |       7.210
```

According to `fitstat`, V(y*) = 7.210, V(error) = 3.29, implying explained variance = 7.21 – 3.29 = 3.92. To confirm that `fitstat` got it right, use `predict` to compute the logit (aka predicted value) for each case and then see what the variance is.

```
. predict yhat if e(sample), xb
. sum yhat

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        yhat |         32   -1.083627     1.97985  -3.685518   2.850418

. display r(Var)
3.9198062
```

6. To sum up:

- In logistic regression, the variance of the residual is typically fixed at 3.29. You need some way to fix the scaling of a latent variable and this approach has several nice mathematical properties, e.g. it is easy to compute odds and probabilities when you do this. However, there are other ways to fix the scale of y*, with the most typical/useful being that you fix V(y*) at 1. The method you use will affect the scaling of the coefficients but not the predicted probabilities.
- The explained variance is the variance of the predicted values. The estimated variance of y* is the sum of the explained and residual variances.
- Probit is similar, except the residuals have a N(0, 1) distribution. Other link functions can also be used.
- The latent variable model for binary regressions is easily extended to many ordinal regression models. Instead of just having a single dividing point, you have multiple cut points/thresholds.