

The Latent Variable Model in Binary Regressions

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

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

$$\text{If } y^* \geq 0, y = 1$$

$$\text{If } y^* < 0, y = 0$$

In logistic regression, $\varepsilon_{y^*} \sim$ 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. Because y^* is unobserved, we have to do something to determine its scaling; 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

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

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 `spost9` set of routines.

```
. use http://www.nd.edu/~rwilliam/xsoc73994/statafiles/glm-logit.dta, clear
. quietly logit grade gpa tuce psi
```

```
. fitstat
```

Measures of Fit for logit of grade

```
Log-Lik Intercept Only:      -20.592   Log-Lik Full Model:      -12.890
D(28):                       25.779   LR(3):                   15.404
                               Prob > LR:                   0.002
McFadden's R2:               0.374   McFadden's Adj R2:       0.180
ML (Cox-Snell) R2:           0.382   Cragg-Uhler(Nagelkerke) R2: 0.528
McKelvey & Zavoina's R2:    0.544   Efron's R2:              0.426
Variance of y*:           7.210 Variance of error:    3.290
Count R2:                    0.813   Adj Count R2:            0.455
AIC:                          1.056   AIC*n:                   33.779
BIC:                          -71.261  BIC':                     -5.007
BIC used by Stata:           39.642   AIC used by Stata:       33.779
```

According to fitstat, $V(y^*) = 7.210$, $V(\text{error}) = 3.29$, implying explained variance = $7.210 - 3.29 = 3.92$. 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 1.97985^2
3.919806
```

5. Suppose that $Z_i = 2$. Since $y^* = Z_i + \varepsilon_{y^*}$, ε_{y^*} would have to be -2 or smaller in order for y^* to fall below the threshold of 0 . Using the formula for the CDF of the standard logistic distribution (Long 1997, p. 42) the probability of this happening is

$$P(\varepsilon_{y^*} \leq -2) = \frac{\exp(-2)}{1 + \exp(-2)} = \frac{.135335}{1.135335} = .1192$$

i.e. there is only about a 12% chance that a person with an expected value of 2 (which is above the 0 threshold) actually falls below the threshold. There is an 88% chance that they fall above it.

Even though it is the residuals that have the logistic distribution, it is equivalent to say $\text{invlogit}(Z_i) = P(Y=1)$ and $1 - \text{invlogit}(Z_i) = \text{invlogit}(-Z_i) = P(Y=0)$.

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 (e.g. log-log, complementary log-log, Cauchit) can also be used.
- As we will see, the latent variable model for binary regressions can easily be extended to many ordinal regression models.