

The Latent Variable Model in Binary Regressions

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

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

If $y^* \geq 0$, $y = 1$

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. 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.

NOTE: Probit is similar, except the residuals have a $N(0, 1)$ distribution. Other link functions (e.g. log-log, complementary log-log, Cauchit) are also sometimes used. These allow for different distributions for the residuals. For example, the Cauchit distribution has tails that are bigger than the normal distribution's, hence the Cauchit link may be useful when you have more extreme values in either direction. See the help files for `oglm` or `gologit2` for more information.

2. $V(y^*) = V(E[y^*]) + V(\varepsilon_{y^*}) = \text{Explained Variance} + \text{Residual Variance}$

In the sample,

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

3. Remember, it is the residuals that have a standard logistic distribution with mean 0 and variance $\pi^2/3$. So, 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

$$\hat{P}(-2) = \frac{\exp(-2)}{1 + \exp(-2)} = \frac{.135335}{1.135335} = .1192$$

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 $1 - \text{invlogit}(Z_i) = \text{invlogit}(-Z_i) = P(Y=0)$.

4. The variance of y^* and ε_{y^*} (which is always 3.29) are reported by the `fitstat` command.

```
. 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

5. 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
```

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.