Lecture 9: Probit and Complementary Log-log Models for Binary Response Data

Lin Chen

Department of Public Health Sciences
The University of Chicago

Part I: Probit Model
Motivating Data for Probit Model: *Biological Assay*

- Biological assay or *Bioassay* is a general type of scientific experiment investigating effects of agents on biological systems
- In one common type of bioassay, different concentrations of a chemical compound, drug, etc. are applied to batches of biological experimental entities (animals, tissues, cells...). The number responding at each dose is then recorded as a response variable.
- This type of experiment typically produces proportions as response data. For example,
  - Number of experimental animals in batch (dose group) i: n_i
  - Number of animals respond (0/1) in batch i: y_i
  - Concentration of chemical compound in batch i: d_i

Tolerance – a latent (unobserved) variable

- Consider the *Toxicity of cypermethrin to moths* example.
  Low dose⟶ few die; high dose ⟶ many die.

Table 1: Mortality of tobacco budworm moths 72 hours after exposure to cypermethrin.

Sex of moth	Dose of cypermethrin	Number affected out of 20
Male	1.0	1
	2.0	4
	4.0	9
	8.0	13
	16.0	18
	32.0	20
Female	1.0	0
	2.0	2
	4.0	6
	8.0	10
	16.0	12
	32.0	16

Tolerance (continued)

- Which ones die? Those with low *tolerance* than dose d_i for cypermethrin will.
- Idea: There exists a latent (unobserved) tolerance U, which has a distribution $f(U)$. Those with tolerance higher than d_i would survive. The binary outcome (death/not) of each moth i depends on whether its latent tolerance (u_i) exceeds the current dose (d_i).

not able to be directly measured

low tolerance, high dose

$$Y = \begin{cases} 1, & \text{if } U_i \le d_i \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

Then the probability of death when exposed to dose d_i is

$$p_i = P(U \le d_i) = \int_{-\infty}^{d_i} f(u)\, du \tag{2}$$

Now consider U as a random variable that we are interested in.

Tolerance: Logistic Distribution

If U follows a *logistic distribution*:

$$f(u) = \frac{\exp\{(u-\mu)/\tau\}}{\tau[1+\exp\{(u-\mu)/\tau\}]^2}, \quad -\infty < u < \infty \tag{3}$$

where $-\infty < \mu < \infty$, and $\tau > 0$. $\mathrm{E}(U) = \mu$, $\mathrm{Var}(U) = \pi^2\tau^2/3$.

$$p_i = \int_{-\infty}^{d_i} f(u)\,du = \int_{-\infty}^{d_i} \frac{\exp\{(u-\mu)/\tau\}}{\tau[1+\exp\{(u-\mu)/\tau\}]^2}\,du = \frac{\exp\{(d_i-\mu)/\tau\}}{1+\exp\{(d_i-\mu)/\tau\}} \tag{4}$$

Let $\beta_0 = -\mu/\tau$ and $\beta_1 = 1/\tau$, we can rewrite Eqn. (3) as:

$$p_i = \frac{\exp(\beta_0+\beta_1 d_i)}{1+\exp(\beta_0+\beta_1 d_i)} \tag{5}$$

This yields the logistic regression model that we are familiar with.

$$\mathrm{logit}(p_i) = \beta_0 + \beta_1 d_i. \tag{6}$$

Note the logit function is the inverse function of logistic function (see Lecture 3, Slide 8).

Tolerance: Normal (Gaussian) Distribution

More naturally, we may consider the tolerance variable U as a continuous variable following a *normal distribution*:

$$f(u) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2}(\frac{u-\mu}{\sigma})^2\}, \quad -\infty < u < \infty \tag{7}$$

Then,

$$p_i = \int_{-\infty}^{d_i} f(u)\, du = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{d_i} \exp\{-\frac{1}{2}(\frac{u-\mu}{\sigma})^2\}\, du = \Phi\left(\frac{d_i-\mu}{\sigma}\right) \tag{8}$$

Let $\beta_0 = -\mu/\sigma$ and $\beta_1 = 1/\sigma$, we can rewrite the above Eqn. as

$$p_i = \Phi(\beta_0 + \beta_1 d_i). \tag{9}$$

This yields the probit regression model,

$$\Phi^{-1}(p_i) = \beta_0 + \beta_1 d_i. \tag{10}$$

The probability of response is linked with a linear combination of predictors via a probit function – the inverse of the cumulative distribution function of the standard normal distribution.

Tolerance - Expressed as Effective Dose

More specifically, we may be interested in

- The dose that is expected to result in a $(100 \times p)\%$ response, for example for $p = 0.50$, 50% of subjects respond
- This quantity ED can be derived from the model. Which dose level can give a $p_i = 50\%$? This can be computed based on Eqn. (9).
- Based on the probit model, the effective dose ED estimate for a response probability of p is given by

$$\widehat{ED} = \frac{1}{\hat{\beta}_1}(\Phi^{-1}(p) - \hat{\beta}_0) \tag{11}$$

- The effective dose estimate for logit model is

$$\widehat{ED} = \frac{1}{\hat{\beta}_1}(\log \frac{p}{1-p} - \hat{\beta}_0) \tag{12}$$

- ED_{50}, ED_{90} are often of interest

Example: *Anti-pneumococcus serum*

Table 2: Number of deaths from pneumonia amongst batches of 40 mice exposed to different doses of a serum

Dose of serum	Number of deaths out of 40
0.0028	35
0.0056	21
0.0112	9
0.0225	6
0.0450	1

Anti-pneumococcus serum: $\text{logit}(p_i) = \beta_0 + \beta_1 \log(d_i)$

First, we fit the logistic model with which we are familiar.

```
. gen ldose=log(dose)
. glm y ldose, family (binomial n) nolog

Generalized linear models                         Number of obs   =          5
Optimization     : ML                             Residual df     =          3
                                                  Scale parameter =          1
Deviance         =  2.80889589                    (1/df) Deviance =   .9362986
Pearson          =  2.917404799                   (1/df) Pearson  =   .9724683

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(u/(n-u))             [Logit]

                                                  AIC             =   4.715478
Log likelihood   = -9.788693777                   BIC             =  -2.019418

------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldose |  -1.829621   .2545469    -7.19   0.000    -2.328524   -1.330719
       _cons |  -9.189392    1.25511    -7.32   0.000    -11.64936   -6.729422
------------------------------------------------------------------------------

. predict yhat_logitmod
(option mu assumed; predicted mean count)
```

Anti-pneumococcus serum: $\text{probit}(p_i) = \beta_0 + \beta_1 \log(d_i)$

Now we fit the probit model by specifying the link function "`link(probit)`". Deviance is still good for evaluating goodness-of-fit.

```
. glm y ldose, family(binomial n) link(probit) nolog
```

default: link(logit)

```
Generalized linear models                         Number of obs   =          5
Optimization     : ML                             Residual df     =          3
                                                  Scale parameter =          1
Deviance         =  3.193092195                   (1/df) Deviance =   1.064364
Pearson          =  3.197015026                   (1/df) Pearson  =   1.065672

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = invnorm(u/n)            [Probit]

                                                  AIC             =   4.792317
Log likelihood   = -9.98079193                    BIC             =  -1.635222

------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldose |  -1.054488   .1328879    -7.94   0.000    -1.314943   -.7940325
       _cons |  -5.276716   .6441273    -8.19   0.000    -6.539183    -4.01425
------------------------------------------------------------------------------
```

Grouped Binary Data GOF ✓

```
. predict yhat_probitmod
(option mu assumed; predicted mean count)
```

Anti-pneumococcus serum: Effective dose

Based on Slide 7, we calculate the effective dose for 50 and 90% response: Probit Model

```
. di (invnormal(0.5)-(-5.276716))/(-1.054488)
-5.004055
. di exp(-5.004055)
.00671068

. di (invnormal(0.9)-(-5.276716))/(-1.054488)
-6.2193857
. di exp(-6.2193857)
.00199047

. gen phat_probit=yhat_probitmod/n
. list y n dose phat_probit
     +----------------------------+
     |  y    n    dose   phat_p~t |
     |----------------------------|
  1. | 35   40   .0028   .8216594 |
  2. | 21   40   .0056   .5756557 |
  3. |  9   40   .0112    .294556 |
  4. |  6   40   .0225   .1010245 |
  5. |  1   40    .045   .0223934 |
     +----------------------------+
```

The effective doses are $ED_{50} = 0.0067$ and $ED_{90} = 0.0020$ based on the log dose model with probit link.

Anti-pneumococcus serum: $\text{probit}(p_i) = \beta_0 + \beta_1 \log(d_i)$

This model still has a binomial outcome even though the link function is a function of the normal distribution

- The link function in a generalized linear model (GLM) answers "How is what is predicted by the linear function related to the mean of y (i.e., response variable)?"
  - Ordinary linear regression has the *identity* link, as the linear predictor directly predicts the mean of y for a given x
  - For binomial outcomes, the default link is the logit or log(p/1-p). Note that indeed p is the mean of y (mean of a sum of 0/1 responses is the proportion responding)
  - Probit regression is another GLM regression for **binary/binomial outcome.** The *probit* link indicates that the linear predictor predicts the probit transform of the mean p. The probit for p is the value from the (standard) normal distribution that is associated with the probability p.

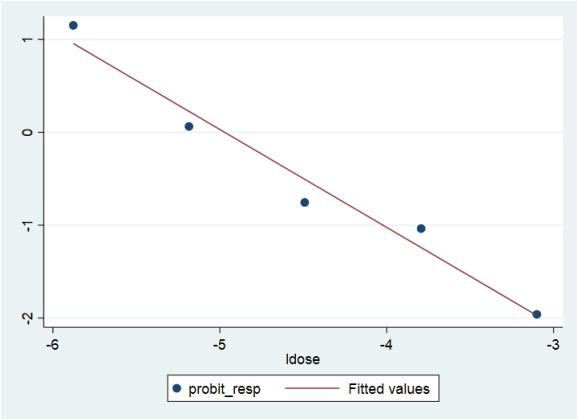
Anti-pneumococcus serum: $\text{probit}(p_i) = \beta_0 + \beta_1 \log(d_i)$

Can we just do the transform ourself and fit the model via ordinary regression. In this instance, we can.

```
. gen probdie=y/n
. gen probit_resp=invnormal(probdie)
. list dose y n probdie ldose probit_resp
```

	dose	y	n	probdie	ldose	probit_~p
1.	.0028	35	40	.875	−5.878136	1.150349
2.	.0056	21	40	.525	−5.184988	.0627067
3.	.0112	9	40	.225	−4.491841	−.755415
4.	.0225	6	40	.15	−3.79424	−1.036433
5.	.045	1	40	.025	−3.101093	−1.959964

```
. twoway scatter probit_resp ldose
```



Anti-pneumococcus serum: $\text{probit}(p_i) = \beta_0 + \beta_1 \log(d_i)$

Here we will run a linear regression (OLS)

```
. regress probit_resp ldose
```

Source	SS	df	MS	Number of obs	=	5
				F(1, 3)	=	95.29
Model	5.3558733	1	5.3558733	Prob > F	=	0.0023
Residual	.168610761	3	.056203587	R-squared	=	0.9695
				Adj R-squared	=	0.9593
Total	5.52448406	4	1.38112101	Root MSE	=	.23707

probit_resp	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ldose	-1.053787	.1079493	-9.76	0.002	-1.39733	-.7102445
_cons	-5.239319	.4961589	-10.56	0.002	-6.818319	-3.66032

Is it surprising? Estimates are very similar (not identical - least squares (LS) estimation used here). This method was widely used before availability of modern computing, because LS estimation can be carried out even without (good) computers.

Part II: The complementary log-log model – Tolerance follows a Gumbel distribution

If the tolerance variable U follows a *Gumbel distribution*:

$$f(u) = \frac{1}{\kappa} e^{(u-\alpha)/\kappa} \exp\{-e^{(u-\alpha)/\kappa}\}, \quad -\infty < u < \infty \tag{13}$$

Where $-\infty < \alpha < \infty$ and $\kappa > 0$ are unknown parameters.

$$p_i = \int_{-\infty}^{d_i} f(u)\, du = \int_{-\infty}^{d_i} \frac{1}{\kappa} e^{(u-\alpha)/\kappa} \exp\{-e^{(u-\alpha)/\kappa}\}\, du = 1 - \exp\{-e^{(d_i-\alpha)/\kappa}\} \tag{14}$$

Let $\beta_0 = -\alpha/\kappa$ and $\beta_1 = 1/\kappa$, and we can rewrite Eqn (12) as

$$p_i = 1 - \exp\left\{-\exp(\beta_0 + \beta_1 d_i)\right\}, \quad \text{or} \tag{15}$$

$$\log\{-\log(1-p_i)\} = \text{cloglog}(p_i) = \beta_0 + \beta_1 d_i \tag{16}$$

This yields the complementary log-log regression model. The cloglog function links the probability with the linear combination of predictors.

The effective dose estimate is $\widehat{ED} = \frac{1}{\hat{\beta}_1}\{\log[-\log(1-p)] - \hat{\beta}_0\}$. (17)

Anti-pneumococcus serum: $\text{cloglog}(p_i) = \beta_0 + \beta_1 \log(d_i)$

Run a cloglog model by specifying the link function "`link(cloglog)`"

```
. glm y ldose, family (binomial n) link(cloglog) nolog

Generalized linear models                         Number of obs   =          5
Optimization     : ML                             Residual df     =          3
                                                  Scale parameter =          1
Deviance         =  1.309885583                   (1/df) Deviance =   .4366285
Pearson          =  1.361478279                   (1/df) Pearson  =   .4538261

Variance function: V(u) = u*(1-u/n)               [Binomial]
Link function    : g(u) = ln(-ln(1-u/n))          [Complementary log-log]

                                                  AIC             =   4.415675
Log likelihood   = -9.039188624                   BIC             =  -3.518428
------------------------------------------------------------------------------
             |                 OIM
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ldose |  -1.397982   .1776859    -7.87   0.000    -1.74624   -1.049724
       _cons |  -7.511787   .9324851    -8.06   0.000   -9.339424    -5.68415
------------------------------------------------------------------------------
. predict yhat_cloglogmod
(option mu assumed; predicted mean y)
```

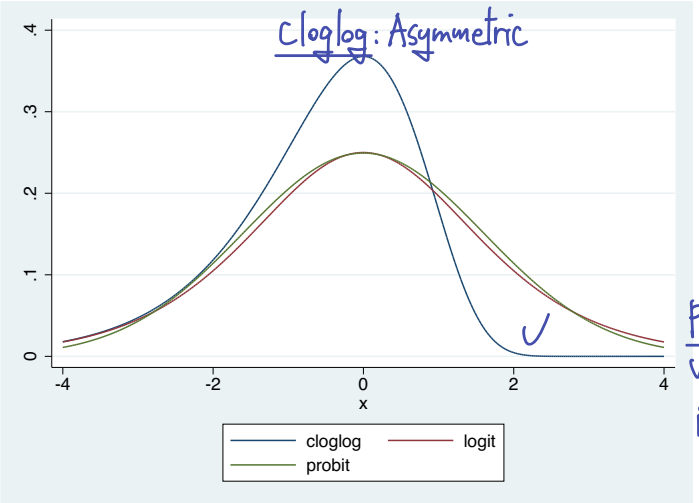
Grouped Binary Data GOF ✓

Deviance can still be used to evaluate the goodness-of-fit of the model.

Tolerance distributions comparison

This figure shows the the tolerance density function for the complementary log-log distribution $(0,1)$, the logistic $(0,1)$, and $\mathrm{N}(0,1.6^2)$. The normal distribution is chosen to have the same mean and variance as the standard logistic distribution.

```
. twoway function cloglog = exp(x)*exp(-exp(x)), range(-4 4)
    || function logit = exp(x)/(1+exp(x))^2, range(-4 4)
    || function probit = 0.39894*exp(-0.5*(x/1.6)^2)/1.6, range(-4 4)
```



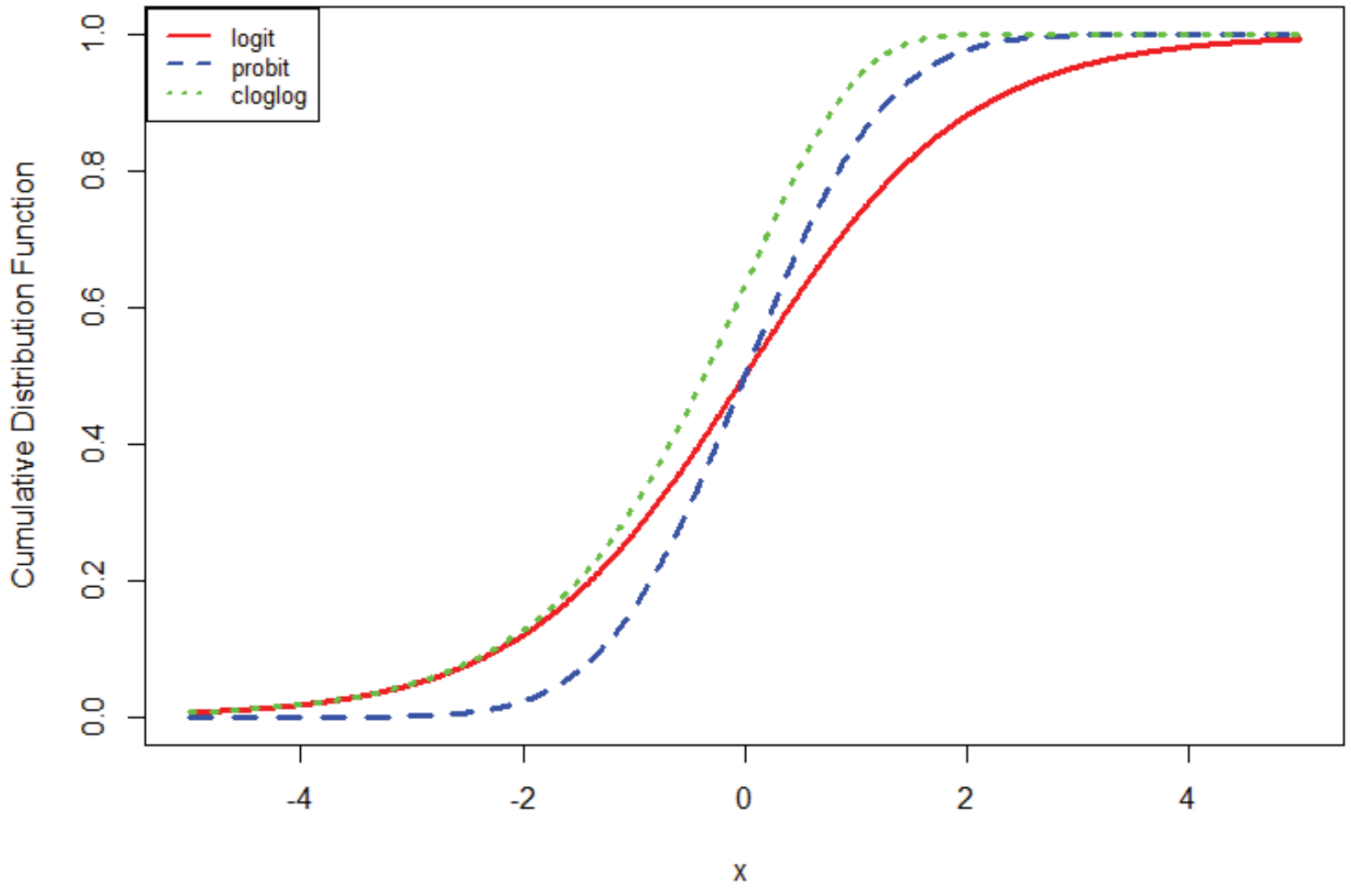
Probit/Logit very similar... usually use interchangeably for rough calculations.

Generally, logit and probit models perform similarly. The normal distribution and the logistic distribution (for the latent tolerance variable) have similar shapes, but the logistic distribution has slightly longer/heavier tails.

Unlike logit/probit, the cloglog function is asymmetric, and may be used when the probability of response is very small or large. Depending on the scentific problems, different tolerance distributions might serve better as the theoretical model.

Comparing Transforms of p

Recall from Lecture 3 and Eqn (4), (8) and (14) in this lecture.



The cloglog function (the green dotted line) approaches 1 sharply but zero slowly and can be used to model rare event.

Examining the Three Models

* Interpretation is more important *

Table 3: **Model Comparison using Deviance.**

Model	Deviance	Deviance d.f.	P-value	$\hat{ED}_{50}$
logit	2.809	1	0.42	.006588
probit	3.193	1	0.36	.006711
cloglog	1.310 ↓	1	0.73	.006029

Harder to interpret

no huge difference

Due to sparse data, but we will not sacrifice interpretation

simply a better fit of cloglog.

* May still choose logit for better interpretation....

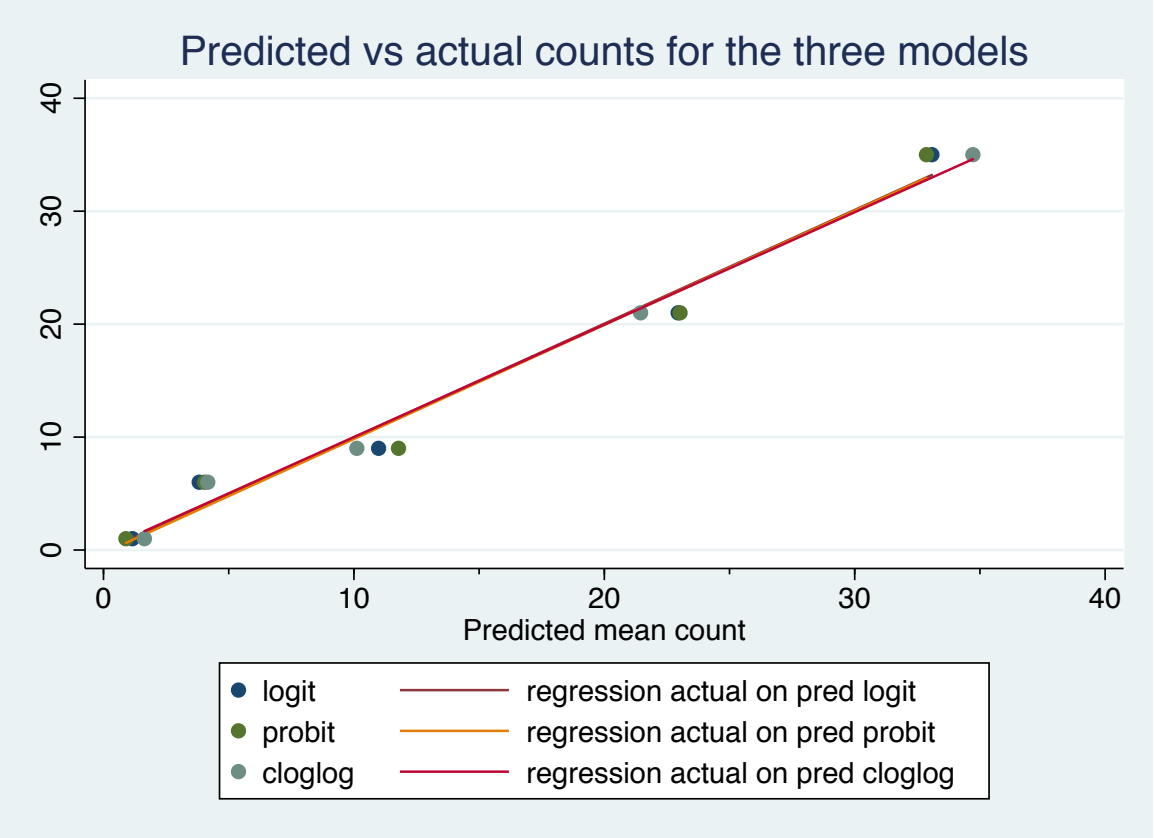
Depends on topic of interest

Predicted counts:

```
. list ldose y n yhat_logitmod yhat_probitmod yhat_cloglogmod
```

	ldose	y	n	yhat_l~d	yhat_p~d	yhat_c~d
1.	−5.878136	35	40	33.08491	32.86637	34.72205
2.	−5.184988	21	40	22.95006	23.02623	21.45233
3.	−4.491841	9	40	10.98707	11.78224	10.11815
4.	−3.79424	6	40	3.823065	4.040979	4.16571
5.	−3.101093	1	40	1.154902	.8957366	1.634881

Examining the Three Models



There is no clear winner - complementary log-log may be a bit better overall. ED_{50} estimate appears closer to the empirical data (50% affected just exceeded for dose .0056)

Summary

Logit, Probit, and Cloglog Models glm for Binary

- For binary outcome data, probit and cloglog are useful alternative models to the logistic regression.
- The probit model has a long history and may have performance similar to logit. It is often used when the binary response (0/1) is whether a latent variable exceed a threshold, and is also known as *the latent variable model*. For example, the tolerance in Bioassay data.
- The Complimentary Log-Log (cloglog) function is asymmetric. It is often used when the probability of an event is very small (or large). sparse
- Deviance can still be used for assessing goodness-of-fit and for model comparisons (formally for nested models of the same link and informally for non-nested models).
- Logit is the most commonly used model for binary outcome data and logit function is the default link option in statistical software.