

Binary Data and Logit Models

This will be our first look at survival and reproduction in fruitflies.

- Survival to 50 days for a male fruitfly
- Treatments:
 - No females
 - One virgin female
 - Eight virgin females
 - One newly pregnant female
 - Eight newly pregnant females
- 25 fruitflies on each of the five treatments.

Data

Num. Females	Pregnant		Virgin		
	1	8	1	8	0
Num < 50	6	6	9	20	7
Num \geq 50	19	19	16	5	18

Linear Predictor

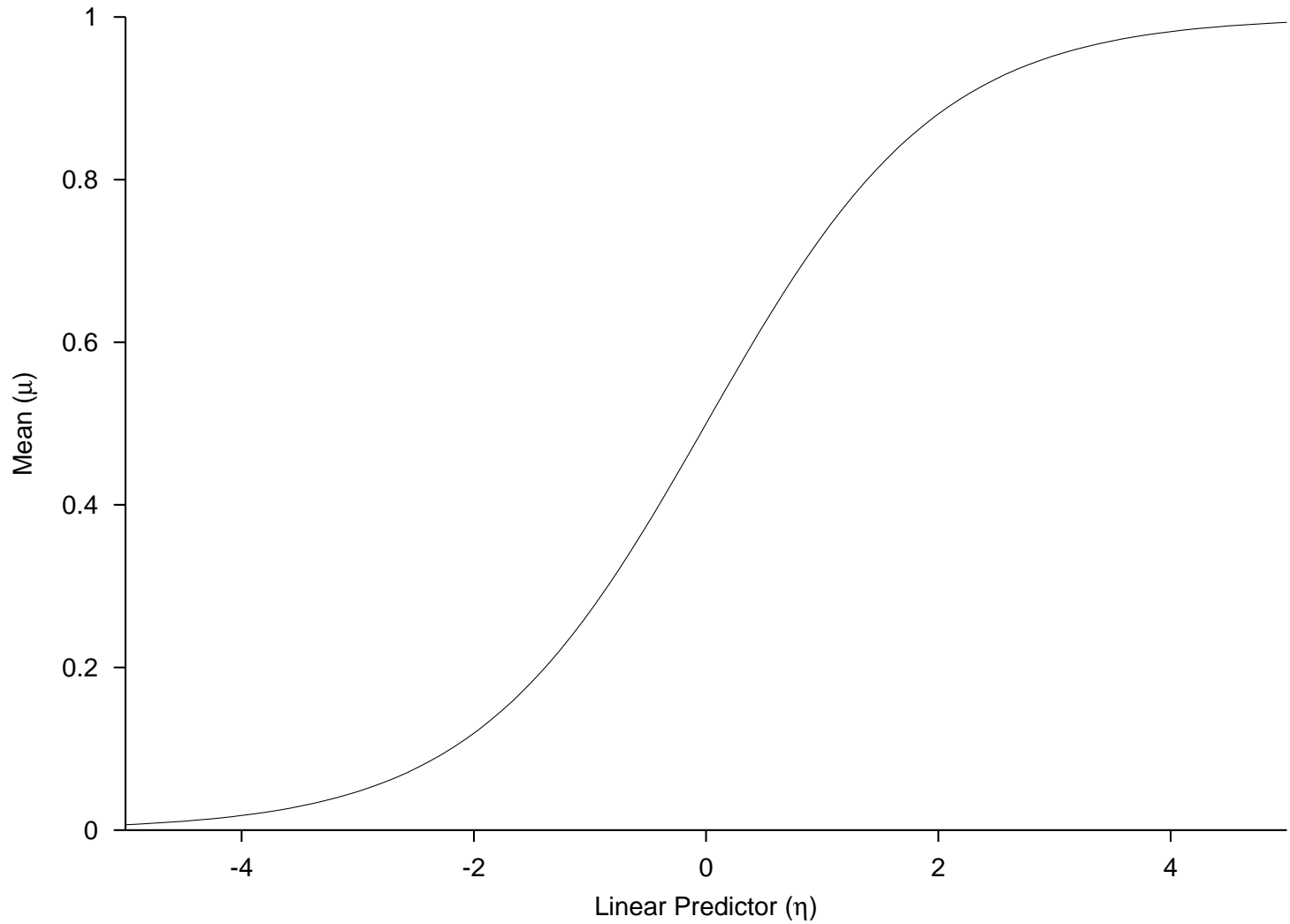
$$\eta_{ijk} = \mu + P_i + V_j + P_iV_j$$

- μ intercept
- P_i main effect for number of partners
- V_j main effect for pregnancy status
- P_iV_j interaction

With a binomial distribution, the common link functions are the logit and the probit.

The inverse link function for the logit is

$$\mu_{ij} = \frac{e^{\eta_{ij}}}{1 + e^{\eta_{ij}}}$$



Estimation

The estimating equations for a GLM is

$$[\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X}] \boldsymbol{\beta}^{[i+1]} = \mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[i]} + \mathbf{H}\boldsymbol{\eta}^{[i]}).$$

We will therefore need an initial estimate for our linear predictor

$$\eta_{ij}^{[0]} = 0.$$

The residual variance is obtained from the specified distribution of \mathbf{y}

$$\mathbf{R} = \text{Diag} \left(\frac{\mu_{ij}(1 - \mu_{ij})}{n_{ij}} \right)$$

$$\mathbf{R}^{-1} = \text{Diag} \left(n_{ij} \frac{1}{\mu_{ij}(1 - \mu_{ij})} \right).$$

Next we need the partial derivative of mean with respect to the linear predictor

$$\mathbf{H} = \text{Diag} (\mu_{ij}(1 - \mu_{ij})).$$

Putting the pieces together we obtain

$$\mathbf{H}' \mathbf{R}^{-1} \mathbf{H} = \text{Diag}(n_{ij} \mu_{ij}(1 - \mu_{ij}))$$

$$\mathbf{H}' \mathbf{R}^{-1} (\mathbf{y} - \boldsymbol{\mu}^{[it]} + \mathbf{H} \boldsymbol{\eta}^{[it]}) = \left\{ c n_{ij} (y_{ij} - \mu_{ij} + \mu_{ij}(1 - \mu_{ij}) \eta_{ij}^{[it]}) \right\}.$$

For the first iterate

$$\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} = \text{Diag}(25 \times .5(1 - .5) = 6.25)$$

$$\mathbf{H}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[0]} + \mathbf{H}\boldsymbol{\eta}^{[0]}) = \begin{pmatrix} 25(.76 - .5 + .25 \times 0) = 6.5 \\ 25(.76 - .5 + .25 \times 0) = 6.5 \\ 25(.64 - .5 + .25 \times 0) = 3.5 \\ 25(.2 - .5 + .25 \times 0) = -7.5 \\ 25(.72 - .5 + .25 \times 0) = 5.5 \end{pmatrix}$$

which yields

$$\hat{\beta}^{[1]} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1.04 \\ 1.04 \\ .56 \\ -1.2 \\ .88 \end{pmatrix}$$

$$\eta^{[1]} = \begin{pmatrix} 0 + \dots + 1.04 = 1.04 \\ 0 + \dots + 1.04 = 1.04 \\ 0 + \dots + .56 = .56 \\ 0 + \dots - 1.2 = -1.2 \\ 0 + \dots + .88 = .88 \end{pmatrix}$$

$$\mu^{[1]} = \begin{pmatrix} \frac{\exp(1.04)}{1+\exp(1.04)} = 0.7388 \\ \frac{\exp(1.04)}{1+\exp(1.04)} = 0.7388 \\ \frac{\exp(.56)}{1+\exp(.56)} = 0.6364 \\ \frac{\exp(-1.2)}{1+\exp(-1.2)} = 0.2315 \\ \frac{\exp(.88)}{1+\exp(.88)} = 0.70688 \end{pmatrix}$$

For the second iterate we use the updated linear predictor from the first iterate:

$$\mathbf{H}'\mathbf{R}^{-1}\mathbf{H} = \begin{pmatrix} 25 \times 0.7388(1 - 0.7388) = 4.8244 & 0 & 0 & & \\ 0 & 4.8244 & 0 & & \\ 0 & 0 & 5.7849 & & \\ & & & \dots & \end{pmatrix}$$

$$\mathbf{H}'\mathbf{R}^{-1}(\mathbf{y} - \boldsymbol{\mu}^{[1]} + \mathbf{H}\boldsymbol{\eta}^{[1]}) = \begin{pmatrix} 25(.76 - .7388 + .1930 \times 1.04) = 5.548 \\ 25(.76 - .7388 + .1930 \times 1.04) = 5.548 \\ 25(.64 - .6364 + .2314 \times .56) = 3.330 \\ \vdots \end{pmatrix}$$

ASREML

Logit Analysis

Type * !A

Partners !I

Day50

Num

ff.dat !SKIP=1

Day50 !BINOMIAL !LOGIT !TOTAL Num ~ mu Type*Partners

0 0 0

predict Type Partners !TDIFF

Using 10 records of 10 read

Model term	Size	#miss	#zero	MinNon0	Mean	MaxNon0	
1 Type	3	0	0	1	1.8000	3	
2 Partners	3	0	0	1	1.8000	3	
3 Day50	Variate	0	5	1.000	0.5000	1.000	
4 Num	Weight	0	0	5.000	12.50	20.00	
5 mu	1						
6 Type.Partners	9	1 Type	:	3	2 Partners	:	3

Forming 16 equations: 16 dense.

Initial updates will be shrunk by factor 0.010

Notice: Algebraic ANOVA Denominator DF calculation is not available
Numerical derivatives will be used.

Distribution and link: Binomial; Logit $\mu = P = 1 / (1 + \exp(-XB))$
 $V = \mu(1 - \mu) / N$

Warning: The LogL value is unsuitable for comparing GLM models

Notice: 11 singularities detected in design matrix.

1	LogL=-59.6161	S2= 1.0000	5 df	1.000
2	LogL=-62.5074	S2= 1.0000	5 df	1.000
3	LogL=-62.5077	S2= 1.0000	5 df	1.000
4	LogL=-62.5077	S2= 1.0000	5 df	1.000
5	LogL=-62.5077	S2= 1.0000	5 df	1.000
6	LogL=-62.5077	S2= 1.0000	5 df	1.000

Final parameter values 1.0000

Solution File:

Type.Partners	P.001		0.000	0.000
Type.Partners	P.008		0.000	0.000
Type.Partners	P.000		0.000	0.000
Type.Partners	V.001		0.000	0.000
Type.Partners	V.008		-1.962	0.9286
Type.Partners	V.000		0.000	0.000
Type.Partners	N.001		0.000	0.000
Type.Partners	N.008		0.000	0.000
Type.Partners	N.000		0.000	0.000
Partners		1	0.000	0.000
Partners		8	0.000	0.6623
Partners		0	0.000	0.000
Type	P		0.000	0.000
Type	V		-0.5773	0.6268
Type	N		-0.2082	0.6463
mu		1	1.153	0.4683

Predict File:

Ecode is E for Estimable, * for Not Estimable

----- 1 -----

Predicted values of Day50

The cells of the hypertable are calculated from all model terms constructed
solely from factors in the averaging and classify sets.

Warning: 4 non-estimable [aliased] cell(s) may be omitted from the table.

The Overall SED statistic includes non-estimable predictions.

Type	Partners	Logit_value	Stand_Error	Ecode	Retransformed_value	approx_SE
P	1	1.1527	0.4683	E	0.7600	0.0953
P	8	1.1527	0.4683	E	0.7600	0.0953
V	1	0.5754	0.4167	E	0.6400	0.1004
V	8	-1.3863	0.5000	E	0.2000	0.0683
N	0	0.9445	0.4454	E	0.7200	0.0978
SED: Standard Error of Difference: Min		0.6099	Mean	0.6509	Max	0.6851

Predicted values with t statistics

1.153				
1.153	0.00			
0.5754	-0.92	-0.92		
-1.386	-3.71	-3.71	-3.01	
0.9445	-0.32	-0.32	0.61	3.48

Interpretation

- The solutions are presented on the logit scale and need to be converted back to the observed scale.
- Least Square Means:

Type	Partners	Logit Scale	Obs. Scale
P	1	$\hat{\mu} + \hat{T}_p + \hat{P}_1 + \hat{TP}_{p1}$	$e^{1.1527} / (1 + e^{1.1527})$
		1.1527	0.76
P	8	1.1527	0.76
V	1	0.5754	0.64
V	8	-1.3863	0.20
N	0	0.9445	0.72

Estimable Functions

- Estimable functions play the same role as they did in linear model.
- For example, if $K'\beta$ is estimable then its MLE is unique.
- Similarly, if $K'\beta$ is not estimable then its MLE is not unique.
- In this model, none of the fixed effects are estimable.
- However, there are estimable functions of the fixed effects.

- For example, the difference in the LS-means are estimable.
- The asymptotic covariance is obtained using

$$\mathbf{K}'(\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{K}$$

- The asymptotic standard error and t-statistic

$$se = \sqrt{\mathbf{K}'(\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{K}}$$

$$t = \frac{\mathbf{K}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{K}'(\mathbf{X}'\mathbf{H}'\mathbf{R}^{-1}\mathbf{H}\mathbf{X})^{-1}\mathbf{K}}}$$

Hypothesis Testing

- Hypothesis testing is based on the likelihood function. The Wald statistic,

$$W = \hat{\beta}' \mathbf{K} (\mathbf{K}' i(\hat{\beta})^{-1} \mathbf{K})^{-1} \mathbf{K}' \hat{\beta},$$

is based on the score function and its asymptotic variance.

- The likelihood ratio statistic,

$$G = 2(\ell(\hat{\beta}) - \ell(\hat{\beta}_R)),$$

is based on the difference in the log likelihoods.

- LRT tests can be obtained using !AOD without a predict statement.