



# Advanced STATS and R

## GLM's

## DAGs and DOODLES



# The Exponential Family

- We have been looking at the linear model of the form:

$$E(Y_i) = \mu_i = x_i' \beta$$

Or  $E(Y) = X\beta$  where  $X$  is the design matrix and  $x_i'$  is the  $i$ th row of the design matrix  $X$

Where

$$Y_i \stackrel{\text{iid}}{\sim} N(\mu_i, \sigma^2)$$

This would be our typical multiple regression model.

# Modification

- We can modify the form of this model to make it a little more general.
- Distribution no longer simply Normal but Exponential family
- $g(\mu_i) = x_i' \beta$
- $g$  is called the link function

# Exponential Family

- Suppose that the distribution of  $Y$  depends on only one parameter  $\theta$
- Then we could define  $Y$  to be distributed within the exponential family of distributions if:

$$f(y|\theta) = s(y)t(\theta)e^{a(y)b(\theta)}$$

- Or equivalently

$$f(y|\theta) = \exp\left((a(y)b(\theta) + c(\theta) + d(y))\right)$$

Another form of the exponential family called the exponential dispersion family (EDF).

$$f(y|\theta, \phi) = \exp \left[ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right]$$

$\theta$  is called the canonical parameter – represents location

$\phi$  is the dispersion parameter and represents the scale

This will have some consequences

# Mean and Variance for EDF

$$E(Y) = \mu = b'(\theta)$$
$$V(Y) = b''(\theta)a(\phi)$$

$b''(\theta)$  describes how the variance of the response relates to the mean.

In the case of the normal  $b''(\theta) = 1$  and hence the variance is independent of the mean.

We shall do some more algebra later on the EDF

# Modeling the GLM - 3 parts

The three components:

- 1. A random response**  $Y$  that is a member of the exponential family (usually EDF)
- 2. A linear predictor**  $\eta = X\beta$
- 3. A link function**

$$g(\mu_i) = \sum_j \beta_j x_{ij}, i = 1, \dots, n$$

$$\mu_i = E(Y_i)$$

$$g(\mu_i) = \eta_i$$

# Canonical link

$$\eta = g(\mu) = \theta$$



# Making a doodle from code

The screenshot displays the OpenBUGS interface, which is split into two main panels. The left panel is a code editor titled 'chap07\_ex3\_wais (2)' containing the following R code for a LOGIT MODEL:

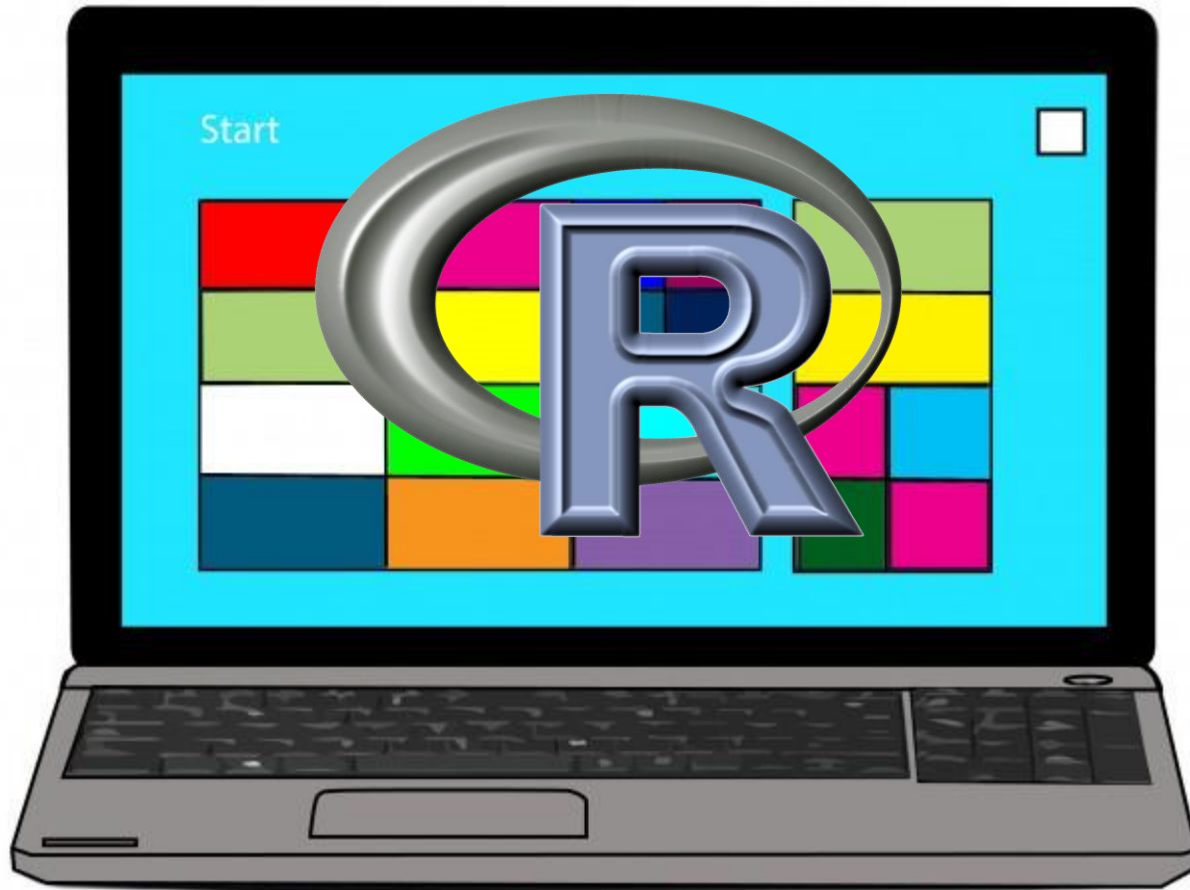
```
# ----- LOGIT MODEL -----  
model{  
  # Senility symptoms data - chapter 8  
  # binary regression example  
  # LOGIT MODEL  
  for (i in 1:n){  
    senility[i] ~ dbern( pi[i] )  
    logit( pi[i] ) <- beta0 + beta1 * wais[i]  
  }  
  # priors  
  beta0~dnorm( 0, 0.001)  
  beta1~dnorm( 0, 0.001)  
  #  
  odds0 <- exp(beta0)  
  OR <- exp(beta1)  
  #  
  # Wais for which pi=1/2  
  wais.half.prob <- - beta0/beta1  
  #  
  # probabilities for all X  
  for (k in 1:21){ logit( pi.model[k] ) <- beta0 + beta1 * (k-1) }  
}
```

The right panel is a doodle editor titled 'untitled1'. It features a table at the top with the following columns: 'name:', 'type:', 'stochastic', and 'density:'. The 'density:' column has a sub-column 'lower bound'. The table contains the following entries:

name:	type:	stochastic	density:
mean	0.0	precision	1.0E-6

Below the table, there are two empty ovals. A larger rectangular box contains a shaded oval and an empty oval. At the bottom of this box, the text 'for( IN : )' is visible, indicating a loop structure in the doodle.

**Bring your laptop or use Network.**



*The University of Oklahoma*

# Get the Data and Book

- Link below:
- [statsandr.oucreate.com](https://statsandr.oucreate.com)



# Courses

- **Bayesian Stats MATH 4803/5803**
- Advanced Applied STATS MATH 4793/5793 (Next Year)
- **Applied Statistical Methods MATH 4753**

