# Settings for a Binary-Binomial response model

- Sample of target population. Each individual is independent and classified by *success, failure*.

```
Response   Predictors for each individual
Y1=0            x11,x12,     ,x1p
Y2=0            x21,x22,     ,x2p
Y3=1            x31,x32,     ,x3p
```

- Individuals with same values for $X_1, X_2, \ldots, X_p$ are grouped.

```
Group   Response
n1         Y1
n2         Y2
n3         Y3
```

- $n_i$ number of individuals, $Y_i$ no. of successes for group $i$.

- Each group has an equal set of predictor variables. (dose-response experiment)
- For example, all individuals in group *i* have the same ”age”.
    - Few replicates of $X_1, X_2, \ldots, X_p$.
    - Data reported as $0 - 1$
- Case 1: *Binary non-replicated* data.
- Case 2: *Grouped data* model or *stratified Binomial* model.
    - Observations grouped at covariate levels
    - Group sizes $n_i$ much larger than 1.
- The MLEs of $\beta_0, \beta_1, \ldots, \beta_p$ do not depend on individuals begin grouped with covariates or not.

# Distributional results on Binary/Binomial models

- If sample size large ($n = n_1 + n_2 + \ldots + n_N$), then

$$\hat{\beta}_i \approx N(\beta_i, Var(\hat{\beta}_i))$$

- Typically $Var(\hat{\beta}_i)$ are obtained by software package.
- An approximate 95% confidence interval for $\beta_i$ is

$$\hat{\beta}_i \pm (1.96) SE(\hat{\beta}_i).$$

- To test $H_0 : \beta_i = 0$ vs $H_a : \beta_i \neq 0$,

$$Z = \frac{\hat{\beta}_i - 0}{SE(\hat{\beta}_i)}$$

  which is approximately a $N(0, 1)$.
- If $z_{obs}$ is the observed value of $Z$, p-value=$2P(Z > z_{obs})$.
- SAS reports $Z^2 \approx \chi^2_{(1)}$

- **Saturated model.** Model with the the maximum number of parameters that can be estimated.
- $Y_1, Y_2, \ldots, Y_N$ are independent and $Y_i \sim Binomial(n_i, \pi_i)$, *log-likelihood* is (except for constant),

$$l(\beta; Y) = \sum_{i=1}^{N} [y_i log(\pi_i) - y_i log(1 - \pi_i) + n_i log(1 - \pi_i)]$$

- Saturated model: All $\pi_i' s$ are different and $\beta = (\pi_1, \pi_2, \ldots, \pi_N)^T$
- The maximum likelihood estimates are $\hat{\pi}_i = y_i / n_i$ ($b_{max}$).
- The max. log-likelihood is

$$l(b_{max}; Y) = \sum_{i=1}^{N} [y_i log(y_i/n_i) - y_i log(1-(y_i/n_i)) + n_i log(1-(y_i/n_i))]$$

- For model with $p < N$ parameters, estimates $\hat{\pi}_i$.
- Fitted values, $\hat{y}_i = n_i \hat{\pi}_i$
- The log-likelihood,

$$l(b; Y) = \sum_{i=1}^{N} [y_i log(\hat{y}_i / n_i) - y_i log(1 - (\hat{y}_i / n_i)) + n_i log(1 - (\hat{y}_i / n_i))]$$

- The deviance (textbook)

$$\begin{aligned} D &= 2[l(b_{max}; Y) - l(b; Y)] \\ &= 2 \sum_{i=1}^{N} \left[ y_i log \left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) log \left( \frac{n_i - y_i}{n_i - \hat{y}_i} \right) \right] \end{aligned}$$

- $D's$ approximate sampling distribution is *chi-square*.

- **R** computes the *null deviance*. For all $i$

$$g(\pi_i) = \beta_0 \to \hat{\pi} = g^{-1}(\hat{\beta}_0)$$

- Fitted values $\hat{y}_i = n_i\hat{\pi}$
- **Null deviance:**

$$-2\sum_{i=1}^{N}\left[y_i log(\hat{\pi}) + (n_i - y_i) log(1-\hat{\pi}) + \binom{n_i}{y_i}\right]$$

with *degrees of freedom* $N-1$.

- **Residual deviance:**

$$g(\pi_i) = \beta_0 + \beta_1 X_{i1} + \ldots + \beta_p X_{pi}$$

- $\hat{\pi}_i$ are estimated with covariates and $\hat{y}_i = n_i\hat{\pi}_i$

- **Residual deviance:**

$$D = -2\sum_{i=1}^{N}\left[y_i log(\hat{\pi}_i) + (n_i - y_i) log(1-\hat{\pi}_i) + \binom{n_i}{y_i}\right]$$

with degrees of freedom $N - (p+1)$.

- *Hypothesis testing:* $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$ vs. $H_1$ : *at least one* $\beta_i \neq 0$.

$$D^* = Null\ deviance - Residual\ deviance$$

$$dof = (N-1) - (N - (p+1)) = p.$$

- $D^*$ approximately follows a $\chi^2_{(p)}$.

- Alternatively if we just focus on the **Residual deviance** (grouped Binomial case),

$$dof = (n_1 + n_2 + \ldots + n_N) - (p + 1)$$

or $dof = no.\ of\ covariate\ combinations - no.\ of\ estimated\ regression\ coefficients$.

- In general, large values of $D$ imply *model lack of fit*.
- $D$ not testing for Binomial assumption of the data.
- $D$ testing if one or more predictors have been omitted from the model.
- p-value for $D$: $P[D > D_{obs}]$ obtained from $\chi^2_{(dof)}$.

# Other 'diagnostics' (summaries)

- *pseudo* $R^2$ statistic,

$$pseudo\ R^2 = \frac{logL_s - logL(\hat{\beta})}{logL_s}$$

where $L_S$ is the max. likelihood for the saturated model and $L(\hat{\beta})$ is the max likelihood for a model with covariates.
- "proportional improvement in log-likelihood".
- Another pseudo $R^2$ statistic (Mc Fadden's)

$$pseudo\ R^2 = \frac{logL(\hat{\beta}_0) - logL(\hat{\beta})}{logL(\hat{\beta}_0)}$$

- Efron's

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{Y})^2}$$

- *Residuals* $Y_i$ number of successes. $n_i$ number of trials.
- $\hat{\pi}_i$ estimated probability of success based on a *glm*
- *Pearson chi-square residuals*

$$r_i = \frac{Y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}; i = 1, 2, \ldots, N$$

- Chi-square statistic,

$$X^2 = \sum_{i=1}^{N} r_i^2$$

  has the same dofs as $D$, $N - (p + 1)$.
- How does deviance work for a Poisson regression?