**Estimation Problems**

- **Q**: if we observe a dataset that has all 0 at lower dose and all 1 at high dose, what does it imply? Is it a good thing?

![Graph showing the effect of dose on probability](image)

1. Clearly, the effect of dose on the probability is very significant.
2. But, cannot fit a line for the probability implies unstable estimation ($\beta$ is unstable). ($\Rightarrow$ problem of unidentifiable in LM)

- The algorithm for MLE (introduced in future lecture) usually converge (approximate) the MLE fast.

- However, difficulties can sometimes arise $\Rightarrow$ convergence fail.

- The following is a condition that cause the failure $\Rightarrow$ the two groups are linearly separable.

![Graph showing linear separability](image)

- Deviance would be small $y \approx \hat{y}$, $\hat{\beta} \approx \frac{y}{n}$ $\Rightarrow$ good fitting.

- Estimation of $\beta$ is very unstable $\Rightarrow$ large standard error $\Rightarrow$ become insignificant.

- This is an “embarrassment of riches” $\Rightarrow$ Perfect fit is possible, but estimation is a problem.

- Lesson for data collection $\Rightarrow$ should get $y_x$ on some $x$’s where $y_x/n_x \neq 0$ or 1. **Q**: why there is no such issue in Normal $y$?

- Alternative fitting approaches
  - Exact logistic regression (Cox, 1970)
  - The work of Cyrus Mehta (1995)
  - Bias reduction method of Firth (1993): remove the $O(1/n)$ term from the asymptotic bias of estimated coefficients.

- Instability in parameter estimation will also occur in datasets that approach linear separability.

$\lim_{n \to \infty} E(\hat{\beta}_{MLE}) = \beta$

$E(\hat{\beta}_{MLE} - \beta) \sim O(\frac{1}{n})$
Alternative Goodness-of-Fit Measure

- Recall: deviance is one measure of how well the model fits the data. \( Q \): are there others?
- Pearson’s \( X^2 \) for count data.

\[
X^2 = \sum_{j=1}^{k} \sum_{i=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}
\]

where \( O_{ij} \): observed counts for \( ij^{th} \) cell
\( E_{ij} \): expected counts for \( ij^{th} \) cell

For a binomial response:

number of successes: \( O_{i1} = y_i \) while \( E_{i1} = n_i \hat{\beta}_i \)

number of failures: \( O_{i2} = n_i - y_i \) while \( E_{i2} = n_i (1 - \hat{\beta}_i) \)

which results in

\[
X^2 = \sum_{i=1}^{k} \frac{(y_i - n_i \hat{\beta}_i)^2}{n_i \hat{\beta}_i (1 - \hat{\beta}_i)}
\]

Some properties of the Pearson’s \( X^2 \) statistic:

- Pearson’s \( X^2 \) typically close in size to the deviance (why?)

\[
D = 2 \sum_{i=1}^{k} \left\{ y_i \log \left( \frac{y_i}{\hat{y}_i} \right) + (n_i - y_i) \log \left( \frac{n_i - y_i}{n_i - \hat{\beta}_i + \hat{\beta}_i} \right) \right\}
\]

\[
= 2 \sum_{i=1}^{k} \left\{ (y_i - \hat{y}_i) + \frac{1}{2 \hat{y}_i} (y_i - \hat{y}_i)^2 + \ldots \right\} + \left\{ (n_i - y_i - \hat{y}_i + \hat{\beta}_i) + \frac{1}{2 (n_i - \hat{\beta}_i)} (n_i - y_i - n_i + \hat{\beta}_i)^2 + \ldots \right\}
\]

\[
\approx \sum_{j=1}^{2} \sum_{i=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = X^2
\]

- Pearson’s \( X^2 \) can be used in the same manner as the deviance

- Alternative versions of hypothesis tests described above: use \( X^2 \) in place of the \( D \) (with same approx. null distribution)

However, some care is necessary because the model is fit (i.e., \( \hat{\beta} \) is estimated) to minimize the deviance and not the Pearson’s \( X^2 \)

It is possible, although unlikely, that the \( X^2 \) could increase as a predictor is added to the model
We can define *Pearson residuals* as:

\[ r_i^P = \frac{y_i - n_i \hat{p}_i}{\sqrt{\text{var}(y_i)}} = n_i \hat{p}_i (1 - \hat{p}_i) \]

which can be viewed as a type of standardized residual.

* Generalized $R^2$

Recall: $R^2$ for Normal linear model is a popular goodness-of-fit measure, which represents the *proportion of variance explained*

**Approach 1:** proportion of deviance explained

\[ \frac{D_{\text{full}} - D_{\text{null}}}{D_{\text{full}}} \]

\[ \frac{D_{\text{full}}}{D_{\text{null}}} \]

**Approach 2:** interpretation of $R^2$ from likelihood viewpoint,

\[ \frac{R^2}{\text{RSS}} = 1 - \left( \frac{\hat{L}_0}{\hat{L}} \right)^{2/K} \]

where

- $K$: number of observations
- $\hat{L}_0$: maximized likelihood of $K$ obs. under null model
- $\hat{L}$: maximized likelihood of $K$ obs. under assumed model

**Generalization to binomial case:**

- $K$: number of Bernoulli observations (i.e., $K = n_1 + n_2 + \ldots + n_k$)
- $\hat{L}_0$: the maximized likelihood under null model $\eta_X = \beta_0$ (class)
- $\hat{L}$: the maximized likelihood of $K$ obs. under assumed model

\[ \eta_X = \beta_0 + \beta_1 h_1(x_1, \ldots, x_m) + \ldots + \beta_{p-1} h_{p-1}(x_1, \ldots, x_m) = X \beta \]

However, Nagelkerke (1991) pointed out that for *discrete models*, i.e., models whose likelihood is a product of probabilities, instead of densities, $0 \leq \text{prob.} \leq 1, 0 \leq \text{pdf}

\[ \max(R^2) = 1 - \left( \frac{\hat{L}_0}{\hat{L}_{\text{full}}} \right)^{2/K} = 1 - \frac{(\hat{L}_0)^{2/K}}{\hat{L}} \leq \text{upper bound} < 1 \]

**Nagelkerke (1991)** suggested (for a fitted model $M$)

\[ \hat{R}^2 = \frac{R^2}{\max R^2} = \frac{1 - (\hat{L}_0/\hat{L})^{2/K}}{1 - (\hat{L}_0)^{2/K}} = \frac{1 - \exp \left( (D_{\text{full}} - D_{\text{null}}) / K \right)}{1 - \exp \left( -D_{\text{null}} / K \right)} \]

\[ \Rightarrow 0 \leq \hat{R}^2 \leq 1 \]

* Reading: F, 2.9
Overdispersion

• Recall 1: if the binomial GLM model specification is correct, i.e.,
  \[ y_x \sim B(n_x, p_x = g^{-1}(\eta_x)) \]

Recall: \( E(y_x) = n_x p_x \quad \text{and} \quad Var(y_x) = n_x p_x (1 - p_x) \leq \frac{n_x}{q} \), max at \( p_x = \frac{1}{2} \)

\[ \Rightarrow \quad D (\text{deviance}) \overset{\Delta}{=} \chi^2_{k-p} \Rightarrow \text{can perform goodness-of-fit test} \]

Q: why no such goodness-of-fit test for Normal linear model if no further assumption/information about \( \sigma^2 \) is offered?

Recall 2: \[ D \approx X^2 = \sum (r_i^P)^2 \]

Q: what cause large \( D \)? \( \Rightarrow \) may suspect \( Var(y_x) \gg n_x p_x (1 - p_x) \)

• Some possible explanation for large deviance (c.f., possible reasons causing \( \sigma^2 \) larger than \( \sigma^2 \) in Normal linear model)

  ➢ Sparse data (i.e., \( n_x \)’s too small) \( \Rightarrow D \overset{\Delta}{=} \chi^2_{k-p} \) is questionable
  ➢ Presence of outlier (can be detected in diagnostics for GLM)

  ▪ For larger number of outliers, we might conclude that they are unexceptional \( \Rightarrow \) something amiss with other structures