VI. Regression models for probabilities instead of rates

B&D2: 4.7

A. Proportional Mortality

1. How do risks of two (mutually exclusive) events compare?
   a. Common offset \( \omega_k \)
   b. Assume \( x_j \sim \mathcal{P}(\lambda_k) \), \( \lambda_k = \exp(\omega_k + z_{kj}) \).
   c. Then \( X_{kj} \mid x_k = x_+ \sim \text{Bin}(\pi_{k1}, x_+) \) for \( \pi_{k1} = \lambda_k/(\lambda_k + \lambda_2) \)
   d. \( \pi_{k1} = \exp(\omega_k + z_{k1})/[\exp(\omega_k + z_{k1}) + \exp(\omega_k + z_{k2})] = \exp((z_{k1} - z_{k2})/\beta)/[\exp((z_{k1} - z_{k2})/\beta) + 1] \)

2. Continuous covariates may also be used
   a. Add in score as a covariate
   b. Consider adding polynomial terms to test whether odds ratio is really constant.
   c. Normal CDF is sometimes used
      i. Mild effect of collinearity
      ii. Impact can be minimized by subtracting out mean.
   d. Recall conditioning on \( X_{kj} \) removes effect of \( \mu \)

3. We have too many parameters
   a. Can decrease \( \mu \) and increase each other \( \alpha_k \) and get same probabilities
   b. Three typical solutions:
      i. Set \( \mu = 0 \): Results in separate log odds fits for each row.
      ii. Set \( \sum_{k=0}^{K-1} \alpha_k = 0 \): Makes \( \mu \) an “average” log odds, and rest are log odds ratios in comparison to average.
      iii. Set \( \alpha_k = 0 \) for some \( k' \in \{0, \ldots, K-1\} \)
         - Makes group \( k' \) the reference group
         - \( \mu \) represents log odds for reference group
         - \( \alpha_k \) is the log odds for group \( k \) with respect to group \( k' \)
         - Typically choose \( k' \) as 1 or \( K \)
   c. Unlike Mantel–Haenszel approach, this approach is not conditional on disease numbers in each table.

5. Approach can be extended to scored categories.
   a. Add in score as a covariate
   b. As with simpler regression models, one should consider the proper scale for continuous covariates
   c. Consider adding polynomial terms
   d. Estimates could be seriously impacted by other variables in model
      i. Mild effect of collinearity
      ii. Impact can be minimized by subtracting out mean.

7. You can use another function instead of logit
   a. Must still map \( \Re \) into \([0, 1]\)
   b. Logit has some mathematical properties we will discuss later
   c. Normal CDF is sometimes used
      i. Called the probit
      ii. Results from discretizing standard multiple regression.
         - Suppose \( Y_j = x_j \beta + \sigma \epsilon_j \), \( \epsilon_j \sim N(0, 1) \)

A: 4.1.4, 4.3

D. Logistic regression for \( K \times 2 \) tables:

1. \( x_{kj} \mid x_k = x_+ \sim \text{Bin}(x_+, 1/(1 + \exp(-\mu - \alpha_k))) \)
2. For \( 2 \times 2 \) table analysis, cohort study (exposed and unexposed group sizes fixed)
   a. Recall notation: \( x_{kj} \) is number of

\[ \begin{align*}
\text{cases} & \quad \text{if } j = 1 \\
\text{exposed} & \quad \text{if } k = 1 \\
\text{none} & \quad \text{if } k = 0
\end{align*} \]

A: 8 pp. 243–262

Fig. 9: Comparison of Probit and Logistic Link Functions