Incorporating Parameter Constraints in GLMs

Often an investigator may have knowledge that a regression coefficient falls within some constrained interval.

For example, one may be willing to assume that the association between smoking and lung cancer was strictly positive in a logistic regression analysis examining other predictors of lung cancer

\[
\text{logit} \Pr(T_i = j \mid T_i \geq j, x_i) = \alpha_j + Smk_i \beta_1 + \text{Radon}_i \beta_2,
\]

where \( Smk_i \) indicates smoking for individual \( i \), and \( \text{Radon}_i \) denotes level of exposure to Radon.
The constraint that smokers have a higher discrete hazard of lung cancer onset in age interval $j$ can be formalized by choosing a prior for $\beta_1$ with support on

$$\Omega = \{\beta : \beta > 0\} \subset \mathbb{R}^1.$$ 

A convenient choice is the truncated normal density,

$$\pi(\beta_1) = \frac{1(\beta_1 > 0) \, N(\beta_1; \mu_1, \sigma_1^2)}{\Phi(\mu_1/\sigma_1)},$$

which assigns zero prior probability to values of $\beta_1$ less than 1.
Suppose we have a normal linear model, 

\[ y_i = \alpha + x_i \beta + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2), \]

and we choose the prior, 

\[ \pi(\alpha) = N(\alpha; \alpha_0, \sigma^2_\alpha), \quad \pi(\beta) \propto 1(\beta > 0)N(\beta; \beta_0, \sigma^2_\beta), \quad \pi(\sigma^2) = G(a_0, b_0). \]

Well, for \( \alpha \) and \( \sigma^{-2} \) we have standard closed form normal and gamma full conditional distributions.
For $\beta$ we have

$$\pi(\beta \mid \alpha, \sigma^2, \text{data}) \propto 1(\beta > 0) \exp \left[ -\frac{1}{2} \left\{ \sigma_{\beta}^{-2}(\beta - \beta_0)^2 + \sigma^{-2} \sum_{i=1}^{n} (y_i^* - x_i\beta)^2 \right\} \right],$$

$$= \frac{1(\beta > 0) \mathcal{N}(\beta; \hat{\beta}, \hat{\sigma}_\beta^2)}{\Phi(\beta / \hat{\sigma})},$$

where $y_i^* = y_i - \alpha$, $y^* = (y_1^*, \ldots, y_n^*)'$, $x = (x_1, \ldots, x_n)'$,

$$\hat{\beta} = \hat{\sigma}_\beta^2 (\sigma_{\beta}^{-2} \beta_0 + \sigma^{-2} x' y^*)$$

and

$$\hat{\sigma}_\beta^2 = (\sigma_{\beta}^{-2} + \sigma^{-2} x' x)^{-1}.$$

Thus, we have a closed form full conditional even under the constraint. This can be used for Gibbs sampling in probit models by modifying Albert and Chib (1993).
A categorical variable is one for which the measurement scale consists of a set of categories.

For example, political philosophy may be classified as (1) liberal, (2) moderate or (3) conservative or pups can be classified as (1) alive & normal; (2) malformed; or (3) dead.

Often, the explanatory variables may also be categorical.
Definitions

• *Nominal Variables*: categorical variables for which levels do not have a natural ordering (a.k.a., unordered categorical). E.g., religious affiliation.

• *Ordinal Variables*: categorical variables for which levels are ordered. E.g., social class, how conservative a person is, amount of physical activity, etc.

• *Interval Variable*: One that has numerical distances between any two levels of the scale. E.g., age or income.
Two-Way Contingency Tables

Let $X$ and $Y$ denote two categorical response variables, $X$ having $I$ levels and $Y$ having $J$ levels.

When we classify subjects on both variables, there are $IJ$ possible combinations.

The response $(X, Y)$ of a subject randomly chosen from some population has a probability distribution.

The *cells* of the table represent the $IJ$ possible outcomes.
Let $\pi_{ij}$ denote the probability that $(X, Y)$ falls in the cell in row $i$ and column $j$.

When the cells contain frequency counts of outcomes, the table is called a *contingency table* or *cross-classification table*.

The probability distribution $\{\pi_{ij}\}$ represents the joint distribution of $X$ and $Y$, while the marginal distributions are row and column totals obtained by summing the joint probabilities,

$$
\pi_{i+} = \sum_{j=1}^{J} \pi_{ij} \quad \text{and} \quad \pi_{+j} = \sum_{i=1}^{I} \pi_{ij}.
$$
Given that a subject is classified in row $i$ of $X$, let $\pi_{j|i}$ denote the probability of classification in row $j$ of $Y$.

Many studies seek to compare the conditional distribution of $Y$ at various levels of $X$.

The conditional cdf

$$F_{j|i} = \sum_{b \leq j} \pi_{b|i}, \quad j = 1, \ldots, J,$$

equals the probability of classification in one of the first $j$ columns, given classification in row $i$.

If $F_{j|h} \leq F_{j|i}$ for $j = 1, \ldots, J$, then the conditional distribution in row $h$ is \textit{stochastically higher} than the one in row $i$. 
Independence

The conditional distribution of $Y$ given $X$ is related to the joint distribution by
\[
\pi_{j|i} = \frac{\pi_{ij}}{\pi_i^+} \quad \text{for all } i \text{ and } j.
\]

The variables are statistically independent if all joint probabilities equal the product of their marginal probabilities,
\[
\pi_{ij} = \pi_i^+ \pi_j^+ \quad \text{for all } i \text{ and } j.
\]

When $X$ and $Y$ are independent,
\[
\pi_{j|i} = \frac{\pi_{ij}}{\pi_i^+} = \frac{(\pi_i^+ \pi_j^+)}{\pi_i^+} = \pi_j^+ \quad \text{for } i = 1, \ldots, I.
\]
Measures of Ordinal Association

Does $Y$ tend to increase as $X$ increases?

When we observe the ordering of two subjects on each of two variables, we can classify the pair as discordant or concordant.

The pair is concordant if the subject ranking higher on $X$ also ranks higher on $Y$, otherwise the pair is discordant or tied.
Consider two independent observations from a joint probability distribution for two ordinal variables.

For that pair of observations,

\[ \Pi_c = 2 \sum_i \sum_j \pi_{ij} \left( \sum_{h>i} \sum_{k>j} \pi_{hk} \right) \quad \text{and} \quad \Pi_d = 2 \sum_i \sum_j \pi_{ij} \left( \sum_{h<i} \sum_{k<j} \pi_{hk} \right) \]

are the probabilities of concordance and discordance.

*Gamma* is the difference between the probabilities of concordance and discordance, given that the pair is untied,

\[ \gamma = \frac{\Pi_c - \Pi_d}{\Pi_c + \Pi_d}. \]

A sample estimate of gamma is \( \hat{\gamma} = (C - D)/(C + D) \), where \( C \) and \( D \) are the total number of concordant and discordant pairs.