Anthony Davison

Problem 1 Suppose that we observe a binary response and want to use a probit link to relate it to the observed value x of a scalar covariate X, i.e., $\mathrm{P}(Y = 1 \mid X = x) = \Phi(\beta_0 + \beta_1 x)$. However, X is measured with error, and we can only observe the surrogate W, which satisfies $W \mid X = x \sim \mathcal{N}(x, \gamma^2)$; put another way, $W = X + \gamma\delta$, where $\delta \sim \mathcal{N}(0, 1)$. For example, W might be stated weekly consumption of dairy products in a food questionnaire, while X is the unobserved true consumption; notice that this model assumes no systematic under- or over-estimation, as $\mathrm{E}(W) = \mathrm{E}(X)$. Suppose below that Y is independent of the measurement error $W - X$.

(a) Show that $\mathrm{P}(Y = 1 \mid W = w) = \Phi(\beta_0' + \beta_1' w)$, where $\beta_0' = \beta_0/\sqrt{(1 + \beta_1^2\gamma^2)}$ and $\beta_1' = \beta_1/\sqrt{(1 + \beta_1^2\gamma^2)}$.

(b) What does this imply about probit regression of Y on the observed value of W? What is the effect on the estimated coefficients? What can be said if γ is unknown?

(c) What if there is a systematic bias in W? For example, true alcohol consumption X might be systematically underestimated by the reported value W.

Hint: Write $\mathrm{P}(Y = 1 \mid X = x) = \mathrm{P}(\varepsilon \le \beta_0 + \beta_1 x)$ for some underlying $\varepsilon \sim \mathcal{N}(0, 1)$.

Problem 2

(a) If $Y_1, \dots, Y_D$ are independent Poisson variables with means $\mu_1, \dots, \mu_D$, show that their joint distribution conditional on $S = \sum_{d=1}^D Y_d = m$ is multinomial with probabilities $\pi_d = \mu_d / \sum_{d'=1}^D \mu_{d'}$ and denominator m, i.e.,

$$\mathrm{P}(Y_1 = y_1, \dots, Y_D = y_D \mid S = m) = \frac{m!}{\prod_{d=1}^D y_d!} \prod_{d=1}^D \pi_d^{y_d}, \quad y_1, \dots, y_D \in \{0, \dots, m\}, \sum_{d=1}^D y_d = m.$$

(b) When m independent individuals are put into categories $1, \dots, D$ with respective probabilities $\pi_1, \dots, \pi_D$, let I_{jd} denote the indicator variable that the jth individual is placed into category d; note that $I_{j1} + \cdots + I_{jD} = 1$, as each individual can only be placed into one category. Find the joint density of $R_1, \dots, R_D$, where $R_d = \sum_{j=1}^m I_{jd}$.

(c) Given that $(R_1, \dots, R_D) \sim \mathrm{Mult}(m; \pi_1, \dots, \pi_D)$, for some $D \ge 5$, find the joint distribution of $(R_1 + R_2, R_3 + R_4, R_5, \dots, R_D)$ and the joint distribution of $(R_1, R_2, R_3 + R_4)$ conditional on $R_5 + \cdots + R_D = n$. Give the general version of this result.