Homework Assignment 3

STAT 544

Due on March 12, 2021

1. Suppose that $n_i Y_i$ has a $\textsc{Binomial}(n_i, \pi_i)$ distribution. Consider the binary GLM $\pi_i = F(\boldsymbol{x}_i'\boldsymbol{\beta})$, where F is the standard cdf of some family of continuous distributions. Find $w_i = (\partial\mu_i/\partial\eta_i)^2/\mathbb{V}Y_i$ and, hence, $\mathbb{V}\hat{\boldsymbol{\beta}}$.

2. Suppose the logistic model holds in which x has a $\textsc{Uniform}(0, 100)$ distribution, and $\text{logit}(\pi_i) = -2 + 0.04\, x_i$. Randomly generate 100 independent observations from this model. Plot the residuals against x and against the fitted values. Why do the residual plots for binary data have this appearance?

3. Let Y_i be a $\textsc{Bernoulli}(\pi_i)$ variate for $i = 1, \ldots, N$. For the model $\text{logit}(\pi_i) = \beta_0 + \beta_1\, x_i$, show that the deviance depends on $\hat{\pi}_i$ but not Y_i. Hence, the deviance is not useful for checking model fit. Explain what this means for ungrouped binary data.

4. An alternative latent variable model for binary data results from early applications of binary response models to toxicology studies of the effect of dosage of a toxin on whether a subject dies, with an unobserved *tolerance distribution*. For a randomly selected subject, ket x_i denote the dosage level and let $Y_i = 1$ if the subject dies. Suppose that the subject has a latent tolerance threshold τ_i for the dosage, with $Y_i = 1$ equivalent to $\tau_i \leq x_i$. Let $F(t) = \mathbb{P}(\tau \leq t)$.

   (a) For a fixed dosage x_i, explain why $\mathbb{P}(Y_i = 1 \mid x_i) = F(x_i)$.

   (b) Suppose F belongs to the Gaussian parametric family, for some μ and σ. Explain why the model has the form

   $$\Phi^{-1}(\pi_i) = \beta_0 + \beta_1\, x_i$$

   and relate β_0 and β_1 to μ and σ.

5. In one of the first studies linking smoking and lung cancer, Richard Doll and Austin Bradford Hill collected data from 20 hospitals in London, England. Each patient admitted with lung cancer in the preceding year was queried about their smoking behavior. For each of the 709 patients admitted, they recorded the smoking behavior of a non-cancer patient at the same hospital and having the same gender and falling within the

same five-year age group. A smoker was defined as a person who had smoked at least one cigarette per day for at least one year. Of the 709 cases having lung cancer, 650 reported being smokers. Specify a relevant logistic regression model, explain what can be estimated and what cannot (and why), and conduct a complete statistical analysis.

6. The sore throat data are from a study about whether a patient having surgery experienced a sore throat on waking ($Y = 1 = \text{yes}, Y = 0 = \text{no}$) as a function of d, the duration of the surgery in minutes, and t, the type of device used to secure the airway ($1 = \text{tracheal tube}, 0 = \text{laryngeal mask}$). Use a model-building strategy to select a GLM for binary data. Interpret parameter estimates and conduct inference about the effects.